Haavelmo, Trygve (1911–1999)

Agnar Sandmo

Keywords

Balanced budget multiplier; Capital theory; Econometrics; Haavelmo, T.; Inequality (global); Investment theory; Neoclassical theory of capital accumulation; Probability

JEL Classifications

B31

Haavelmo was born in Skedsmo, Norway. He graduated from the University of Oslo in 1933 and joined Ragnar Frisch's newly created Institute of Economics as a research assistant. He spent the war years working for the Norwegian government in the United States. After a year's stay at the Cowles Commission at the University of Chicago, he returned to Norway in 1947, becoming professor of economics at the University of Oslo in 1948. He retired from his chair in 1979. In 1989 he was awarded the Nobel Memorial Prize in Economics, the Nobel citation referring to 'his clarification of the probability theory foundations of econometrics and his analyses of simultaneous economic structures'.

Haavelmo first made his name by a series of path-breaking contributions to the theory of econometrics, most of which were written during his years in the United States. His 1943 article in *Econometrica* was the first to consider the statistical implications of simultaneity in economic models. This paper was one of the main sources of inspiration for the extensive work carried out in this area over the next decade, particularly at the Cowles Commission. Haavelmo developed his ideas further in the famous 1944 supplement to *Econometrica*; the main general contribution of this work was to base econometrics more firmly on the foundations of probability theory.

After his return to Norway, Haavelmo turned away from econometrics to economic theory as his main field of interest. In his 1957 presidential address to the Econometric Society (published the next year) he emphasized the need for a more solid theoretical foundation for empirical work as well as the need for theory to be inspired by empirical research.

Haavelmo's *Study in the Theory of Economic Evolution* (1954), is a broad exploration of the contributions that analytical economics can make to the understanding of global economic inequality. As an early contribution to growth theory it is less notable for simple models and precise theorems than for its imaginative and experimental attitude towards hypotheses concerning population growth, education, migration and the international struggle for redistribution. The open-mindedness of the approach is very characteristic of the author.

Similar remarks apply to his 1960 book, *A Study in the Theory of Investment.* Its main

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

objective is to provide a firmer microeconomic foundation for the macroeconomic theory of investment demand. To this end Haavelmo probes deeply into capital theory, emphasizing strongly, however, that a theory of optimum capital use does not in itself provide a theory of investment. This insight, and his clear statement of what has since been known as the neoclassical theory of capital accumulation, has been a major influence on late work in this area, both theoretical and applied.

Of Haavelmo's other contributions to economic theory, special mention should be made of his 1945 analysis of the balanced budget multiplier. The expansionary effect in a Keynesian unemployment situation of a balanced increase of public expenditure and taxes had been pointed out before, but Haavelmo was the first to provide a rigorous theoretical analysis of it.

Haavelmo was also been very active as a teacher. His lecture notes on a wide range of topics in economic theory exerted a formative influence on generations of Norwegian economists.

Selected Works

1943. The statistical implications of a system of simultaneous equations. *Econometrica* 11(January): 1–12.

1944. The probability approach in econometrics. Supplement to *Econometrica* 12 (July): S1–115.

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

1947a. Methods of measuring the marginal propensity to consume. *Journal of the American Statistical Society* 42(237): 105–122.

1947b. (With M.A. Girshick.) Statistical analysis of the demand for food: examples of simultaneous estimation of structural equations. *Econometrica* 15(April): 79–110.

1954. *A study in the theory of economic evolution*. Amsterdam: North-Holland.

1958. The role of the econometrician in the advancement of economic theory. *Econometrica* 26(July): 351–357.

1960. *A study in the theory of investment.* Chicago: University of Chicago Press.

1970. Some observations on welfare and economic growth. In *Induction, growth and trade: essays in honour of Sir Roy Harrod,* ed. W.A. Eltis, M.F.G. Scott, and J.N. Wolfe. Oxford: Clarendon Press.

Habakkuk, John Hrothgar (1915–2002)

Ravi Mirchandani

Keywords

Britain, economics in; Capital-intensive technology; Falling rate of profit; Habakkuk, J. H.; Labour supply; Technical change; United States, economics in

JEL Classifications

B31

Born in Wales in 1915, Habakkuk graduated from Cambridge in 1936, where he was a Fellow of Pembroke College from 1938 until 1950. He held the Oxford chair of economic history from 1950 to 1967, when he became Principal of Jesus College, Oxford. He retired in 1984. As a member of the Advisory Council on Public Records (1958–1970), the Royal Commission on Historic Manuscripts, and the British Library Organizing Committee, amongst other bodies, he was active in the field of public records; he was knighted in 1974.

His major contribution was to the study of the rates of technological change in Britain and America in the 19th century and the reasons for the much more rapid development and use of manufacturing technology in the United States. In his book, *American and British Technology in the Nineteenth Century* (1962), American industrial development is roughly divided into two important stages, the period before the first wave of immigration in the 1840s, which laid the ground for future development, and the period after 1870 when abundant natural resources and rapid growth of market demand provided the stimulus for growth.

Habakkuk argues that American technological development in the early period, by contrast with Britain, was stimulated by the high cost of labour relative to capital and the relative inelasticity of labour supply. The expanding manufacturer, to avoid a falling marginal rate of profit, was more likely than his British counterpart to look to capital-intensive and labour-saving technology. Though Habakkuk was also keen to stress the importance of social factors, the suspicion of British employers and the hostility of British labour to new techniques, his explanation of the disparity is grounded in economic relationships.

Habakkuk's thesis has come under considerable scrutiny; recent research has tended to suggest that there was considerable diversity, both on a regional basis and between different industries, in development on both sides of the Atlantic. Economic historians have also questioned the timing of significant development in the States and chosen to put greater stress on non-economic explanations.

Habakkuk also made notable contributions to the debates on British population growth in the late 18th century and on the changing pattern of landholding as smaller holdings gave way to larger units in the same period.

Selected Works

1953. English population in the eighteenth century. *English Historical Review* 6:117–33.

1962. *American and British technology in the nineteenth century: The search for labour-saving inventions*. Cambridge: Cambridge University Press.

1968. *Industrial organisation since the industrial revolution*. Southampton: University of Southampton Press.

1971. *Population growth and economic development since 1750*. Leicester: Leicester University Press.

1979–1981. The rise and fall of English landed families, 1600–1800. *Transactions of the Royal Historical Society* I 29: 187–207. II, 30: 199–221; III, 31: 195–217.

Haberler, Gottfried (1900–1995)

John S. Chipman

Keywords

Barone, E.; Business cycles; Comparative costs; Cost-of-living index; Factor mobility; Fixed exchange rates; Free trade; Haberler, G.; Homothetic preferences; Indirect utility functions; Inflation; International trade theory; Laspeyres index; Paasche index; Pigou effect; Production-possibility frontier; Purchasing power parity; Real-balance effect; Specific factor; Stagflation; Transfer problem

JEL Classifications

B31

Gottfried Haberler was born on 20 July 1900 in Purkersdorf, near Vienna. He studied economics at the University of Vienna under Friedrich von Wieser and Ludwig von Mises, where he received doctorates in law (1923) and economics (1925). After two years in the United States and Britain he returned to Vienna, received his habilitation in 1928, and was appointed lecturer, later professor, of economics, at the University of Vienna, from 1928 to 1936. He was appointed professor at Harvard University in 1936, where he remained until his retirement in 1971, after which he was a resident scholar at the American Enterprise Institute, Washington, DC. He was President of the International Economic Association (1950–1), the National Bureau of Economic Research (1955), and the American Economic Association (1963). In 1980 he was awarded the Antonio Feltrinelli prize.

Haberler's first major work was his habilitation thesis (1927), *The Meaning of Index Numbers*, summarized in Koo (1985, pp. 546–9). This work stimulated a great deal of subsequent research on the theory of the price or cost-of-living index. Haberler defined the 'true change in the price level' as 'the ratio of the money income in the first period to the money income in the second period that would leave the individual indifferent' (Koo 1985, p. 547). Haberler's main concern was to find conditions under which this 'true price index' would be bounded by the Laspeyres and Paasche price indices. Some of the difficulties with this approach (and with the similar, earlier approach of Konüs (1924)) were discussed by Bortkiewicz (1928), Neisser (1929), Staehle (1935), and Frisch (1936). Frisch remarked (p. 25) that Haberler's definition of the 'true change of the price level' involved an implicit assumption of expenditure proportionality (homothetic preferences), and attributed this point (but apparently without justification) to Bortkiewicz; he also interpreted Haberler (1929) in his reply to Neisser and Bortkiewicz as accepting this point. In terms of contemporary concepts we may say that homothetic preferences characterize indirect utility functions of the form $Y/C(p)$ where Y is income and $C(p)$ is a homogeneous-of-degree-1 function of prices.

Undoubtedly Haberler's most significant contribution was his reformulation of the theory of comparative costs (Haberler 1930a), which revolutionized the theory of international trade. Prior to this paper, the Ricardian theory still held sway, but had been so amended with ill-defined concepts such as 'real cost' and 'units of productive power' taking the place of labour allocation that it had lost all its simplicity and elegance. Haberler introduced the production 'substitution curve' (now usually known as the production-possibility frontier), allowing for several factors of production, and taken to be concave to the origin as a result of diminishing returns. This laid the foundations for Ohlin's theory, as well as Lerner's and Samuelson's. True, as recently brought to light by Maneschi and Thweatt (1987), a footnote contained in the posthumous edition of Barone's *Principi* (1936, pp. 170–3), depicting a (non-concave) production-possibility frontier and a community indifference curve, was actually present in the first (1908) edition – but not subsequent ones; hence Barone must be accorded priority. But Haberler's independent discovery – and the use to which he put it – is what transformed the theory of international trade. Haberler also introduced the concept of a 'specific factor' – one that is completely immobile among industries – and used this concept with great effect in Haberler (1950) to illustrate the proposition that the gains from trade do not depend on the assumption of factor mobility.

Haberler made numerous other contributions to international economics, including (1) his synthesis and clarification of the Keynes–Ohlin debate on the transfer problem (Haberler 1930b); (2) his judicious use of purchasing-power-parity calculations to set exchange rates (Haberler 1945); (3) his introduction of the concept of supply and demand schedules for foreign exchange (1936) and his subsequent use of them (Haberler 1949) in qualified support of the proposition that a devaluation in a pegged-rate regime could improve a country's balance of payments – but subject to the important proviso (1949, p. 213) that it would, through monetary expansion, likely shift these schedules; (4) his advocacy of free trade as the best policy for developing countries (Haberler 1959); (5) numerous contributions to past and current history of international economic relations (cf. Koo 1985).

The third area in which Haberler made major contributions is business-cycle theory (Haberler 1937, 1942). His classic synthesis, notably in the third edition of *Prosperity and Depression* (1941), introduced the important 'real-balance effect', initially called the 'Pigou effect' by Patinkin (1948), although Patinkin in his 1951 revision acknowledged Haberler's priority over Pigou (1943). In the 1970s and 1980s Haberler furnished trenchant analyses of the phenomenon of worldwide inflation and the political economy of stagflation (cf. Koo 1985), displaying the unique combination of clarity and wisdom that are characteristic of his writings.

Information on Haberler's life and work may be found in Schuster (1979), Chipman (1982), Baldwin (1982), Officer (1982), and Willett

(1982). A complete bibliography of his writings is contained in Koo (1985).

Selected Works

1927. *Der Sinn der Indexzahlen.* Tübingen: J.C.B. Mohr (Paul Siebeck).

1929. Der volkswirtschaftliche Geldwert und die Preisindexziffern. *Weltwirtschaftliches Archiv* 30(July): 6**–14**.

1930a. Die Theorie der komparativen Kosten und ihre Auswertung für die Begründung des Freihandles. *Weltwirtschaftliches Archiv* 32(July): 350–370. Trans. as 'The theory of comparative costs and its use in the defense of free trade' in Koo (1985).

1930b. Transfer und Preisbewegung. *Zeitschrift für Nationalökonomie* 1: 547–554; 2, 100–2. Trans. as 'Transfer and price movements' in Koo (1985).

1933. *Der internationale Handel. Theorie der weltwirtschaftlichen Zusammenhänge sowie Darstellung und Analyse der Aussenhandelspolitik.* Berlin: Julius Springer. Translated (revised by the author) as The theory of international trade with its applications to commercial policy. London: William Hodge & Co., 1936.

1937. *Prosperity and Depression.* Geneva: League of Nations. 3rd edition enlarged by Part III, 1941. 5th and 6th editions, Cambridge, MA: Harvard University Press, 1964.

1942. *Consumer Instalment Credit and Economic Fluctuations.* New York: NBER.

1945. The choice of exchange rates after the war. *American Economic Review* 35, 308–318.

1949. The market for foreign exchange and the stability of the balance of payments. *Kyklos* 3(3): 193–218. Reprinted in Koo (1985).

1950. Some problems in the pure theory of international trade. *Economic Journal* 60: 223–40. Reprinted in Koo (1985).

1955. *A survey of international trade theory.* Special papers in economics no. 1. Princeton: International Finance Section, Princeton University. Revised and enlarged edn, 1961. Reprinted in Koo (1985).

1959. *International trade and economic development.* Cairo: National Bank of Egypt. Reprinted in Koo (1985).

Bibliography

Baldwin, R.E. 1982. Gottfried Haberler's contributions to international trade theory and policy. *Quarterly Journal of Economics* 97: 141–159.

Barone, E. 1908. *Principi di economia politica*. Rome: Tipografia Nazionale di Giovanni Bertero e C. German translation from the 3rd (1913) edn, *Grundzüge der theoretische Nationalökonomie,* Bonn: Kurt Schroeder, 1927. Posthumous edition, *Le opere economiche*, vol. 2, Bologna: Nicola Zanichelli Editore, 1936.

Chipman, J.S. 1982. Salute to Gottfried Haberler on the occasion of his 80th birthday. *Journal of International Economics* [Supplement], January, 25–30.

Frisch, R. 1936. Annual survey of general economic theory: The problem of index numbers. *Econometrica* 4: 1–38.

Konüs, A.A. 1924. The problem of the true index of the cost of living. *Economic Bulletin of the Institute of Economic Conjuncture* 9–10, October, 64–71. Translated from the Russian in *Econometrica* 7(1939), 10–29.

Koo, A.Y.C., ed. 1985. *Selected essays of Gottfried Haberler.* Cambridge, MA: MIT Press.

Maneschi, A., and W.O. Thweatt. 1987. Barone's 1908 representation of an economy's trade equilibrium and the gains from trade. *Journal of International Economics* 22: 375–382.

Neisser, H. 1929. Der volkswirtschaftliche Geldwert und die Preisindexziffern. *Weltwirtschaftliches Archiv* 29 (Part I): 6**–18**. Schlusswort, Part II, July, 14**–7**.

Officer, L.H. 1982. Prosperity and depression – And beyond. *Quarterly Journal of Economics* 97: 149–159.

Patinkin, D. 1948. Price flexibility and full employment. *American Economic Review* 38: 543–564. Revised version in *Readings in monetary theory,* ed. F.A. Lutz and L.W. Mints, Philadelphia: Blakiston, 1951.

Pigou, A.C. 1943. The classical stationary state. *Economic Journal* 53: 343–351.

Schuster, H., ed. 1979. Univ.-Prof. Dr. Gottfried Haberler. In *Österreicher, die der Welt gehören.* Vienna: Mobil Oil Austria AG.

Staehle, H. 1935. A development of the economic theory of price index numbers. *Review of Economic Studies* 2 (June): 163–188.

von Bortkiewicz, L. 1928. Review of *Der Sinn der Indexzahlen* by Gottfried Haberler. *Magazin der Wirtschaft* 4 (11): 427–429.

Willett, T.D. 1982. Gottfried Haberler on inflation, unemployment, and international monetary economics: An appreciation. *Quarterly Journal of Economics* 97: 161–169.

Habit Persistence

Stephanie Schmitt-Grohé and Martin Uribe

Abstract

This article reviews the concept of habit persistence and its application in macroeconomics and finance. Special attention is given to the role of habit persistence in explaining the equity premium puzzle, observed business-cycle fluctuations and inflation dynamics, and in generating a theory of counter-cyclical markups of prices over marginal costs.

Keywords

Asset pricing; Business-cycle fluctuations; Currency pegs; Deep habits; Equity premium puzzle; Euler equation; Habit persistence; Imperfect competition; Inflation dynamics; Investment adjustment costs; Partial equilibrium; Risk aversion; Sticky prices; Variable capacity utilization

JEL Classifications

E21; D11; D91

Habit persistence, or 'habit formation' in its most common representation, is a preference specification according to which the period utility function depends on a quasi-difference of consumption. Specifically, if the utility function without habit formation is given by $\sum_{t=0}^{\infty} \beta^t U(c_t)$, where c_t denotes consumption in period t, U denotes the period utility function, and $\beta \in (0,1)$ denotes the subjective discount factor, then the utility function with habit persistence is given by $\sum_{t=0}^{\infty} \beta^t U(c_t - \alpha c_{t-1})$. The parameter $\alpha \in (0, 1)$ denotes the intensity of habit formation and introduces non-separability of preferences over time. Under habit persistence, an increase in current consumption lowers the marginal utility of consumption in the current period and increases it in the next period. Intuitively, the more the consumer eats today, the hungrier he wakes up tomorrow. It is in this sense that this type of preferences captures the notion of habit formation.

In the habit-forming preferences given above, past consumption represents the consumer's stock of habit in period t. More general specifications allow for the stock of habit to be a function of possibly all past consumptions. In this case, the period utility function is given by $U(c_t - \alpha S_{t-1})$, where $S_{t-1} = S(c_{t-1}, c_{t-2}, \ldots)$ denotes the stock of habit in period t. Often, the stock of habit is assumed to follow an autoregressive law of motion of the form $S_t = (1-\delta)S_{t-1} + \lambda c_t$. The parameter δ governs the rate of depreciation of the stock of habit, and the parameter λ measures the sensitivity of the stock of habit to current consumption.

A common variant of the habit persistence model is to treat habits as external to the consumer. When habits are external, the stock of habit depends on the history of aggregate past consumption as opposed to the consumer's own past consumption. Early formulations of the habit formation model, for example Pollak (1970), were cast in the external form. Since the work of Abel (1990), external habit formation has become known as 'catching up with the Joneses'. The external form of habit persistence simplifies the optimization problem of the consumer because the evolution of the stock of habit is taken as exogenous by the individual.

Another variation of the habit formation model is relative habit persistence, which features a quasi-ratio of consumption rather than a quasi-difference of consumption, as the argument of the period utility function (Duesenberry 1949; Abel 1990).

Habit Persistence and the Equity Premium Puzzle

Habit persistence has been proposed in financial economics as a possible solution to the equity premium puzzle first identified in the seminal work of Mehra and Prescott (1985). The equity premium puzzle is that, under the assumption of power utility and no habit persistence, observed excess returns of stocks over less risky assets, such as commercial paper, are too high to be consistent with actual consumption behaviour unless households are assumed to be extremely risk averse. At the heart of the equity premium puzzle lies the low volatility of observed consumption growth. To see this, note that a risky asset commands a high rate of return if it provides poor insurance against consumption fluctuations by paying plenty in periods of high consumption growth and little in periods of low consumption growth. If fluctuations in consumption growth are small (as is observed in the data), then high returns on risky assets can be supported only if one assumes that even minute consumption fluctuations are very painful to consumers. In other words, one must assume that consumers are extraordinarily risk averse.

With this intuition in mind, one can readily see why habit persistence has the potential to solve the equity premium puzzle. Habit-forming consumers dislike variations in habit-adjusted consumption, $c_t - \alpha S_{t-1}$, rather than variations in consumption itself, c_t. A given percentage change in consumption produces a much larger percentage change in habit-adjusted consumption than in consumption itself. In this way, small fluctuations in consumption growth can generate large variations in habit-adjusted consumption growth and hence explain sizable excess returns on risky assets even for moderate values of the degree of risk aversion. Early studies of the ability of habit persistence to resolve the equity premium puzzle include Sundaresan (1989), Abel (1990), and Constantinides (1990). Subsequent work has refined the habit-formation model to account for additional asset-pricing puzzles, such as the risk-free-rate puzzle and the forecastability of excess returns (see, for example, Campbell and Cochrane 1999).

Habit Persistence and the Business Cycle

In the asset-pricing literature, most applications of habit persistence are conducted within the context of partial equilibrium settings, in which private consumption is assumed to be exogenous. This assumption is not inconsequential. Indeed, it has been shown that, once a general equilibrium approach is adopted, in which consumption decisions are endogenous, the ability of habit persistence to reconcile the behaviour of asset prices and consumption is diminished. This is because habit formation induces excess smoothness in consumption expenditure (Lettau and Uhlig 2000). Boldrin et al. (2001) show that habit formation can help explain salient aspects of asset prices and business cycles only in combination with severe inflexibilities in factor markets.

Habit persistence features prominently in the literature devoted to the estimation of medium-scale macroeconomic models (for example, Christiano et al. 2005; Smets and Wouters 2004). The goal of this literature is to build dynamic general equilibrium models capable of explaining the observed behaviour at business-cycle frequency of a large number of macroeconomic variables. To this end, this literature has brought together in a single model most of the theoretical advances in business-cycle theory since the mid-1980s. Thus, these models include not only habit persistence but also other rigidities such as investment adjustment costs, variable capacity utilization, sticky product and factor prices, money demand by households and firms, and imperfect competition in product and factor markets. In the data, the response of consumption to expansionary shocks of various natures is hump-shaped, with the peak response occurring several quarters after the innovation. Such a response is hard to replicate in the absence of habit formation. For in this case consumption has a tendency to peak immediately after the shock and then to decline to its long-run level.

In the applications of habit formation discussed thus far, it matters little whether habits are of the internal or external type. The distinction is of importance in situations in which the consumer expects a regime shift of some nature in the future. A case in point is the consumption dynamics associated with temporary exchange- rate-based inflation stabilization programmes. Exchange-rate-based stabilization programmes, or currency pegs, constitute the most commonly used policy to control high inflation in emerging-market countries. It is well documented that, in response to the announcement of a currency peg, consumption rises initially, reaches a peak and then declines. Importantly, the observed eventual decline in consumption typically takes place before the currency peg is abandoned. Habit formation, be it of the internal or external type, can explain the observed gradual increase in consumption after the implementation of the stabilization plan (Uribe 2002). However, Uribe shows that the observed contraction in consumption that begins well before the collapse of the stabilization programme can be rationalized with internal habit formation but not with the external form of habits. In effect, maintaining a high consumption habit after the collapse of the stabilization programme is expensive because high inflation acts as a tax on consumption expenditures. When consumers internalize the habitual nature of consumption they start reducing their stock of habit – by cutting back consumption – before the price of consumption increases to mitigate the transition to a lower stock of habit. By contrast, when consumers do not internalize the habitual nature of their consumption, they continue to take advantage of the temporarily low price of consumption until the last day of the stabilization programme.

Deep Habits

All of the models of habit persistence discussed thus far in this article assume that habits are formed at the level of a single aggregate consumption good. An important consequence of this assumption is that the introduction of habit formation alters the propagation of macroeconomic shocks only in so far as it modifies the consumption Euler equation and possibly the household's labour supply schedule. Ravn et al. (2006), hereafter RSU, propose a general equilibrium model of habit formation on a good-by-good basis. They refer to this type of habit formation as 'deep habits'. They have in mind environments in which consumers can form habits separately over narrowly defined categories of goods, such as clothing, vacation destinations, music, and cars.

The assumption that agents can form habits on a good-by-good basis has two important implications for macroeconomic dynamics. First, the demand side of the macroeconomy – in particular the consumption Euler equation – is indistinguishable from that pertaining to an environment in which agents form habits over a single aggregate good. Second, and more significantly, the assumption of deep habit formation alters the supply side of the economy in fundamental ways. Specifically, when habits are formed at the level of individual goods, firms take into account the fact that the demand they will face in the future depends on their current sales. This is because higher consumption of a particular good in the current period makes consumers, all other things equal, more willing to buy that good in the future through the force of habit. Thus, when habits are deeply rooted, the optimal pricing problem of the firm becomes dynamic.

RSU embed the deep-habit-formation assumption in an economy with imperfectly competitive product markets. This combination results in a model of endogenous, time-varying markups of prices over marginal cost. A central result of RSU's work is that in the deep habit model markups behave counter-cyclically in equilibrium. In particular, expansions in output driven by demand shocks are accompanied by declines in markups. This implication of the deep-habit model is in line with the existing empirical literature. In addition, RSU show that, because of the strong counter-cyclical movements of markups, the deep-habit theory is capable of explaining increases in wages and consumption in response to a positive demand shock as is observed in the data. This latter empirical regularity has proved difficult to explain with standard models of the transmission of demand shocks.

In the deep-habit model it is of great consequence whether habits are internal or external. RSU show that the firm's pricing problem is time consistent under external habit persistence but time inconsistent under internal habit persistence.

See Also

▶ Consumption Externalities

Bibliography

Abel, A. 1990. Asset prices under habit formation and catching up with the Joneses. *American Economic Review* 80: 38–42.

Boldrin, M., L. Christiano, and J. Fisher. 2001. Habit persistence, asset returns, and the business cycle. *American Economic Review* 91: 149–166.

Campbell, J., and J. Cochrane. 1999. By force of habit: A consumption-based explanation of aggregate stock market behavior. *Journal of Political Economy* 107: 205–251.

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.

Constantinides, G. 1990. Habit formation: A resolution of the equity premium puzzle. *Journal of Political Economy* 98: 519–543.

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

Lettau, M., and H. Uhlig. 2000. Can habit formation be reconciled with business cycle facts? *Review of Economic Dynamics* 3: 79–99.

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

Pollak, R. 1970. Habit formation and dynamic demand functions. *Journal of Political Economy* 78: 745–763.

Ravn, M., S. Schmitt-Grohé, and M. Uribe. 2006. Deep habits. *Review of Economic Studies* 73: 195–218.

Smets, F., and R. Wouters. 2004. Shocks and frictions in US business cycles: A Bayesian DSGE approach. Mimeo, European Central Bank.

Sundaresan, S.M. 1989. Intertemporally dependent preferences and the volatility of consumption and wealth. *Review of Financial Studies* 2: 73–89.

Uribe, M. 2002. The price-consumption puzzle of currency pegs. *Journal of Monetary Economics* 49: 533–569.

Hadley, Arthur Twining (1856–1930)

Robert B. Ekelund Jr.

Keywords

American Economic Association; Cartels; Economic evolution; Exhaustible resources; Hadley, A. T.; Monopoly; Period analysis; Price discrimination; Profit maximization; Property rights; State capture; Transport markets

JEL Classifications

B31

American economist, educator and public servant, Hadley was educated at Yale and at the University of Berlin, where he studied under German historicists. In a remarkable career, Hadley was, in turn, a freelance writer and lecturer on railway economics, a professor of political economy at Yale (1891–9), president of the American Economic Association, president of Yale University (1899–1921), chairman of the Railroad Securities Commission providing the Hadley Report on Railway finances in 1911, and was widely sought after as a political candidate for high political office in the United States. An inveterate traveller, Hadley died aboard ship in Kobe harbour in 1930.

Hadley was an extremely prolific and eclectic writer, but the bulk of his important work in economics was completed before the turn of the 20th century. His reputation rests essentially on two works, *Railway Transportation* (1885) and a basic text, *Economics: An Account of the Relations between Private Property and Public Welfare* (1896), which received high praise from his friend and colleague, Irving Fisher.

In *Railway Transportation* Hadley revealed himself as the most creative railway economist of the day through an integration of sophisticated (certainly for the time) economic analysis with the problems of railway organization. Among other theoretical insights Hadley formalized a theory of monopoly and price discrimination; developed, in the mathematical terms of Cournot, a marginal rule for profit maximization; and anticipated the period analysis of Marshall's *Principles.* More importantly, perhaps, he developed a modern and complete theory of cartels, showing that, in the presence of open competition, such unsanctioned behaviour on the part of railroads, would lead to the benefits of competition without the attendant disadvantages. In another perspicacious insight Hadley correctly characterized railway regulation as resulting from the capture, by the industry, of legal sanctions to obtain rate stability. In the main, Hadley viewed regulation as representing a low-cost cartel enforcement device.

In *Economics* Hadley went further than Marshall by explicitly developing the interrelations between property rights, economic evolution and economic efficiency. Hadley utilized the real world examples of the fisheries and mining to demonstrate the impact of ill-defined property rights on depletable resources, emphasizing the necessity of altered systems to obtain optimal resource use and allocation. This contribution, along with his prophetic analyses of transport market structure, establishes Hadley as one of the most inventive pre–20th-century American economists.

Selected Works

1885. *Railway transportation: Its history and its laws.* New York.

1890. The prohibition of railroad pools. *Quarterly Journal of Economics* 4: 158–171.

1896. *Economics: An account of the relations between private property and public welfare.* New York.

Bibliography

Cross, M.L., and R.B. Ekelund Jr. 1980. A.T. Hadley on monopoly theory and railway regulation: An American contribution to economic analysis and policy. *History of Political Economy* 12: 214–233.

Cross, M.L., and R.B. Ekelund Jr. 1981. A.T. Hadley: The American invention of the economics of property rights and public goods. *Review of Social Economy* 39: 37–50.

Fisher, I. 1930. Obituary: Arthur Twining Hadley. *Economic Journal* 40: 526–533.

Locklin, D.P. 1933. The literature on railway rate theory. *Quarterly Journal of Economics* 47: 167–230.

Hagen, Everett Einar (1906–1993)

Robert R. Nathan

Keywords

Economic development; Keynesianism; New deal; Population growth; Technical change; Unemployment

JEL Classifications

B31

Hagen was born in Holloway, Minnesota. He graduated from St Olaf College (BA, 1927) and the University of Wisconsin (MA, 1932; Ph.D., 1941). After a short period at the University of Illinois (1948–51) he became professor of economics at the Massachusetts Institute of Technology (1953–72); from 1970 to 1972 he was Director of the Center for International Studies at MIT.

Since the Second World War, developing nations have received unprecedented attention from economists and large financial resources from the industrialized world. Dr. Hagen was an important contributor to analysing key problems and processes of economic development.

Before concentrating on economic development, Hagen served in the Bureau of the Budget as a close associate of Gerhard Colm in the application of Keynesian principles to US fiscal policies. His firm commitment to Keynes's concepts was a factor in his transfer to the MIT from the University of Illinois, where more traditionalist faculty and top officialdom were hostile to the views of Keynes and of the New Deal.

In his book *On the Theory of Social Change* (1962), Hagen correctly concluded that economics alone could not provide the theoretical or policy directions for economic development. He studied deeply the role of human behaviour based on studies of anthropologists, sociologists and political scientists. Hagen's multidisciplinary approach provided invaluable insights for formulating development plans and policies.

In his fourth edition of *The Economics of Development* (1986), Hagen continued to elaborate on theoretical aspects as well as policies and implementation processes essential for development progress. Hagen updates the most promising lessons from successful nations replicable in the lagging nations.

Hagen disputes the common view that high population growth rates are a major deterrent to development. He also documents the thesis that protectionism is helpful to the developing world. He sets forth a strong case for attributing considerable unemployment to technological change. These somewhat unorthodox views are persuasively articulated and documented.

Of major importance are Hagen's conceptual formulations, his analyses based on personal experiences, and his challenges to economists and members of other disciplines to work jointly to overcome the persistent barriers to significant progress in the lagging nations.

Selected Works

1962. *On the theory of social change*. Homewood: Dorsey Press.
1963. ed. *Planning economic development*. Homewood: Richard D. Irwin.
1968. *The economics of development*. Homewood: Richard D. Irwin. Revised, 1980, 1986.

'Hahn Problem'

F. H. Hahn

Harrod (1939), who inaugurated the postwar concern with growth theory, distinguished between three growth rates: the natural, the warranted and the actual. True to his Keynesian heritage he argued that there were circumstances in which the warranted rate of growth permanently exceeds the natural rate. More importantly from the point of view of this essay he claimed that the warranted growth path was highly unstable – he called it a 'knife-edge'. By this he meant that small disturbances of the warranted growth path would lead to a cumulative divergence of actual from warranted growth. The argument was simple. Suppose, for instance, that for some exogenous reason the actual growth rate fell a little below the warranted rate. By virtue of the accelerator mechanism, savings would exceed investment (exante) and income would be given a further impulse taking it below its warranted level. This leads to further reductions in investment and to further downward displacement of the actual path. This process continues. Hicks (1950) quickly saw that this theory could easily serve as an explanation of cycles.

Many economists, however, took the view that Harrod had underestimated the prevalence of stabilizers in a market economy. In particular his theory had little to say about the behaviour of relative prices and had ruled out substitution possibilities by assuming fixed coefficients of production. Not only did he thereby overdetermine the long run equilibrium system (the equation: natural rate = warranted rate had only exogenously given variables on both sides) but he allowed no scope to the price mechanism to stabilize the economy against small shocks. This argument found its clearest expression in a famous article by Solow (1956).

H

For a fuller discussion of Solow's work the reader should consult the entry on Neoclassical Growth Theory, here it is very briefly summarized. Let y = output per man and k = capital per man and let

$$y = f(k)$$

be the production function which is concave and has the property

$$f'(0) = \infty, \quad f'(k) > 0 \quad \text{all } k \in (0,\infty).$$

Let n be the rate of population growth and s the propensity to save. For an equilibrium, saving per man must equal investment per man, write it as i. But

$$i = \dot{k} + nk$$

so we require

$$\dot{k} + nk = sy. \tag{1}$$

In steady state $\dot{k} = 0$ and we must solve

$$nk = sf(k) \quad \text{or} \quad n = s\frac{f(k)}{k}$$

which is Harrod's equation. Given the assumptions on $f(k)$ there always exists k^* which solves the equation. This then answers one of Harrod's arguments to the effect that it may not be possible to bring the natural rate (n) into equality with the warranted rate [$sf(k)/k$].

Now divide both sides of (1) by k and rearrange to give

$$\frac{\dot{k}}{k} = \frac{sf(k)}{k} - n. \tag{2}$$

By the concavity of $f(k)$, $f(k)/k$ is a diminishing function of k. Hence starting at any $k(0) \neq k^*$ and following a path for which (a) employment grows at the rate n and (b) savings are always equal to investment (call this a 'warranted' path), the economy will be driven to the steady state k^*, (where $k = 0$). This was the gist of Solow's argument.

It will be noticed straight away that this argument has no bearing on Harrod's knife-edge claim. Harrod had not proposed that warranted paths diverge from the steady state but that actual paths did. The latter are neither characterized by a continual equality of ex ante investment and savings nor by continual equilibrium in the market for labour. Thus although Solow thought that he was controverting the knife-edge argument he had only succeeded in establishing the convergence of warranted paths to the steady state.

However, even here it was not at all clear how robust *that* conclusion was to a relaxation of some of its rather strong assumptions. In particular it was widely agreed that the aggregate production function in terms of an aggregate capital input was a 'fable' (Samuelson 1962). The question was whether this fable was instructive or misleading. An attempted answer which was closely related to the pioneering work on turnpikes by Dorfman et al. (1958) was christened the 'Hahn problem', although it was not really a problem nor was Hahn's analysis of startling novelty.

Before giving a precise account it will be helpful to have a bird's eye view.

Suppose that there are many different capital goods used in their own production as well as in the production of a single consumption good. Let $t = 0$ be the initial date at which we take the capital stock as determined by past history up to that date. (For simplicity capital goods are assumed to be infinitely durable.) Let agents have expectations concerning the change in relative prices between $t = 0$ and $t = 0 + \epsilon$. These expectations together with the technological conditions of production will determine investment in the various capital goods. This will have the property that everyone is, at the margin, indifferent between investing in one good rather than another. Once that has been determined the economy is, as it were, on rails from which it cannot deviate if we require expectations to be correct and production to be intertemporarily efficient. For the correctness of the price expectations for $t = 0 + \epsilon$ imply what prices must be in all subsequent time periods. However, the 'rails' which the economy gets onto depend on the arbitrarily postulated expectations at $t = 0$. There are in fact an infinity of such rails depending on initial expectations. Most of these, however, lead away from the steady state and not to it (in the example of Hahn 1966, all of them except one lead the economy away from the steady state). There are thus many warranted paths and they do not conform to the Solow proposition for the single capital good. There seems to be both indeterminacy and instability of the steady state under warranted paths deviations. However, it may be that the rails which lead the economy away from the steady state are also leading it into an abyss. That is, the paths may eventually become infeasible because some capital good needed in production has disappeared. However, if we postulate some form of myopia in expectations, by which is meant no more that that agents cannot predict prices into the infinitely distant future, there is nothing to prevent the economy following such errant warranted paths for a 'long time'. However, we return to this matter below after the technical discussion.

The story which has just been told informally exemplifies the difficulties which arise in an economy which does not have a full set of Arrow–Debreu markets. Such an economy must act on the basis of price expectations and these in turn open up the possibility of 'bootstrap' warranted paths: the economy evolves the way in which it does because expectations are what they are and not for any 'real' reason. In the conclusion we return to these intuitive explanations. But first we must demonstrate the existence of many warranted paths which do not seek the steady state.

Let there be m capital goods whose quantities *per man* are denoted by the vector $k = (k_1, \ldots, k_m)$ and let $y = (y_1, \ldots, y_m)$ be the output vector (per man) of the capital goods. The output of consumption good per man is written as y_0. Let p_0 be the price of the consumption good and $p = (p_1, \ldots, p_m)$ the price vector of capital goods. All prices are reckoned in unit of account. There are constant returns to scale and one defines

$$A(k) = \{(y, y_0) | F(y_0, y \cdot k) \geq 0\}$$

as the production possibility set of the economy given k. In this definition $F(\cdot)$ is assumed C^2, strictly concave function with the property:

$$\begin{aligned} \frac{\partial F}{\partial k_i} &> 0 && \text{for } k_i < \infty, \\ \frac{\partial F}{\partial k_i} &< +\infty, && \text{for } k_i = 0 \ \text{all}\, i. \end{aligned}$$

A competitive economy in equilibrium will at all dates behave as if it solved the problem:

$$\max_{a(k)} \ (p \cdot y + p_0 y_0).$$

Let R be this maximized sum. Then we can write

$$R = R(p_0, p, k).$$

Classical duality theory gives

$$R_i(p_0, p, k) = y, \quad i = 0, \ldots, \ m.$$

where $R_i = \partial B / \partial p_i$. Moreover we know that R is convex in (p_0, p) and concave in k. If we suppose that population is growing at the geometric rate n then the evolution of the capital stock per man is given by the differential equation

$$R_i(p_0, p, k) - nk_i = k_i, \quad i = 0, \ldots, \ m. \qquad (3)$$

But if the economy has perfect foresight so that the expected rate and actual rate of all price changes coincide then it must satisfy arbitrage equations which ensure that investment in all directions is equally profitable. If we let $R_{m+i} = \partial R / \partial k_i$ this means that there is at each date a scalar, r, such that

$$R_{m+i}(p_0, p, k) + \dot{p}_i = rp_i, i = 1, \ldots, m. \qquad (4)$$

Since we can choose one good as numeraire (say the consumption good) we need only one more equation to be able to trace the evolution of all variables from given initial conditions. That equation must refer to the common rate of return r. This will depend on the savings decisions of agents and on technology and so on (p_0, p, k, r). Write

$$\dot{r} = c(p_0, p, k)$$

In steady state: $\dot{r} = k = \dot{p} = 0$. Let r^*, p^*, k^* be the solution of (3), (4), (5) in such a steady state. (On present assumptions such a solution exists.) To study the warranted growth path of the economy near the steady state we take a first order Taylor expansion of these three equations at (r^* p^*, k^*). We write: $\tilde{p} = p - p^*$, $\overline{k} = k^*$, $\overline{r} = r - r^*$ and set P^*_0 identically equal to unity. Also

$$R_{ij} = \frac{\partial R_i(p^*, k^*)}{\partial P_j}, \quad R_{im+j} = \frac{\partial R_i(p^*, k^*)}{\partial P_j} \ \text{etc.}$$

and

$$c_r = \frac{\partial c(r^*, p^*, k^*)}{\partial r}.$$

We obtain

$$\sum_j R_{im+j} \overline{k}_j - n\overline{k}_j + \sum_j R_{ij} \overline{p}_j = \dot{\overline{k}}_i, \quad i = 1, \ldots, n \qquad (6)$$

$$\overline{r} p_i^* \sum_j R_{m+im+j} \overline{k}_j - j\overline{k}_j - \sum_j R_{m+ij} \overline{p}_j + r^* \overline{p}_i = \dot{\overline{p}}_i, \quad i = 1, \ldots, n \qquad (7)$$

$$\overline{r} c_r + \sum_j c_{m+j} \overline{k}_j + \sum_j c_p \overline{p}_j = \dot{\overline{r}} \qquad (8)$$

Let R_{pk} be the $n \times n$ matrix of elements $[R_{im+j}]$, R_{pp} then $n \times n$ matrix of elements $[R_{ij}]$, R_{kk} the $n \times n$ matrix of elements $[R_{m+j,\ m+j}]$ and I the $n \times n$ identify matrix. Then the above equations can be written compactly as

$$\begin{bmatrix} c_r & \{c_{m+j}\} & \{c_j\} \\ \{0\} & R_{pk} - nI & R_{pp} \\ p^* & -R_{kk} & -R_{kp} + r^*H \end{bmatrix} \begin{bmatrix} \bar{r} \\ \bar{k} \\ \bar{p} \end{bmatrix} = \begin{bmatrix} \dot{\bar{r}} \\ \dot{\bar{k}} \\ \dot{\bar{p}} \end{bmatrix} \tag{9}$$

Note that $R_{pk} = R'_{kp}$. Let us consider the unbordered matrix:

$$A = \begin{bmatrix} R_{pk} - nI & R_{pp} \\ -R_{kk} & -R_{kp} + p^*I \end{bmatrix}$$

If we make the assumption that profits are all saved and wages are all spent then $r^* = n$. Make this assumption: Let

$$T\begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix}$$

so that $T' = -T$. Then and B is a symmetric matrix. Now let $Ax = \lambda x$ be the characteristic equation for A with eigenvalue λ. Then

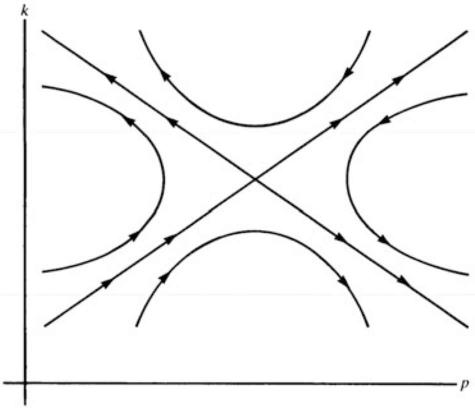
$$\lambda Tx = Bx = B'x \tag{10}$$

But $TA = A'T' = B'$ and $A'T' = A'\ (-\ T)$. Let $Tx = y$. Then from (10)

$$\lambda y = -A'y \quad \text{or} \quad -\lambda y = -A'y$$

Hence if λ is a root of A so is – λ. One says that A has the saddle point property. The phase diagram for p and k in two dimensions is given in Fig. 1.

If $\dot{r}$ remained constant at its steady state value r* then Figure I would show all the warranted paths of the economy. It will be seen that only one of these approaches the steady state. On the other hand all the other paths may eventually become infeasible – they lead to one of the axes. Infinite perfect foresight would rule all these paths out of consideration. However, the postulate of such foresight seems farfetched.



'Hahn Problem', Fig. 1

When the whole system of equation (9) is considered matters are more complicated. One way out of the complication is to suppose that the economy behaves as if it were solving an infinite 'Ramsey problem'. The behaviour of r would then be fully determined by the Euler–Lagrange equations for this problem. But once again, in the absence of discounting, all paths but the convergent one would be ruled out and the 'Hahn problem' would disappear. But also once again the realism and relevance of such a postulate must be in doubt (see Hahn 1968; Kurz 1968).

The alternative is to proceed by way of a model of overlapping generations or simply by a descriptive savings function. Work along these lines (and also with more than one consumption good), has been undertaken by a number of economists. Shell and Cass (1976) have provided a good general treatment of systems such as (9). The main conclusion is that in addition to the divergent paths there is also a manifold of paths which converge. This is interesting since now even with infinite perfect foresight and convergence, there is nothing to tell us which of the convergent paths the economy will choose. The same difficulty has been encountered in the overlapping generations' literature which has been very ably summarized by Woodford (1984). However in both approaches divergent warranted paths remain and some of these (in the case of overlapping generations) are viable over infinite time.

There are a number of technical matters which have not been considered in the above account – for instance the relation of the problem to turnpike theory (Samuelson 1960) and the role of intertemporal efficiency conditions which have been encapsulated in the dual formulation here adopted. But enough has been provided to allow an intditive summary.

The requirement of perfect foresight equilibrium is certainly too weak to determine the path of an economy if perfect foresight is not over the infinite future. Moreover there are then many paths which do not converge to the steady state. This is due to the circumstances that arbitrage, given initial expectations, imposes a particular path on the economy if expectations are to be fulfilled in the future, and if the arbitrage equations are to hold. Thus the invisible hand may for a long time provide coherence in the economy even while it is guiding it to eventual disaster. But even when there is perfect foresight over the infinite future and that future is discounted, there will be many paths that do not converge to the steady state. (Kurz 1968, gives a case where (9) has been converted into a Ramsey problem with discounting and where all paths diverge from the steady state.) It would seem that in general the price-mechanism even with correct expectations will not bear out the rather optimistic conclusion of Solow with which this essay started.

Two matters remain to be mentioned. Warranted paths which do not converge may yet be Pareto-efficient (see Cass 1972), provided of course they are feasible over infinite time. However this does not mean that such paths do not provide an occasion for policy since they may be associated with very undesirable inter-temporal distributions of welfare between generations. (This applies to models in which agents are not infinitely lived and in which agents do not value their descendants' utility as they do their own.)

The second matter is this: one may ask whether the steady state would not be stable if one allowed for false expectations, that is if one considered actual and not warranted paths. This question was posed by Shell and Stiglitz (1967) and is also discussed in Hahn (1969). Although Shell and Stiglitz did indeed find that with relatively inelastic expectations the steady state was stable their model was very special, particularly in the manner in which it incorporates the heterogeneity of capital goods. In a more general model Hahn (1969) found no general presumption that the steady state was stable unless expectations were completely inelastic, as was postulated by Morishima (1964). In that latter case there are no expected capital gains and losses and the arbitrage equation takes on a degenerate form. Nonetheless it remains an interesting question which set of circumstances leads to false expectations being stabilizing. No general answers are now available. But the 'Hahn problem' was concerned with correct (albeit myopic) expectations.

See Also

- ▶ Hamiltonians
- ▶ Sunspot Equilibrium
- ▶ Tulipmania
- ▶ Turnpike Theory

Bibliography

Cass, D. 1972. On capital overaccumulation in the aggregative, neoclassical model of economic growth. *Journal of Economy Theory* 4: 200–223.

Dorfman, R., P.A. Samuelson, and R.M. Solow. 1958. *Linear programming and economic analysis*. New York: McGraw-Hill.

Hahn, F.H. 1966. Equilibrium dynamics with heterogeneous capital goods. *Quarterly Journal of Economics* 80: 633–645.

Hahn, F.H. 1968. On warranted growth paths. *Review of Economic Studies* 35: 175–184.

Hahn, F.H. 1969. On some equilibrium paths. In *Models of economic growth*, ed. J. Mirrlees and N.H. Stern. London: Macmillan, 1973.

Harrod, R.F. 1939. An essay in dynamic theory. *Economic Journal* 49: 14–33.

Hicks, J.R. 1950. *A contribution to the theory of the trade cycle*. Oxford: Clarendon Press.

Kurz, M. 1968. The general instability of a class of competitive growth processes. *Review of Economic Studies* 35: 155–174.

Morishima, M. 1964. *Equilibrium, stability and growth*. Oxford: Clarendon Press.

Samuelson, P.A. 1960. Efficient paths of capital accumulation in terms of the calculus of variations. In *Stanford symposium on mathematical methods in the social sciences*, ed. K.J. Arrow, S. Karlin, and P. Suppes. Stanford: Stanford University Press.

Samuelson, P.A. 1962. Parable and realism in capital theory: The surrogate production function. *Review of Economic Studies* 29: 193–206.
Shell, K., and D. Cass. 1976. The structure and stability of competitive dynamical systems. *Journal of Economic Theory* 12: 31–70.
Shell, K., and J. Stiglitz. 1967. The allocation of investment in a dynamic economy. *Quarterly Journal of Economics* 81: 592–610.
Solow, R.M. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70: 65–94.
Woodford, M. 1984. *Indeterminancy of equilibrium in the overlapping generation model: A survey.* Mimeo: Columbia University.

Halévy, Elie (1870–1937)

M. Donnelly

Elie Halévy was one of the foremost historians of 19th-century English thought and politics. He was born at Etretat, France, and educated in Paris at the Lycée Condorcet and the Ecole Normale. His early training was philosophical, and he remained throughout his life associated with the *Revue de métaphysique et de morale.* He passed his *agrégation* in 1892, and was invited to lecture at the Ecole des Sciences Politiques on the evolution of political ideas in England; this was to establish the course of his career. In 1900–1903 he published his first major work. *La formation du radicalisme philosophique en Angleterre*, a study tracing the development of the utilitarian doctrine from 1776 to 1832. As an offshoot of this project he also published a short study, *Thomas Hodgskin* (1903), which presents Hodgskin as a precursor of Marx. Halévy's major historical writings were the volumes of his *Histoire du peuple anglais au XIXe siècle* (1912–1932), most notably vol. I, *England in 1815,* and vol. V, *Imperialism and the Rise of Labour.*

La formation du radicalisme philosophique is, among other things, a signal contribution to the history of economic thought. Halévy's subject is less utilitarianism in general than the application of utilitarian principles to criticize the established order and to justify grand proposals of reform: in sum. Philosophic Radicalism, or what Bentham referred to as the exposure of 'political fallacies'. The book offers a detailed exposition, at once historical and analytical, of works by Bentham and James Mill, and to a somewhat lesser extent the classical economists (Smith, Malthus, Ricardo). Halévy is at pains to demonstrate the connection between utilitarianism as a moral and political doctrine, and classical political economy. Indeed he summarizes his argument in the formula. 'The morality of the Utilitarians is their economic psychology put into the imperative' (Halévy 1928, p. 478), a formula which nicely captures utilitarianism's debt to economics as well as the ambiguities inherent in the doctrine.

Bentham held that only the principle of utility – the principle of promoting the greatest happiness for the greatest number – can offer a satisfactory criterion for evaluating action. Not only is this principle commonly acceptable to reasonable men and women, it is moreover grounded in and reinforced by human psychology. Human beings are creatures who cannot but pursue pleasure and avoid pain. The difficulty in the argument arises in the comparison and summing up of individual pleasures. Is the greatest happiness of the greatest number simply a summation of individual happinesses egoistically pursued? Or does it require that an individual's pursuit of his private pleasure coincide with a pursuit of the greatest happiness of the greatest number? For his own part Bentham was ambiguous on this point: on the one hand, he acknowledged the (potential and actual) conflict between private interests and the public interest, and hence the need for molding or transforming human nature. This is the sphere of the 'artificial identification of interests', where as Halévy puts it, 'the science of the legislator must intervene to identify interests which are naturally divergent' (p. 508). On the other hand, Bentham argued, more optimistically, that there is social order 'realised spontaneously, by the harmony of egoisms' (p. 508). This is the part of the argument utilitarianism shares most closely with, or borrows from, classical economics. It provides the climax of Halévy's history:

> insensibly, the progress of the new political economy had determined the preponderance within [Utilitarianism] of another principle, the principle according to which egoisms harmonise of themselves in a society which is in conformity to nature. From this new point of view, the fundamental moral notion for the theorists of Utilitarianism is no longer that of obligation, but that of exchange . . ..The Utilitarian moralist dispenses the legislator from intervening just in so far as, by his advice and by his example, he tends, in conformity with the hypothesis of the political economists, to realise in society the harmony of egoisms. (p. 478)

In the event, as Halévy shows in conclusion, this synthesis was precarious. The harmony of egoisms was too tenuous a factual basis for utilitarian morality, and the ambiguities of Philosophic Radicalism were supplanted by new and simplified versions of utilitarianism, like the 'Manchester philosophy'.

Selected Works

1912–1932. *History of the English people in the nineteenth century*, 6 vols. Trans. E.I. Watkin and D.A. Barker. London: Benn, 1924–1949.

1928. *The growth of philosophic radicalism.* Trans. Mary Morris. London: Faber.

Hamilton, Alexander (1755–1804)

Henry W. Spiegel

Keywords

Credit; Hamilton, A.; Implied powers; Jefferson, T.; Protection; Public credit; Public debt; Subsidies

JEL Classifications

B31

One of the founding fathers of the United States and Secretary of the Treasury in President Washington's cabinet, in which Thomas Jefferson served as Secretary of State. The two great men differed widely in their views about the destiny of the young nation. Jefferson wanted to preserve the position of the states and assign to the national government not much more than authority over foreign affairs. Hamilton favoured a strong and active central government. Jefferson was eager to preserve the rural economy in which he had grown up in Virginia. Hamilton proposed to promote economic development, especially manufacture, and vest in the national government the function of actively fostering such development. Jefferson took a dim view of public debts, paper money and financial institutions. Hamilton favoured them all. Jefferson was more of an egalitarian and had greater faith in the common man than Hamilton, who placed his trust in an alliance of government and the aristocracy of wealth: neither could flourish without the support of the other. Hamilton died in a duel with a political adversary during Jefferson's presidency, but his ideas were strong enough to survive him. The exigencies of the time caused Jefferson himself to adopt a number of Hamiltonian policies.

Thus Hamilton became the architect of what in *The Federalist* (1787, No. XI) he had called 'the great American system', later to be buttressed by such economic writers as the Careys, Daniel Raymond and Frederick List, and by Henry Clay in politics. He set forth his economic ideas in a series of state papers, published under his name when serving as Secretary to the Treasury. These papers are the first and second *Report on the Public Credit* (1790a; 1795), the *Report on a National Bank* (1790b), and the *Report on Manufactures* (1791). The state papers are justly famous, not as repositories of economic analysis, but as a masterly presentation of a case of which Hamilton, who had been trained in the law, was an eloquent advocate.

The apotheosis of credit found in Hamilton's reports refers to public and private credit as well. According to Hamilton, credit is a substitute of capital almost as useful as gold and silver. It has, and Hamilton has no doubt about this, a tendency to lower the interest rate. If the public credit is in a bad state, it can only have deleterious effects on private credit. The preservation of a healthy credit system is thus an important task. Foreign creditors should enjoy the same protection as domestic ones. Domestic holders of the public debt should be protected against the imposition of taxes on the public funds, as foreign creditors should be protected from repudiation or expropriation.

With the help of the public debt it will be possible to promote the economic development of the country. Scrupulous attention must be paid to the rights of the creditors, both for the sake of public expediency and as a moral obligation. The public debt of the United States should be funded, that is, arrangements should be made for the service of the debt by putting aside funds for the payment of interest and principal. A funded debt has great benefits. It will facilitate the use of instruments of debt as money and bring about lower interest rates, and will result in an increase in land values, which have declined in consequence of the scarcity of money.

Hamilton also proposed that the Union assume responsibility for the debts of the states, and that the funding of the debt should be financed in part from new duties on imported spirits and taxes on domestic ones and on stills. These proposals met considerable opposition because of the windfall gains that would accrue to speculators who had purchased instruments of the debt at low prices. To obtain Jefferson's support for this measure, Hamilton had to agree that the future capital of the nation would be located in the South, that is, in what is now Washington, DC.

The national bank, which Hamilton proposed to establish, was designed to aid in the expansion of the money supply, thereby facilitating the payment of taxes, the reduction of interest rates, the fulfilment of public functions, and the development of the national economy. The bank was to be under private rather than public direction, with the government playing the role of a minor shareholder. When the question was raised whether the Constitution granted the federal government the authority to establish a bank, Hamilton resolved it by referring to 'implied powers', that is, the power to employ suitable means to pursue constitutional ends. This solution was to have far-reaching consequences for the future development of constitutional law.

The *Report on Manufactures* goes into considerable detail examining the relative merits of agriculture and industry. Hamilton underlines the merits of both and the benefits which each derives from the other. He stresses that both are productive, a point that had to be made, and made forcefully, in view of the teachings of the Physiocrats. Hamilton demonstrates great ingenuity in enumerating the factors that are responsible for favourable effects of industrial development on the national income. Among these factors he mentions the division of labour, the more extensive use of machinery, the utilization of manpower that is not suited for agricultural pursuits, the promotion of immigration, the widening of opportunities for the exercise of entrepreneurial talent, and the strengthening of demand for agricultural products.

As far as international trade is concerned, Hamilton holds that the benefits from free trade are more imaginary than real because of the obstacles which foreign countries place in the way of United States exports. Moreover, foreign governments support domestic industries in various ways, and the United States should adopt similar policies by imposing protective and prohibitive duties, granting subsidies to domestic industries, and promoting internal improvements that facilitate the flow of commerce. Subsidies are liable to be abused, but their advantages outweigh the disadvantages. Lastly, Hamilton proposes that a board be established to promote economic development by bringing in skilled workers from abroad, rewarding useful improvements and inventions, paying premiums to importers of machinery, and similar means.

Selected Works

1790a. *Report on the public credit.* Washington, DC: Government Printing Office, 1908.

1790b. *Report on a national bank.* Washington, DC: Government Printing Office, 1908.

1791. *Report on manufactures.* Washington, DC: Government Printing Office, 1913.

1795. *Report on the public credit.*

1934. *Papers on public credit, commerce and finance*, ed. S. McKee, Jr. New York: Columbia University Press.

Bibliography

Dorfman, J. 1946. *The economic mind in American civilization 1606–1865*. Vol. 1. New York: Viking.

Spiegel, H.W. 1960. *The rise of American economic thought*. Philadelphia: Chilton.

Bibliographic Addendum

Chernow, R. 2004. *Alexander Hamilton.* New York: Penguin, has established itself as a standard biography. F. McDonald, *Alexander Hamilton,* New York: W. W. North, 1979, is a spirited defence of Hamilton's vision for American development. See also E. J. Ferguson, *The Power of the Purse: A History of American Public Finance*. Chapel Hill: University of North Carolina Press, 1961.

Hamilton, Earl Jefferson (Born 1899)

Donald N. McCloskey

Hamilton was born on 17 May 1899 in Houlka, Mississippi. After graduating from the University of Mississippi (and coaching football and playing minor league baseball), he received in 1929 his doctorate in economics at Harvard. He taught at Duke, Northwestern, and finally for twenty years at the University of Chicago, during the banishment and eventual rehabilitation of the quantity theory of money.

Though he has worked on several topics in the early history of the Atlantic economy, his main contribution to economic science is the documentation of the dependence of the price level on the quantity of precious metals, 1351–1800. Spain through its centuries of prosperity and decline was his field of study, and he is accounted a major historian of that country (hon. Dr University of Madrid, 1967). Involved nearly from its beginnings in the 1920s with the International Committee on Price History, Hamilton constructed indexes of prices, wages and money from primary sources. The historical weight and economic ingenuity of his volumes, 1934, 1936, 1947, made them central to the modern quantity theory. Various attempts to revise his history of prices (attaching it to population, for instance) have had difficulties with the sheer mass of evidence that Hamilton accumulated, Kepler-like. Hamilton, further, is prominent in the thin, bright stream of historical economists *avant la lettre*. His combination of economic and historical erudition is a model of cliometrics, exhibited best in his lucid reply to his revisers (1960).

Selected Works

1934. *American treasure and the price revolution in Spain, 1501–1650*. Cambridge, MA: Harvard University Press.

1936. *Money, prices, and wages in Valencia, Aragon, and Navarre, 1351–1500*. Cambridge, MA: Harvard University Press.

1947. *War and prices in Spain, 1651–1800*. Cambridge, MA: Harvard University Press.

1960. The history of prices before 1750. In *XI^{e} Congrès International des Sciences Historiques*, Stockholm.

1968. John Law. In *International encyclopedia of the social sciences*, vol. 9, New York: Macmillan.

Hamiltonians

Karl Shell

Abstract

Hamiltonian dynamics arises not only in economic optimization problems but also in descriptive economic models in which there is perfect foresight about asset prices. Hamiltonian dynamics applies in discrete time as well as in continuous time. In discrete time, the system of differential equations is replaced by a closely related system of difference equations. The theory accommodates differential correspondences or difference correspondences, which naturally arise in economics. The Hamiltonian approach through the

Hamiltonian function has proved remarkably successful in establishing sufficient conditions for the saddle-point property and related stability questions in a class of optimal economic growth models.

Keywords

Continuous and discrete time models; Duality; Hamilton, W. R.; Hamiltonian dynamical system; Hamiltonian function; Hamiltonians; Lyapunov functions; Optimal-growth theory; Overlapping generations models; Poincaré, J. H.; Pontryagin's maximum principle; Saddle-point property; Transversality conditions

JEL Classifications

C6

The laws of motion for a perfect-foresight economy, whether centrally planned or competitive, can be described by a Hamiltonian dynamical system or by a simple perturbation thereof. The Hamiltonian dynamical system and the Hamiltonian function which generates it are named for their inventor, the great Irish mathematician William Rowan Hamilton (1805–1865).

Hamilton's differential equations serve as the basic mathematical tool of classical particle mechanics (including celestial mechanics). Let $x(t) = (x_1(t) , ... , x_i(t) , ... , x_m(t)$ and $y(t) = (y_1(t) , ... , y_i(t) , y_m(t)$ be m-vectors dependent on time t. Let H be a continuous, differentiable function of x, y, and t, H: $R^m \times R^m \times R \rightarrow R$. Think of H as the Hamilton's function (HF) which generates Hamilton's differential equations,

$$\mathrm{d}x_i(t)/\mathrm{d}t = \partial H(x(t), y(t), t)/\partial y_i(t)$$

and

$$\mathrm{d}y_i(t)/\mathrm{d}t = \partial H(x(t), y(t), t)/\partial x_i(t)$$

for $i = 1,\ldots,m$. If the Hamiltonian function H depends on time only through the variables $x(t)$ and $y(t)$, i.e., $\partial H/\partial t \equiv 0$, then the corresponding Hamiltonian dynamical system (HDS) is said to be *autonomous.* These differential equations are frequently interpreted in physics as solutions to some extremization problem. In mechanics for example, HDS is implied by the principle of least action. Since economic planning and many other economic problems involve maximization or minimization over time, it is unsurprising that the Hamiltonian formalism has substantial application in economics. Its appeal to economists goes much further than this. There is a duality (conjugacy, in the language of mechanics) between $x_i(t)$ and $y_i(t)$ which allows us to interpret one as a (primal) economic flow and the other as a (dual) economic price. Given this point of view, the Hamiltonian function (HF) itself has important economic interpretations. Hamiltonian dynamics not only arises in economic optimization problems but it also arises in descriptive economic models in which there is perfect foresight about asset prices. Hamiltonian dynamics applies in discrete time as well as in continuous time. In discrete time, the system of differential equations is replaced by a closely related system of difference equations. The right side of the equations describing Hamilton's law of motion need not be single-valued. The theory accommodates differential correspondences or difference correspondences, which naturally arise in economics.

Consider first the application of Hamiltonian approach to the theory of economic growth; see, for example, the Cass-Shell (1976a) volume. A large class of economic growth models can be described by simple laws of motion based on the instantaneous production set T, with feasible production satisfying.

$$(c, z, -k, -l) \in T \subset \{(c, z, -k, -l) | (c, k, l) \geq 0\},$$

Where c denotes the vector of consumption-goods outputs, z the vector of net investment-goods outputs, k the vector of capital-goods inputs, and l the vector of primary-goods inputs. There is an equivalent representation of static technological opportunities that is better suited to dynamic analysis: the representation of the static technology by its Hamiltonian function H.

Let p be the vector of consumption-goods prices and q be the vector of investment-goods prices. Define the Hamiltonian function $H(p, q, k, l)$ by

$$H(p,q,k,l) = \max_{(c',z')} \left\{pc' + qz' \middle| \left(c', z', -k, -l\right) \in T\right\},$$

H is defined on the non-negative orthant and can be interpreted as the maximized value of net national product at the output prices (p, q) given input endowments (k, l).

Obviously, if we know the set T, then we know precisely the function H. If T is closed, convex, and permits free disposal, then H is continuous, convex and homogeneous of degree one in the output prices (p, q), and concave in the input stocks (k, l). If H is a function of (p, q, k, l) which is continuous, convex and homogeneous of degree one in (p, q), and concave in (k, l), then H corresponds to a unique T among closed, convex technologies permitting free disposal. In many dynamic applications, it is only the H representation which matters. Relax, for example, the free-disposal assumptions on T. For a given function H, the set T might be unique, but the dynamics would be independent of the particular set T which generated the function H. Relax, as another example, the assumption that T is convex. Given an H which is convex in (p, q), and concave in (k, l), the set T will not be unique, but the continuous dynamics (HDS) will not be altered in an essential way.

Representation of the static technology by the Hamiltonian function permits one to describe the economic laws of motion as a Hamiltonian dynamical system. In continuous time, the motion is described by

$$\dot{K}(t) \in \partial H(p(t), q(t), k(t), l(t))/\partial q(t)$$

(HDS)

$$\dot{q}(t) \in -\partial H(p(t), q(t), k(t), l(t))/\partial k(t)$$

where $\dot{k}(t)$ and $\dot{q}(t)$ are vectors of time derivatives and $(\partial H/\partial q)$ and $(\partial H/\partial k)$ are gradients (derivatives when H is differentiable). The first line of (HDS) is immediate from the definition of net investment since it reduces to $\dot{k}(t) = z(t)$, where $z(t)$ is the vector of net investment. The second line is an equal-asset-return condition which reduces to $\dot{q}(t) + r(t) = 0$, where $r(t)$ is the dual vector of shadow rental rates.

For discrete time, the Hamiltonian dynamical system is

$$k_{t+1} \in k_t + \partial H(p_t, q_t, k_t, l_t)/\partial q_t$$

(HDS)′

$$q_{t+1} \in q_t - \partial H\left(p_{t+1}, q_{t+1}, k_{t+1}, l_{t+1}\right)/\partial k_{t+1}.$$

Line 1 is equivalent to $k_{t+1} = k_t + z_t$ and line 2 is equivalent to $q_{t+1} - q_t - r_{t+1} = 0$, where z_t is the time $r(t)$ gross investment vector and r_{t+1} is the dual vector of shadow capital-goods rental rates in period $(t + 1)$.

For openers, let us analyse the case where H is autonomous. This occurs if $p(t) = \overline{p}$ and $l(t) = \overline{l}$ for (HDS) or $p_t = \overline{p}$ and $l_t = \overline{l}$ for (HDS)′. Let (q^*, k^*) be a rest point to (HDS) or (HDS)′. Hence, we have

$$0 \in \partial H\left(\overline{p}, q^*, k^*, \overline{l}\right)/\partial q, 0 \in \partial H\left(\overline{p}, q^*, k^*, \overline{l}\right)/\partial k.$$

Consider the linear approximations about (q^*, k^*) of (HDS) and (HDS)′ (taken, for example, as if H were quadratic). Study the characteristic roots to the linearized systems. A simple but remarkable theorem due to Poincaré tells us that if λ is a root for the linearized, *autonomous* version of (HDS) then so is − λ. For the linearized, *autonomous* version of (HDS)′, we have if λ is a root, then so also is 1/λ. If for (HDS), we could rule out pure imaginaries (Re $\lambda \neq 0$), then we would have: The dimension of the manifold in (q, k) – space of solutions tending to (q^*, k^*) as $t \to \infty$ is equal to the dimension of the manifold of solutions tending to (q^*, k^*) as $t \to -\infty$ This is the *saddle-point property,* where the manifold of forward solutions and the manifold of backward solutions each have dimension equal to half the total dimension of the space. Similarly, we would have the saddle-point property for (HDS)′, if the modulus |λ| is unequal to unity.

Poincaré's result nearly gives us the saddle-point property. In the autonomous cases, the saddle-point property can be assured if the geometry of the Hamilton function is correct. We need to add very little to the convexity–concavity assumption (see Cass and Shell 1976b and Rockafellar 1976). Strict convexity in q and strict concavity in k will do the trick. So will a weaker uniform Hamiltonian steepness condition, which reduces to a value-loss condition; see, for example, McKenzie (1968) and Cass-Shell (1976b).

What about non-autonomous systems, such as optimal economic growth with the constant, positive discount rate ρ? Here $c(t)$ or c_t, is a scalar called felicity and usually denoted in optimal-growth problems by $u(t)$ or u_t. In this case, present prices must satisfy

$$-\dot{p}(t)/p(t) = \rho$$

or

$$-(p_t - p_{t-1})/p_t = \rho.$$

For simplicity, allow only for a single fixed factor and adopt the convention $l(t) = 1$, or $l_t = 1$.

It is natural then to re-express the systems (HDS) and (HDS)′ in terms of current prices $Q \equiv q/p$, rather than in terms of present prices q. We then have

$$\dot{k} \in \partial H(Q, k)/\partial Q$$

(PHDS)

$$\dot{Q} \in -\partial H(Q, k)/\partial k + \rho Q$$

and

$$k_{t+1} \in k_t + \partial H\left(Q_t, k_t\right)/\partial Q_t$$

(PHDS)′

$$Q_{t+1} \in Q_t - \partial H\left(Q_{t+1}, k_{t+1}\right)\partial k_{t+1} + \rho Q_t.$$

The systems (PHDS) and (PHDS)′ are *perturbed* Hamiltonian dynamical systems.

We no longer have Poincare's root-splitting theorems in pure form: the roots split but not about 0 for (HDS) nor 1 for (HDS)′. The trick here is to strengthen the geometry of H to give a saddle-point property or something like it.

This is the basics of the approach taken by Cass and Shell (1976b), Rockafellar (1976) and Brock and Scheinkman (1976). Conditions are found on H which assure that either (PHDS) or (PHDS)′ along with transversality conditions defines a globally stable system. It suffices to strengthen the convexity-concavity of H by an amount dependent on p or (weaker) to strengthen the steepness of H by an amount dependent on p. (The Lyapunov function which does the trick is $V = (Q - Q^*)\ (k - k^*)$ in the continuous-time model.)

The Hamiltonian approach *through the Hamiltonian function* has proved remarkably successful in establishing sufficient conditions for the saddle-point property and related stability questions in a class of optimal economic growth models. The parallel programme of using the Hamiltonian formalism in optimal- growth theory to yield sufficient conditions for cycling or other dynamic configurations has not yet been pursued in a systematic fashion but should prove equally successful when applied. The success of the Hamiltonian approach in decentralized and descriptive growth theory has so far been very limited; see Cass and Shell (1976b, Section 4). This has been disappointing. I still hope to see the Hamiltonian approach playing a pivotal technical role in, say, the dynamical analysis of overlapping-generations models, but there has not been much tangible encouragement for this hope.

Many of us first met Hamiltonian dynamical systems as necessary conditions for intertemporal maximization in the form of Pontryagin's maximum principle; see Pontryagin et al. (1962). See Shell (1967) for applications to economics and references.

Samuelson and Solow (1956) were probably the first in economics to mention the Hamiltonian formalism. For some of the history of Hamiltonian dynamics, in economic, mathematics, and physics, and for some of the classical references, see Magill (1970).

Bibliography

Brock, W.A., and J.A. Scheinkman. 1976. Global asymptotic stability of optimal economic systems with applications to the theory of economic growth. In Cass and Shell (1976a).

Cass, D., and K. Shell. 1976a. *The Hamiltonian approach to dynamic economics*. New York: Academic Press. [Reprinted from the *Journal of Economic Theory* 12, February 1976, Symposium: 'Hamiltonian dynamics in economics'.]

Cass, D., and K. Shell. 1976b. The structure and stability of competitive dynamical systems. In Cass and Shell (1976a).

Magill, M.J.P. 1970. *On a general economic theory of motion*. Berlin: Springer-Verlag.

McKenzie, L.W. 1968. Accumulation programs of maximum utility and the von Neumann facet. In *Value, capital and growth*, ed. J.N. Wolfe. Edinburgh: Edinburgh University Press.

Pontryagin, L.S., G. Boltyanskii, R.V. Gamkrelidze, and E.F. Mischenko. 1962. *The mathematical theory of optimal processes*. New York: Interscience.

Rockafellar, R.T. 1976. Saddlepoints of Hamiltonian systems in convex Lagrange problems having nonzero discount rate. In Cass and Shell (1976a).

Samuelson, P.A., and R.M. Solow. 1956. A complete model involving heterogeneous capital goods. *Quarterly Journal of Economics* 70: 537–562.

Shell, K., ed. 1967. *Essays on the theory of optimal economic growth*. Cambridge, MA: MIT Press.

Shell, K. 1969. Applications of Pontryagin's maximum principle to economics. In *Mathematical systems theory and economics*, ed. H.W. Kuhn and G.P. Szegö, vol. 1. Berlin: Springer-Verlag.

Hammarskjöld, Dag (1905–1961)

Björn Hansson

Hammarskjöld was born in Jönköping, Sweden, and died in an aircrash near Ndola in Zambia. He came from a family (which was knighted early in the 17th century) with a long tradition of public service; his father Hjalmar Hammarskjöld was Prime Minister of Sweden in 1914–17. Hammarskjöld's achievements were many and varied: a PhD at the University of Stockholm, 1933; Decent at the University of Stockholm, 1933–41; Under Secretary of the Ministry of Finance, 1936–45; Chairman of the Board of Governors of the Bank of Sweden, 1941–8; Under Secretary of the Foreign Office, 1946–51; Vice-chairman of the executive committee of Organization for European Economic Cooperation, 1948–9; an expert non-party member of the Swedish Cabinet as Minister without portfolio in charge of foreign economic relations, 1951–3; Vice-chairman of Sweden's delegation to the UN General Assembly in 1952 and chairman in 1953; elected Secretary General of the UN for 1953–8 and re-elected for a five-year term in 1958; Fellow of the Swedish Academy of Letters, 1954; posthumously awarded the Nobel Peace Prize, 1961.

While Hammarskjöld is mainly known for his outstanding career as a civil servant and international statesman, he also made a contribution to economics. After his MA in 1928 he became secretary to the Royal Commission on Employment (1927–35). Hammarskjöld's dissertation (1933) was published as a report to the committee and he also wrote the theoretical introduction to its final report. The aim of the dissertation was to show the determinants of the price level of consumer goods for an expired period, which implies an ex post analysis. Hammarskjöld went beyond the earlier formulas of Keynes and Lindahl, since his construction had profit as the mechanism by which given changes in prices and purchasing power are transmitted to the next period, i.e. a form of disequilibrium process. In this context, he was the first among Swedish economists actually to define the length of a period, namely, the duration of time for which plans are unchanged determines the length of the unit period. However, his work was not particularly influential among his colleagues, since his exposition was extremely complicated. During his career as a civil servant he published very little in the economic field.

See Also

▶ Stockholm School

Selected Works

1932. Utkast till en algebraisk metod för dynamisk prisanalys. (A sketch of an algebraic

method for a dynamic analysis of prices.). *Ekonomisk Tidskrift* 34: 157–176.

1933. *Konjunturspridningen. En teoretisk och historisk undersökning*. (The propagation of business cycles. A theoretical and historical investigation.). Stockholm: P.A. Norstedt.

Hammond, John Lawrence le Breton (1872–1949) and Lucy Barbara (1873–1961)

Peter Clarke

Lawrence Hammond was born in Yorkshire in 1872. He married Barbara Bradby in 1901; they had no children. Both the Hammonds received a classical education, which they drew on in their literary work. At Oxford, Lawrence was a Scholar of St John's College and Barbara of Lady Margaret Hall, where she made a striking impression as an early feminist. She became active in social work in London at the turn of the century, while Lawrence was making his career as a Liberal journalist and later as Secretary of the Civil Service Commission (1907–13). But their increasingly precarious health (hers tubercular, his mainly coronary) led to a steady withdrawal to a life of authorship in the country, punctuated by Lawrence's intermittent work for the *Manchester Guardian* in later years. It is as pioneer social historians that they are remembered, especially for their 'labourer' trilogy. Their account of how agricultural workers fared under the enclosure measures of the period 1760–1830 opened up a far-reaching debate. They did not deny the economic rationality of the process, but pointed to the way in which its costs were borne by the rural poor (Hammond and Hammond 1911). Their work was given contemporary salience by the inception of Lloyd George's Land Campaign in 1913. In turning their attention to the urban working class, the Hammonds helped establish the 'pessimistic' view of the Industrial Revolution. Again, they did not disparage industrialization itself but focused on its exploitative effects, given the prevailing ideologies of an age which took social inequality for granted (Hammond and Hammond 1917). Published amid wartime planning for reconstruction, their findings once more fed current political debate. Finally, the Hammonds analysed the impact of technological change in making skilled craftsmen redundant in the early 19th century, and offered a new understanding of the Luddite movement (Hammond and Hammond 1919).

The Hammonds' view of the Industrial Revolution was countered in the 1920s by that of J.H. Clapham, with all his authority as Professor of Economic History at Cambridge. He mounted an 'optimistic' case on the standard of living by constructing a real wage index which showed substantial gains by industrial workers (Clapham 1926). The Hammonds publicly bowed on this point in face of the apparent weight of the new statistical evidence (Hammond 1930), though subsequent research has shown that Clapham's claims themselves depended upon a flawed price index. Insofar as the Hammonds rested their interpretation upon a quantitative assessment, therefore, it was by no means overturned; and its main thrust, in fact, was qualitative in its concern for the impact of economic change upon the lives of ordinary people. While they depicted the 'bleak age' of early industrialization, they also pointed to the civilizing process which urban life underwent from the middle of the 19th century (Hammond and Hammond 1930). Though often identified as socialists, the Hammonds remained liberal reformists in their outlook. Their work came to serve as a straw man, the object of ritual slights from a new generation of 'optimists' among professional economic historians; but its scholarly credentials have survived with remarkable resilience.

Selected Works

1911. *The village labourer, 1760–1832*. London: Longman, Green.

1917. *The town labourer, 1760–1832*. London: Longman, Green.

1919. *The skilled labourer, 1760–1932*. London: Longman, Green.

1925. *The rise of modern industry*. London: Methuen.
1930a. *The age of the chartists*. Revised as *The bleak age*, Harmondsworth: Penguin, 1947.
1930b. The industrial revolution and discontent. *Economic History Review* 2: 215–228.

Bibliography

Clapham, J.H. 1926. *An economic history of modern Britain*, The early railway age, 1820–1850, vol. 1. Cambridge: Cambridge University Press.

Hannan, Edward J. (1921–1994)

P. M. Robinson

Abstract

We review the research of the late Edward J. Hannan. Hannan contributed deeply and influentially to the development of econometric time series analysis, both in the elegance and incisiveness of his technical work and in his invention of new methodology.

Keywords

ARMA models; ARMAX models; Band-spectrum regression; Econometrics; Errors in variables; Hannan, E.; Hannan-efficient estimation; Heteroscedasticity; Maximum likelihood; Semi-parametric estimation; Statistics and economics; Time series analysis; Whittle estimation

JEL Classifications

B31

Ted Hannan, who died in 1994 at the age of 72, made contributions to statistical time series analysis of considerable depth and originality. His research by no means focused entirely on problems with econometric motivation, and he stopped publishing in econometric journals in the 1970s. However, Hannan introduced important econometric methodology and theory, and some of his ideas anticipated themes that later became important in econometrics. I focus on his most econometric-related contributions rather than attempt to survey the breadth of his research (of which an account can be found in Robinson, 1994, on which we draw upon here).

Hannan in fact started out as an economic researcher at the Reserve Bank of Australia, in Sydney, following an undergraduate commerce degree. While visiting the Australian National University (ANU), Canberra, in 1953 he was 'discovered' by the then Professor of Statistics P. A. P. Moran, who encouraged him to undertake doctoral research in statistics. Hannan quickly completed a Ph.D. and within a few years achieved professorial status at the ANU, retiring in 1986 but continuing to be very productive in research up to his death. Altogether he published over 130 articles and five books (all of which are definitely in the research monograph category).

Hannan's intellectual development was unusual in that his mathematical abilities and taste for abstraction increased throughout his career, suggesting that a different early training might well have led to a career in pure mathematics. This partly explains why it is the earlier part of his career in which he did most of his econometric-related work.

In the early 1950s testing for zero autocorrelation was a major theme of time series research. Two of Hannan's first papers, published in 1955 in *Biometrika*, concerned exact tests for autocorrelation. However he quickly realized the limitations of finite-sample theory, and began the research on asymptotic theory which he developed with such originality and power during the rest of his years. His first published contribution to asymptotics, in 1956, concerned Pitman efficiencies.

Soon thereafter he wrote an early, and widely uncredited, contribution to a topic that has, since the mid-1980s, been actively pursued in econometrics, so-called 'heteroscedasticity-and-autocorrelation consistent variance estimation'. It had already been noticed by Grenander and Rosenblatt that the variance of the sample mean of a weakly dependent time series is approximately proportional to the spectral density at zero frequency. On the other hand, Parzen and others had recently developed consistent smoothed estimation of nonparametric spectral densities. Hannan put these ideas together in a 1957 paper in the *Journal of the Royal Statistical Society*. The awareness he displayed here of the importance of bandwidth choice was notable for the time. Other early contributions included bias-correction in spectrum estimation for detrended data.

One major innovation for which Hannan does receive credit is what econometricians know as 'Hannan-efficient estimation'. The problem is one of efficiently estimating regression coefficients when the disturbances have nonparametric autocorrelation, that is, to do as well asymptotically as if one correctly assumed the disturbance followed a particular parametric model, such as an autoregression (as Cochrane and Orcutt had assumed). Based on the property of unitary transformation of a stationary time series to a heteroscedastic, approximately uncorrelated, one, Hannan, in a paper published in the 1963 Brown Symposium proceedings, proposed a frequency-domain generalized least squares procedure involving inverse weighting by the disturbance spectral density. Moreover, he established its asymptotic normality and efficiency. This was perhaps the first instance of justifying smoothed nonparametric estimation in a semi-parametric model. The technical difficulties here, of establishing parametric convergence rates despite a slowly converging nonparametric nuisance function, are now familiar, but Hannan was perhaps the first to solve them. He and others subsequently developed the approach in more general models, but it is most notable that his 1963 paper preceded by over 20 years work in the analogous problem of adapting to heteroscedasticity of nonparametric form, and by over ten years work in adapting to distribution of unknown form in location and regression models, though papers on these topics rarely mention his work. A related paper, also published in 1963, in *Biometrika*, is also insufficiently cited. There, Hannan introduced what subsequently became known as 'band-spectrum regression', omitting from a frequency-domain formulation of the least squares estimate non-degenerate bands of frequencies, with the aim of reducing errors-in-variables bias.

Around the same time Hannan introduced new ideas in the seasonal analysis of time series, using operators in estimating seasonality in the presence of trend and stationary noise, and modelling the seasonal component by a cosinusoid whose coefficients form stationary processes. He developed this model in a 1967 *Journal of Applied Probability* paper, allowing the coefficients to have roots on the unit circle, years before unit roots became the focus of so much econometric interest. Indeed, Hannan's model resembles the random component models that subsequently became popular. One non-time series contribution to econometrics was his work on the relation between canonical correlation and simultaneous equations estimation, which led to a 1967 *Econometrica* paper.

Hannan had published a short but influential 1960 monograph on time series analysis, and in 1970 he developed this into the major work *Multiple Time Series*. This constituted a detailed and rigorous account of time series, mainly in a multivariate setting, and covering continuous-time as well as discrete-time processes. It has been an invaluable reference for researchers, stimulating new research on a number of aspects.

Like the earlier book, *Multiple Time Series* put considerable stress on the frequency domain, but not exclusively and certainly not focusing particularly on nonparametric spectral methods. The frequency domain is sometimes misguidedly identified with nonparametric spectral estimation, but, just as nonparametric time series analysis can be considered in the time domain, so frequency-domain inference on parametric models is possible, and indeed the frequency domain is basic in much theoretical discussion of time series. Hannan's work showed how combining time-and frequency-domain assumptions can lead to an incisive theory, and also demonstrated immense resourcefulness in using techniques from Fourier analysis and other areas of mathematics in his proofs.

These qualities stand out in Hannan's work on linear time series models, which were the focus of much of his effort from around 1970 on. The publication of Box and Jenkins' 1970 book had greatly increased interest, especially among econometricians, in autoregressive moving average (ARMA) modelling. Hannan had already become interested in the topic, as *Multiple Time Series* shows. There, and in 1969 and 1971 papers, he addressed the difficult problem of identification in multivariate ARMA models, which he built on in several subsequent pieces of work.

In a 1973 *Journal of Applied Probability* paper, Hannan gave a rather definitive treatment of the asymptotic theory of various forms of Whittle estimation of scalar ARMA processes. This paper is notable for several aspects, which typify much of his work. First, while ARMA models are the linear time series models of leading practical importance, he showed that models with much stronger autocorrelation can be handled. Indeed, his proof of (strong) consistency actually covered long-range dependent processes, though these had not really been identified as a class at that time. Moreover, he showed that second moments suffice not only for consistency but for asymptotic normality (for parameters describing only autocorrelation). Another feature was his use of martingale difference rather than independence assumptions on innovations. From a methodological viewpoint, Hannan proposed a discrete-frequency version of Whittle estimation, which has computational advantages over Gaussian maximum likelihood or continuous-frequency Whittle, in that it makes nice use of the neat, explicit form of the spectral density for ARMA and some other models, and makes direct use of the fast Fourier transform algorithm. Later, in a 1976 paper with Dunsmuir, Hannan extended the 1973 paper to multivariate ARMA models, and in a 1980 paper with Deistler and Dunsmuir, to models with lagged explanatory variables ('ARMAX' models). From 1979 onwards, Hannan made several major contributions to the practically important problem of order determination in ARMA and ARMAX models. His 1988 book with Deistler, *The Statistical Theory of Linear Systems* collects much of his work on linear time series models.

Ted Hannan is mainly thought of as an outstanding technician, with the ability to elegantly solve highly challenging problems under conditions that are at the same time incisive and comprehensible, but he also repeatedly demonstrated a keen practical sense, inventing new methodology, involving interesting tricks, and well understanding computational issues. His influence in econometrics has been profound and lasting, but had he chosen to devote much more of his brilliance and energy to econometric problems it would be difficult to overstate the further benefits to the econometric profession that could have accrued.

See Also

▶ Econometrics
▶ Law(s) of Large Numbers
▶ Serial Correlation and Serial Dependence
▶ Statistics and Economics
▶ Time Series Analysis

Selected Works

1955a. Exact tests for serial correlation. *Biometrika* 42: 133–142.

1955b. An exact test for correlation between time series. *Biometrika* 42: 316–326.

1956. The asymptotic powers of certain tests based on multiple correlations. *Journal of the Royal Statistical Society B* 18: 227–233.

1957. The variance of the mean of a stationary process. *Journal of the Royal Statistical Society B* 19: 282–285.

1960. *Time Series Analysis*. London: Methuen. (Published in Russian with an Appendix by Yu. B. Rozanov in 1964; published in Japanese.)

1963a. Regression for time series. In *Time Series Analysis*, ed. M. Rosenblatt. New York: Wiley.

1963b. Regression for time series with errors of measurement. *Biometrika* 50: 293–302.

1967a. The measurement of a wandering signal amid noise. *Journal of Applied Probability* 4: 90–102.

1967b. Canonical correlation and multiple equation systems in economics. *Econometrica* 35: 123–138.

1969. The identification of vector, mixed autoregressive-moving average, systems. *Biometrika* 56: 223–225.

1970. *Multiple Time Series*. New York: Wiley. (Published in Russian in 1974.)

1971. The identification problem for multiple equation systems with moving average errors. *Econometrica* 39: 751–765.

1973. The asymptotic theory of linear time series models. *Journal of Applied Probability* 10: 130–145.

1976. (With W. Dunsmuir.) Vector linear time series models. *Advances in Applied Probability* 8: 339–364.

1979. (With B. Quinn.) The determination of the order of an autoregression. *Journal of the Royal Statistical Society B* 41: 190–195.

1980. (With W. Dunsmuir and M. Deistler.) Estimation of vector ARMAX models. *Journal of Multivariate Analysis* 10: 275–295.

1988. (With M. Deistler.) *The Statistical Theory of Linear Systems*. New York: Wiley.

Bibliography

Robinson, P. 1994. Edward J. Hannan, 1921–1994. *Journal of Time Series Analysis* 15: 563–576.

Hansen, Alvin (1887–1975)

Richard A. Musgrave

Keywords

American Economic Association; Bretton Woods system; Commons, J. R.; Council of economic advisers; Ely, R. T.; Fiscal expansion; Full employment; Great depression; Hansen, A.; International monetary fund; Keynesianism; Multiplier–accelerator interaction; Neoclassical synthesis; Saving and investment; Stabilization; Stagnation

JEL Classifications

B31

Alvin Hansen grew up in Viborg, South Dakota, a small rural community with a one-room school house and traditional values. Preferring academic pursuits to farm work, he proceeded to Sioux Falls for his high school education and then to Yankton College for his BA degree. Several years of high school teaching followed, with rapid advancement to principal and superintendent. The financial basis for his graduate work thus laid, Hansen entered the University of Wisconsin in 1914, where John R. Commons and R.T. Ely were to impress him with the importance of data and their institutional setting. In 1919 he moved to Brown University as assistant professor. There he completed his dissertation, later published as *Cycles of Prosperity and Depression* (1921). He then accepted a position at the University of Minnesota, where he remained for nearly 20 years. His major works during the 1920s included a solid *Principles* text, co-authored with F.B. Garver (1928), and an historical study of *Business Cycle Theory* (1927). Ranging from Malthus to Spiethoff and Hawtrey, stress was on structural shifts in investment rather than on monetary factors, and special attention was given to the interaction of short cycles with longer waves of economic development.

A Guggenheim fellowship in 1928 permitted extensive travel abroad, an experience that he continued to cherish and renew in later years. The early 1930s also brought a growing policy involvement outside the campus, activities that in subsequent years were to claim an increasing share of his time. Such early activities included that of Director of Research for the Committee of Inquiry on International Economic Relations (1933–34) and service as adviser on trade agreements to Secretary Cordell Hull.

In 1936 Harvard University had received a grant to establish the Littauer School of Public Administration, and Hansen was appointed as its first Lucius S. Littauer Professor of Political Economy. As fortune had it, his arrival at Harvard in the fall of 1937 closely followed the appearance of Keynes's *General Theory.* Hansen, distressed by the wastes of the Great Depression, soon (though with some initial hesitation) adopted the Keynesian approach. With Harvard's Fiscal Policy Seminar as his base, Hansen became the leading analyst and expositor of Keynesian economics in the United States. Driven by his enthusiasm for new ideas, his determination to find policy solutions, and his eagerness to learn as well as to teach, the seminar left a deep impact on the course of macroeconomics. The still obscure components of the Keynesian system had to be sorted out and new tools, such as the concept of governmental net contribution, the multiplier-accelerator model and the balanced budget theorem, were forged. With the application of these new tools to the setting of the US economy as its challenge, the seminar thus became the training ground for a generation of US policy economists.

The output of these years may be traced in Hansen's writings, beginning with the two key volumes of *Full Recovery and Stagnation* (1938) and *Fiscal Policy and Business Cycles* (1941). Other volumes followed, including *Business Cycles and National Income* (1951) and his widely used *A Guide to Keynes* (1953). The persistent theme was that of unemployment, caused by a failure of private investment to match the level of saving at a full employment income. With the effectiveness of monetary policy reduced by inelastic investment and high liquidity preference in a sluggish economy, the required level of aggregate demand would have to be provided by fiscal expansion responded to in the private sector by a multiplieraccelerator process. The need for expansionary fiscal policy, however, would not be one of pump-priming only. Linking back to his earlier interest in the long waves of the cycle, the weakness of the economy was seen as the downward phase of a long wave, with the declining population growth the most depressing factor. The stagnation thesis, offered in Hansen's presidential address before the American Economic Association (1937), placed the Keynesian model in a historical perspective and once more emphasized the strategic role of expansionary budget policy. Events, to be sure, proved different. The Second World War generated massive budgetary expansion and a strengthened post-war economy called for a correction, combining the traditional role of monetary policy with that of fiscal controls. Hansen the pragmatist welcomed the neoclassical synthesis of the mid-1960s.

While macro policy and stabilization remained his major concern, his activities during the Harvard years covered a much wider range. As a member of the Advisory Council on Social Security in 1937–38, he helped to shape the Social Security System. During 1941–43 he served as Chairman of the US–Canadian Joint Economic Commission, and from 1940 to 1945 he acted as Economic Advisor to the Federal Reserve Board. At the close of the war he participated in the monetary reconstruction of Bretton Woods and the birth of IMF. At the same time, he played a strategic role in the creation of the Full Employment Act of 1946 and the Council of Economic Advisers. After retiring from Harvard in 1956, Hansen remained in Belmont, Massachusetts, until 1972, when he joined his daughter in Virginia. He died there in 1975.

Throughout Hansen's work, the goal of full employment was central, as was the need for fiscal action to achieve it. His social philosophy was expressed 'in the democratic ideal of providing for all individuals a reasonable approach to equality of opportunity'. Beyond this, he was pragmatic and non-ideological in approach. For him, economics – as James Tobin put it when presenting him with the Walker Medal at his 80th birthday – was a science for the service of mankind.

Selected Works

1921. *Cycles of prosperity and depression in the United States, Great Britain and Germany: A study of monthly data 1902–1908.* Madison: University of Wisconsin Press.

1927. *Business-cycle theory*: *Its development and present status.* Boston: Ginn.

1928. (With F.B. Garver.) *Principles of economics.* Boston: Ginn.

1938. *Full recovery or stagnation?* New York: Norton.

1941. *Fiscal policy and business cycles*. New York: Norton.

1951. *Business cycles and national income*. New York: Norton.

1953. *A guide to Keynes.* New York: McGraw-Hill.

Hansen, Lars Peter (Born 1952)

Esther-Mirjam Sent

Abstract

Lars Peter Hansen is the 2013 recipient of the Nobel Prize in Economics along with Eugene Fama and Robert Shiller. Hansen has been instrumental in developing the Generalised Method of Moments, a statistical method that is particularly well suited to testing rational expectations economics. He is also known for his contributions to macroeconomics, in which he focuses on the linkages between the financial and real sectors of the economy. Five distinct phases may be identified in Hansen's research.

Keywords

Adaptive expectations; Asset pricing; Dynamic models; Econometrics; Empirical analysis; Fama, E.; Generalized method of moments estimators; Heaton, J.; Hodrick, R.; Instrumental variables; Jagannathan, R.; Lo, A.; Lucas, R.; Luttmer, E.; Rational expectations; Richard, S.; Risk; Rrobustness; Sargent, T.; Scheinkman, J.; Shiller, R.; Sims, C.; Singleton, K.; Stochastic discount factor models; Time series econometrics; Uncertainty

JEL Classification

B31; C1

Copyright © Nobel Foundation 2013. Photo: Alexander Mahmoud.

In 2013, Lars Peter Hansen received the Nobel Prize in Economic Sciences along with Eugene Fama and Robert Shiller 'for their empirical analysis of asset prices' (Press Release announcing the Nobel Prize 2013). All three recipients address puzzles related to financial market data. Fama does so from the efficient markets perspective, Shiller from the behavioural finance viewpoint, and Hansen with an econometric focus. According to the Nobel Committee, Hansen's contributions to the explanation of asset prices lie in developing a statistical method that is particularly well suited to testing rational theories of asset pricing.

Hansen was born in Champaign, Illinois, USA, on 26 October 1952. He was the son of Gaurth Hansen, Professor of Biochemistry and former Provost of Utah State University, and Anna Lou Hansen, a homemaker and active volunteer. After obtaining a BS in Mathematics and Political Science from Utah State University in 1974, he went on to the University of Minnesota to pursue a PhD degree in Economics, which he received in 1978. Hansen took classes from Thomas Sargent, and Christopher Sims served as his primary advisor. The topic of his dissertation was exhaustible resources.

After obtaining his doctorate, Hansen served as Assistant and then as Associate Professor at Carnegie Mellon University. He joined the University of Chicago's Department of Economics in 1981, where he is the David Rockefeller Distinguished Service Professor in Economics and Statistics and has served as department chairperson and director of graduate studies. He also serves as the Research Director of the Becker Friedman Institute for Research in Economics. Hansen is married to fellow University of Chicago economist Grace Tsjiang and they have a son, Peter. He has been a visiting professor at Harvard, MIT and Stanford. Hansen is a Fellow of the Econometric Society, National Academy of Sciences and American Finance Association and Distinguished Fellow of the Macro Finance Society. He is also a member of the American Academy of Arts and Sciences and served as President of the Econometric Society in 2007 and Vice President of the American Economic Association in 2011. He was selected to deliver the Third Toulouse Lectures in Economics at the Université de Toulouse in

2005, the distinguished Fisher Schultz Lecture to the Econometric Society in 2006, the Ely Lecture to the American Economic Association in 2007, the Tjalling C. Koopmans Memorial Lectures at Yale University in 2008 and the Princeton Lectures in Finance at the Benheim Center for Finance in 2010. Hansen is the co-winner of the Frisch Medal with Kenneth Singleton in 1984 and was awarded the Erwin Plein Nemmers Prize in Economics from Northwestern University in 2006 and the CME Group-MSRI Prize in Innovative Quantitative Applications in 2008. In 2010, he won the BBVA Foundation Frontiers of Knowledge Award in Economics, Finance, and Management 'for making fundamental contributions to our understanding of how economic actors cope with risky and changing environments'. He also received an honorary doctorate from Utah State University in 2012.

Overall Focus

The 2013 Nobel Award for Hansen was the third to recognise rational expectations economics, with Robert Lucas having received the prize in 1995 'for having developed and applied the hypothesis of rational expectations, and thereby having transformed macroeconomic analysis and deepened our understanding of economic policy' (Press Release announcing the Nobel Prize 1995) and Thomas Sargent and Christopher Sims in 2011 'for their empirical research on cause and effect in the macroeconomy' (Press Release announcing the Nobel Prize, 2011).

Economics experienced the so-called rational expectations revolution during the 1960s (Begg 1982; Guzzardi 1978; Kim 1988; Klamer 1983). This was a direct response to the adaptive expectations hypothesis, according to which people form their expectations about what will happen in the future based on what has happened in the past. Once a forecasting error is made by agents, due to a stochastic shock, the adaptive expectations hypothesis posits that they incorporate only part of their errors. As a result, agents will be unable to correctly forecast the price level again even if the price level experiences no further shocks. This backward nature of expectation formulation and the resultant systematic errors made by agents was unsatisfactory to rational expectations economists. Indeed, the central idea behind the rational expectations revolution was that individuals should not make systematic mistakes. Economic agents are not stupid: they learn from their mistakes and draw intelligent inferences about the future from what is happening around them.

As Hansen said in his interview with Jeff Sommer (2013): '[I]t's important to ask what happens if people actually think and have expectations about policy. Once you say, you can't just fool and trick people, you kind of ask, what's left of that policy? So I thought that was a tremendous insight'. Starting from this insight, Hansen worked on the boundary between rational expectations economics and statistics, in the field called rational expectations econometrics. Here his interest was to use statistics in productive ways to analyse dynamic rational expectations models. The challenge Hansen saw was that rational expectations econometrics requires a complete model specification, including a specification of the information available to the economic agents inside the model. What he tried to do was to create and apply statistical and quantitative methods that were able to study these complicated dynamic systems with limited information. Or in the phrase he used during his Nobel lecture: 'doing something without doing everything'. In the process, Hansen paid specific attention to the impact of uncertainty on dynamic economic models, as elaborated in the final section of this entry. Before returning to this overall focus, I identify the five different phases of Hansen's research, during which the focus was sometimes on developing methods and sometimes on studying a variety of economic phenomena ranging from consumption to investment and from exchange rates to asset pricing.

Phase 1: Generalised Method of Moments Estimators (1980–88)

Hansen's initial research concentrated on the development of the large sample properties of Generalised Method of Moments (GMM) Estimators (Hansen 1982b). GMM is a generic method for estimating parameters in statistical models that is designed for estimating dynamic models that are partially specified through conditional moment restrictions. Maximum likelihood estimation is not applicable in these situations, for the full shape of the distribution function of the data is not known. GMM merely requires that a certain number of moment conditions is specified for the model. As such, the strength of GMM estimation lies in the ability it offers to learn about something without needing to learn about everything. At the same time, an appeal to partial specification limits the questions that can be answered by an empirical investigation. And herein lies the weakness of GMM estimation, for the analysis of hypothetical interventions or policy changes typically requires a fully specified model.

GMM estimators opened the way to thinking about studying and testing a variety of different asset pricing models and models that link the macroeconomy and securities markets. Hansen himself elaborated the applications of GMM estimation in macroeconomics and finance in papers written with his colleagues at Carnegie Mellon University. Along with Robert Hodrick, his colleague at Carnegie Mellon at the time, Hansen investigated forward exchange rates as predictors of future spot rates (Hansen 1980b, 1983a). Along with Kenneth Singleton, who came to Carnegie Mellon after Hansen had left, Hansen studied nonlinear consumption-based intertemporal asset pricing models (Hansen 1982d, 1983e, 1988). GMM has proven particularly valuable for the estimation of rational expectations models, because it facilitates estimation without the need to impose strong explicit distribution assumptions. With his rational expectations colleague Thomas Sargent, Hansen explored the nature of the cross equation nonlinear restrictions that emerged from linear rational expectations models (Hansen 1980a, 1981, 1982c, 1983c, d).

Phase 2: Stochastic Discount Factor Models (1989–95)

As a natural outgrowth of his initial applied papers and his interest in GMM estimation in a dynamic setting, Hansen explored stochastic discount factors that simultaneously discount the future and adjust for risk. The discount factor is the factor by which a future cash flow must be multiplied in order to obtain the present value. A stochastic factor has a random probability distribution or pattern that may be analysed statistically but may not be predicted precisely. Stochastic discount factors represent market valuations of risky cash flows. In stochastic discount models, the price of an asset can be computed by 'discounting' the future cash flow by a stochastic factor and then taking the expectation. These have been used constructively in applied economic research in asset pricing. Along with Scott Richard and Ravi Jagannathan, Hansen expanded the methodological underpinnings of stochastic discount factor models and characterised some of the resulting empirical implications (Hansen 1987, 1991a).

Phase 3: Econometric Evaluation of Asset Prices (1995–2000)

Subsequently, Hansen advanced the empirical investigation of asset pricing models and considered estimation in the presence of model misspecification on the part of the econometrician, along with Ravi Jagannathan (Hansen 1997b) and John Heaton and Erzo Luttmer (Hansen 1995c). In this work, Hansen advocated the use of weighting matrices that are suboptimal from a statistical point of view, but which have desirable properties in financial applications.

The theory of GMM computes a weighting matrix based on the available data set. With the optimal weighting matrix, the resulting estimator is asymptotically efficient. This matrix is obtained by asking the following statistical question: given that a finite set of moment conditions are satisfied, what is the most efficient linear combination to use in estimating a parameter vector of interest? In this phase of his research, Hansen asked a different question, from an applied perspective: How can one keep pricing errors small if the model is misspecified?

Phase 4: Expansion Empirical Analysis of Macroeconomics–Finance Relationship (1995–2000)

Next, Hansen studied the term structure of macroeconomic risk in the presence of stochastic growth, along with John Heaton and Jose Scheinkman (Hansen 1996, 1997a, 1998). As described earlier, stochastic growth has a random probability distribution or pattern that may be analysed statistically but may not be predicted precisely. Macroeconomic risk refers to macroeconomic factors that influence the volatility over time of investments, assets, portfolios and the intrinsic value of companies. The term structure of risk, finally, refers to the entire range, from the short-term to the long-term.

In a dynamic setting with stochastic growth, risk prices depend on the time horizon of investment opportunities. That is, if one plans to invest for a long period of time, one can make more risky investments. Of course, returns play a role in one's investment decisions as well, and there is a so-called risk–return tradeoff. This is because potential return rises with an increase in risk. More precisely, low levels of risk tend to be associated with low potential returns, whereas high levels risk tend to be associated with high potential returns. According to the risk–return tradeoff, invested money can render higher profits only if it is subject to the possibility of being lost. From the perspective of rational expectations economics, long-run risk–return tradeoffs are encoded in equilibrium prices. During the fourth phase of his research career identified here, Hansen offered novel characterisations of these tradeoffs and explored the resulting measurement challenges.

Phase 5: Rational Agents Guarding Against Model Misspecification (2000–Present)

Recently, Hansen has intensified his collaboration with Thomas Sargent to study models in which economic agents are capable of robust decision-making (Hansen 2001a, b, 2002b, 2003a, b, 2005, 2006, 2007b, c, 2011). That is, they are smart enough to take model misspecification into account while making decisions. There are several reasons for acknowledging model misspecification on the part of agents with rational expectations. First, because econometricians face specification doubts, the agents inside the model might too. Second, rational expectations models tend to under-predict prices of risk from asset market data. Finally, while a long tradition dating back to Friedman (1953) casts doubts about model specification, these may be formalised by means of robustness.

In extending results from the field of robust control theory, Hansen studied how the concerns of economic actors about their limited knowledge of the future affect macroeconomic outcomes. In particular, Hansen investigated how this can be reflected in security market values and how learning in the presence of misspecified statistical models can contribute to price dynamics. While GMM – Hansen's focus during the first phase in his research career – is designed to estimate partially specified models, the robustness approach – Hansen's focus during the most recent phase in his research career – estimates fully specified dynamic rational expectations models using a misspecified likelihood function.

Concluding Comments

Seemingly highly technical, Hansen's contributions have profound implications for the modelling and understanding of the global financial crisis of 2008 (Hansen 2012d). This is for two central concerns that inspired Hansen throughout the research phases identified in this entry. First, he sought to integrate financial market performance with the macroeconomy. In particular, he developed statistical methods for exploring the interconnections between macroeconomic indicators and assets in financial markets. And in Hansen's words: '[T]he financial crisis exposed gaps in the existing models that were about the financial-macro linkages' (Interview by Sommer 2013). Second, Hansen's contributions were centred on uncertainty outside and inside economic models, which is also the title of his

Nobel lecture. This uncertainty poses serious challenges when seeking to conceptualise and quantify systemic risk in financial markets. This is the risk of the collapse of the entire financial system, as opposed to the risk associated with any one individual element of the system, which can be contained therein without harming the entire financial system. And mitigating systemic risk appears to be a common defence underlying the need for macro-prudential policy initiatives in response to the financial crisis.

In Hansen's view, uncertainty is related to risk and ambiguity. Risk concerns the probabilities assigned by a given model and ambiguity involves not knowing which among a family of models should be used to assess risk. The result is scepticism about the model specification. This uncertainty is felt by both researchers and investors. That is, researchers outside a model must estimate unknown parameters and assess model implications. And investors inside a model interact with economic agents (consumers, enterprises and policymakers) that cope with uncertainty and must acknowledge this uncertainty in their interactions.

The uncertainty outside and inside economic models identified by Hansen has implications for the policy response to the so-called great recession. This is because this uncertainty causes serious difficulties in measuring systemic risk in a meaningful way. In Hansen's opinion, this challenges the value of systemic risk as a guiding principle. According to Hansen, the best way forward is simplicity rather than trying to devise complicated solutions. Indeed, complicated problems do not necessarily require complicated solutions. For example, Hansen would rather have simple and transparent capital requirements for banks.

This is not to suggest that the entire global financial crisis of 2008 can be understood and solved with the contributions of Hansen, of course. Contrary to the rational expectations assumptions, markets do not clear, unsold goods are left and unemployed workers do exist. At the same time, the financial system that came tumbling down was based on the assumption of rational individual behaviour and market discipline. In particular, the financial crisis was a direct consequence of the deregulation of financial markets that was urged by the rational expectations approach. Models based on these insights turned out to have too much faith in financial markets and too little interest in the inner workings of the financial system. Indeed, convenience, not conviction, appears to dictate the choices that economists such as Hansen make. His models and those of his rational expectations colleagues are influential precisely because of their simplicity. At the same time, Hansen is more than willing to learn from these mistakes. Along with economist Andrew Lo, MIT Sloan Professor of Finance, Hansen co-directs the Macro Financial Modeling Group, a network of macroeconomists working to develop improved models of the linkages between the financial and real sectors of the economy in the wake of the 2008 financial crisis. In the process, this crisis may in the end prove to be a spur to Hansen's creativity.

See Also

- ▶ Capital Asset Pricing Model
- ▶ Lucas, Robert (Born 1937)
- ▶ Rational Expectations
- ▶ Sargent, Thomas J. (Born 1943)
- ▶ Time Series Analysis
- ▶ Uncertainty

Selected Works

1980a. (with T. Sargent) Formulating and estimating dynamic linear rationalexpectations models. *Journal of Economic Dynamics & Control* 2: 7–46.

1980b. (with R. Hodrick) Forward exchange-rates as optimal predictors of future spot rates – An econometric-analysis. *Journal of Political Economy* 88(5): 829–853.

1981. (with T. Sargent) A note on Wiener–Kolmogorov prediction formulas for rational expectations models. *Economic Letters* 8(3): 255–260.

1982a. Consumption, asset markets, and macroeconomic fluctuations – A comment. *Carnegie–Rochester Conference Series on Public Policy* 17: 239–250.

1982b. Large sample properties of generalized method of moments estimators. *Econometrica* 50(4): 1029–1054.

1982c. (with T. Sargent) Instrumental variables procedures for estimating linear rational expectations models. *Journal of Monetary Economics* 9(3): 263–296.

1982d. (with K. Singleton) Generalized instrumental variables estimation of nonlinear rational expectations models. *Econometrica* 50(5): 1269–1286.

1983a. (with R. Hodrick) Risk averse speculation in the forward foreign exchange market: An econometric analysis of linear models. In *Exchange rates and international macroeconomics*, ed. J. Frenkel. Chicago: University of Chicago Press.

1983b. (with R. Avery and V. Hotz) Multiperiod probit models and orthogonality condition estimation. *International Economic Review*24(1): 21–35.

1983c. (with T. Sargent.) Aggregation over time and the inverse optimal predictor problem for adaptive expectations in continuous time. *International Economic Review*24(1): 1–20.

1983d. (with T. Sargent) The dimensionality of the aliasing problem in models with rational spectral densities. *Econometrica* 51(2): 377–387.

1983e. (with K. Singleton) Stochastic consumption, risk aversion, and the temporal behavior of asset returns. *Journal of Political Economy* 91(2): 249–265.

1985. A method for calculating bounds on asymptotic covariance matrices of generalized method of moments estimators. *Journal of Econometrics* 30, 203–238.

1987. (with S. Richard) The role of conditioning information in deducing testable restrictions implied by dynamic asset pricing-models. *Econometrica* 55(3): 587–613.

1988. (with M. Eichenbaum and K. Singleton) A time-series analysis of representative agent models of consumption and leisure choice under uncertainty. *Quarterly Journal of Economics* 103(1): 51–78.

1990a. (with M. Eichenbaum) Estimating models with intertemporal substitution using aggregate time-series data. *Journal of Business & Economic Statistics* 8(1): 53–69.

1990b. (with A. Gallant and G. Tauchen) Using conditional moments of asset payoffs to infer the volatility of intertemporal marginal rates of substitution. *Journal of Econometrics* 45: 141–179.

1991a. (with R. Jagannathan) Implications of security market data for models of dynamic economies. *Journal of Political Economy* 99(2): 225–262.

1991b. (with T. Sargent) Exact linear rational expectations models: Specification and estimation. In *Rational expectations econometrics,* ed. L. Hansen and T. Sargent. Boulder/Oxford: Westview Press.

1991c. (with T. Sargent) *Rational expectations econometrics: Underground classics in economics*. Boulder/Oxford: Westview Press.

1993. (with T. Sargent) Seasonality and approximation errors in rational-expectations models. *Journal of Econometrics* 55: 21–55.

1995a. (with T. Sargent) Discounted linear exponential quadratic Gaussian control. *IEEE Transactions on Automatic Control* 40(5): 968–971.

1995b. (with J. Scheinkman) Back to the future: Generating moment implications for continuous time Markov-processes. *Econometrica* 63(4): 767–804.

1995c. (with J. Heaton and E. Luttmer) Econometric evaluation of asset pricing models. *Review of Financial Studies* 8(2): 237–274.

1996. (with J. Heaton and A. Yaron) Finite-sample properties of some alternative GMM estimators. *Journal of Business & Economic Statistics* 14(3): 262–280.

1997a. (with T. Conley, E. Luttmer and J. Scheinkmann) Short-term interest rates as subordinated diffusions. *Review of Financial Studies* 10(3): 525–577.

1997b. (with R. Jagannathan) Assessing specification errors in stochastic discount factor models. *Journal of Finance* 52(2): 557–590.

1998. (with J. Scheinkman and N. Touzi) Spectral methods for identifying scalar diffusions. *Journal of Econometrics* 86(1): 1–32.

1999. (with T. Sargent and R. Tallarini, Jr) Robust permanent income and pricing. *Review of Economic Studies* 66(4): 873–907.

2001a. (with T. Sargent) Robust control and model uncertainty. *American Economic Review* 91(2): 60–66.

2001b. (with T. Sargent) Acknowledging misspecification in macroeconomic theory. *Review of Economic Dynamics* 4(3): 519–535.

2001c. Generalized method of moments estimation: A time series perspective (published title "Method of Moments"). In *Methodology: Statistics, international encyclopedia of the social and behavior sciences*, ed. N. Smelser, P. Bates, S. Fienberg, and J. Kadane. Oxford: Pergamon.

2002a. (with M. Cagetti, T. Sargent and N. Williams) Robustness and pricing with uncertain growth. *Review of Financial Studies* 15(2): 363–404.

2002b. (with T. Sargent and N. Wang) Robust permanent income and pricing with filtering. *Macroeconomic Dynamics* 6(1): 40–84.

2003a. (with E. Anderson and T. Sargent) A quartet of semigroups for model specification, robustness, prices of risk and model detection. *Journal of the European Economic Association* 1(1): 68–123.

2003b. (with T. Sargent) Robust control of forward-looking models. *Journal of Monetary Economics* 50(3): 581–604.

2005. (with T. Sargent) Robust estimation and control under commitment. *Journal of Economic Theory* 124(2): 258–301.

2006. (with T. Sargent, G. Turmuhambetova and N. Williams) Robust control and model specification. *Journal of Economic Theory* 128(1): 45–90.

2007a. (with J. Heaton, J. Lee and N. Roussanov) Intertemporal substitution and risk aversion. *Handbook of Econometrics* 6(1): 3967–4056.

2007b. (with T. Sargent) Recursive robust estimation and control without commitment. *Journal of Economic Theory* 136(1): 1–27.

2007c. (with T. Sargent) *Robustness*. Princeton: Princeton University Press.

2009a. (with J. Scheinkman) Long term risk: An operator approach. *Econometrica* 77(1): 177–234.

2009b. (with X. Chen and J. Scheinkman) Nonlinear principal components and long run implications of multivariate diffusions. *Annals of Statistics* 37(6B): 4279–4312.

2010. (with Y. Ait-Sahalia and J. Scheinkman) Operator methods for continuous-time Markov processes. *Handbook of Financial Econometrics* 1(1): 1–66.

2011. (with T. Sargent) Robustness and ambiguity in continuous time. *Journal of Economic Theory* 146(3): 1195–1223.

2012a. (with J. Scheinkman) Pricing growth-rate risk. *Finance and Stochastics* 16(1): 1–15.

2012b. (with J. Scheinkman) Recursive utility in a Markov environment with stochastic growth. *Proceedings of the National Academy of Sciences* 109(30): 11967–11972.

2012c. Proofs for large sample properties of generalized method of moments estimators. *Journal of Econometrica* 170(2): 325–330.

2012d. Challenges in identifying and measuring systemic risk. *NBER Working Paper No. 18505*.

Bibliography

Begg, D. 1982. *The rational expectations revolution in macroeconomics*. Baltimore: Johns Hopkins University Press.

Eagleson, G. 1975. On Gordin's central limit theorem for stationary processes. *Journal of Applied Probability* 12(1): 176–179.

Fama, E. 1970. Efficient capital markets: A review of theory and empirical work. *Journal of Finance* 25(2): 383–417.

Friedman, M. 1953. The methodology of positive economics. In *Essays in positive economics*, 1–43. Chicago: University of Chicago Press.

Ghysels, E., A. Hall, and L. Hansen. 2002. Interview with Lars Peter Hansen. *Journal of Business & Economic Statistics Twentieth Anniversary Issue on the Generalized Method of Moments* 20(4): 442–447.

Gordin, M. 1969. The central limit theorem for stationary processes. *Soviet Mathematics Doklady* 10: 1174–1176.

Guzzardi, W. 1978. The new down-to-earth economics. *Fortune*, 21 December: 72–79.

Hall, P., and C. Heyde. 1980. *Martingale limit theory and its applications*. Boston: Academic Press.

Kim, K. 1988. *Equilibrium business cycle theory in historical perspective*. Cambridge: Cambridge University Press.

Klamer, A. 1983. *Conversations with economists*. Totowa: Rowman and Littlefield.

Lucas, R. 1972. Expectations and the neutrality of money. *Journal of Economic Theory* 4(2): 103–123.

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

Sargan, J. 1958. The estimation of economic relationships using instrumental variables. *Econometrica* 26: 393–415.

Sargan, J. 1959. The estimation of relationships with autocorrelated residuals by the use of the instrumental variables. *Journal of the Royal Statistical Society B* 21: 91–105.

Sargent, T. 1972. Rational expectations and the term structure of interest rates. *Journal of Money, Credit, and Banking* 4(1): 74–97.

Shiller, R. 1981. Do stock prices move too much to be justified by subsequent movements in dividends? *American Economic Review* 71(3): 421–436.

Sommer, J. 2013. A talk with Lars Peter Hansen, Nobel laureate. *The New York Times*, 16 November. http://economix.blogs.nytimes.com/2013/11/16/a-talk-with-lars-peter-hansen-nobel-laureate.

Happiness, Economics of

Carol Graham

Abstract

The economics of happiness assesses welfare by combining economists' and psychologists' techniques, and relies on more expansive notions of utility than does conventional economics. The research highlights factors other than income that affect well-being. It is well suited to informing questions in areas where revealed preferences provide limited information – for example, the welfare effects of inequality and of inflation and unemployment. Despite the potential contributions for policy, a note of caution is necessary because of the potential biases in survey data and the difficulties in controlling for unobservable personality traits.

Keywords

Behavioural economics; Bounded rationality; Capabilities-base approach to poverty; Democracy; Easterlin paradox; Easterlin, R.; Expectations; Expressed preferences; Gender; Income; vs. utility; Inequality; Interdependent utility; Misery index; Order bias; Poverty; Procedural utility; Psychologists; Revealed preferences; Race; Sen, A.; Status; Utility maximization; Walras, L

JEL Classifications

D31

The economics of happiness is an approach to assessing welfare which combines the techniques typically used by economists with those more commonly used by psychologists.

While psychologists have long used surveys of reported well-being to study happiness, economists only recently ventured into this arena. Early economists and philosophers, ranging from Aristotle to Bentham, Mill, and Smith, incorporated the pursuit of happiness in their work. Yet, as economics grew more rigorous and quantitative, more parsimonious definitions of welfare took hold. Utility was taken to depend only on income as mediated by individual choices or preferences within a rational individual's monetary budget constraint.

Even within a more orthodox framework, focusing purely on income can miss key elements of welfare. People have different preferences for material and non-material goods. They may choose a lower-paying but more personally rewarding job, for example. They are nonetheless acting to maximize utility in a classically Walrasian sense.

The study of happiness or subjective well-being is part of a more general move in economics that challenges these narrow assumptions. The introduction of bounded rationality and the establishment of behavioural economics, for example, have opened new lines of research. Happiness economics – which represents one new direction – relies on more expansive notions of utility and welfare, including interdependent utility functions, procedural utility, and the interaction between rational and non-rational influences in determining economic behaviour.

Richard Easterlin was the first modern economist to revisit the concept of happiness, beginning in the early 1970s. More generalized interest took hold in the late 1990s (see, among others, Easterlin 1974, 2003; Blanchflower and Oswald 2004; Clark and Oswald 1994; Frey and Stutzer 2002a; Graham and Pettinato 2002; Layard 2005).

The Approach

The economics of happiness does not purport to replace income-based measures of welfare but instead to complement them with broader measures of well-being. These measures are based on the results of large-scale surveys, across countries and over time, of hundreds of thousands of individuals who are asked to assess their own welfare. The surveys provide information about the importance of a range of factors which affect well-being, including income but also others such as health, marital and employment status, and civic trust.

The approach, which relies on expressed preferences rather than on revealed choices, is particularly well suited to answering questions in areas where a revealed preferences approach provides limited information. Indeed, it often uncovers discrepancies between expressed and revealed preferences. Revealed preferences cannot fully gauge the welfare effects of particular policies or institutional arrangements which individuals are powerless to change. Examples of these include the welfare effects of inequality, environmental degradation, and macroeconomic policies such as inflation and unemployment. Sen's capabilities-based approach to poverty, for example, highlights the lack of capacity of the poor to make choices or to take certain actions. In many of his writings, Sen (1995) criticizes economists' excessive focus on choice as a sole indicator of human behaviour. Another area where a choice approach is limited and happiness surveys can shed light is the welfare effects of addictive behaviours such as smoking and drug abuse.

Happiness surveys are based on questions in which the individual is asked, 'Generally speaking, how happy are you with your life' or 'How satisfied are you with your life', with possible answers on a four-to-seven point scale. Psychologists have a preference for life satisfaction questions. Yet answers to happiness and life satisfaction questions correlate quite closely. The correlation coefficient between the two – based on research on British data for 1975–92, which includes both questions, and Latin American data for 2000–1, in which alternative phrasing was used in different years – ranges between .56 and .50 (Blanchflower and Oswald 2004; Graham and Pettinato 2002).

This approach presents several methodological challenges (for a fuller description of these, see Bertrand and Mullainathan 2001; Frey and Stutzer 2002b). To minimize order bias, happiness questions must be placed at the beginning of surveys. As with all economic measurements, the answer of any specific individual may be biased by idiosyncratic, unobserved events. Bias in answers to happiness surveys can also result from unobserved personality traits and correlated measurement errors (which can be corrected via individual fixed effects if and when panel data are available). Other concerns about correlated unobserved variables are common to all economic disciplines.

Despite the potential pitfalls, cross-sections of large samples across countries and over time find remarkably consistent patterns in the determinants of happiness. Many errors are uncorrelated with the observed variables, and do not systematically bias the results. Psychologists, meanwhile, find validation in the way that people answer these surveys based in physiological measures of happiness, such as the frontal movements in the brain and in the number of 'genuine' – Duchenne – smiles (Diener and Seligman 2004).

Micro-econometric happiness equations have the standard form:

$W_{it} = \alpha + \beta x_{it} + \varepsilon_{it}$, where W is the reported well-being of individual i at time t, and X is a vector of known variables including socio-demographic and socioeconomic characteristics. Unobserved characteristics and measurement errors are captured in the error term. Because the answers to happiness surveys are ordinal rather than cardinal, they are best analysed via ordered logit or probit equations. These regressions typically yield lower R-squares than economists are used to, reflecting the extent to which emotions and other components of true well-being are driving the results, as opposed to the variables that we are able to measure, such as income, education, and marital and employment status. (Cross-section work also typically yields low R-squares.)

The availability of panel data in some instances, as well as advances in econometric techniques, are increasingly allowing for sounder analysis (Van Praag and Ferrer-i-Carbonell, 2004). The coefficients produced from ordered probit or logistic regressions are remarkably similar to those from OLS regressions based on the same equations. While it is impossible to measure the precise effects of independent variables on true well-being, happiness researchers have used the OLS coefficients as a basis for assigning relative weights to them. They can estimate how much income a typical individual in the United States or Britain would need to produce the same change in stated happiness that comes from the well-being loss resulting from, for example, divorce ($100,000) or job loss ($60,000) (Blanchflower and Oswald 2004).

The Easterlin Paradox

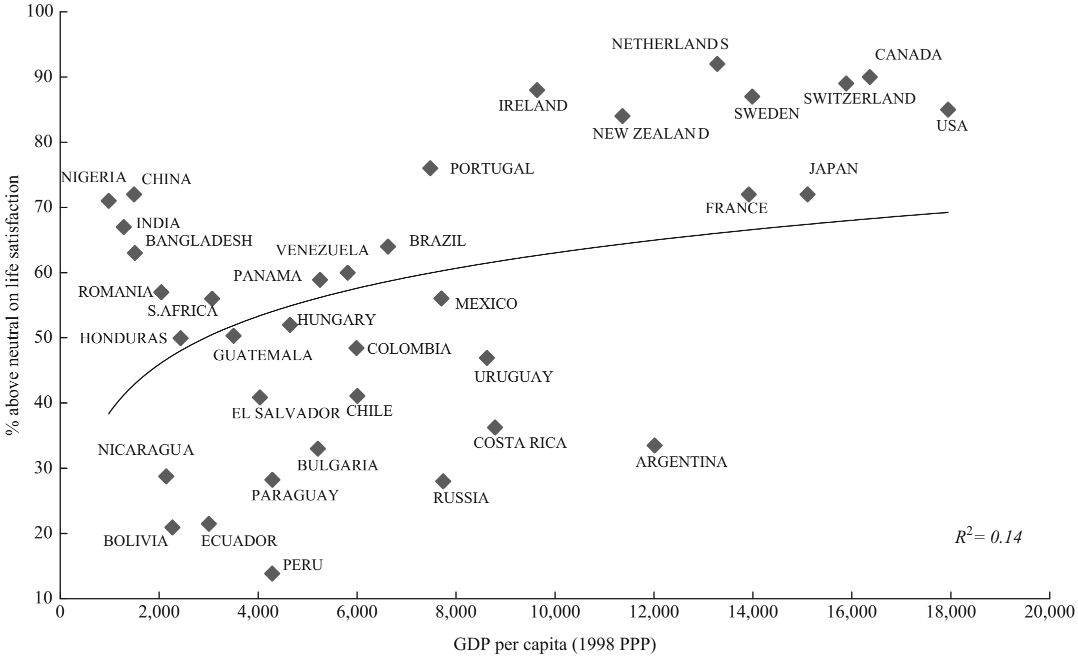
In his original study, Easterlin revealed a paradox that sparked interest in the topic but is as yet unresolved. While most happiness studies find that *within* countries wealthier people are, on average, happier than poor ones, studies across countries and over time find very little, if any, relationship between increases in per capita income and average happiness levels. On average, wealthier countries (as a group) are happier than poor ones (as a group); happiness seems to rise with income up to a point, but not beyond it. Yet even among the less happy, poorer countries, there is not a clear relationship between average income and average happiness levels, suggesting that many other factors – including cultural traits – are at play (see Fig. 1).

Within countries, income matters to happiness (Oswald 1997; Diener et al. 2003, among others). Deprivation and abject poverty in particular are very bad for happiness. Yet after basic needs are met other factors such as rising aspirations, relative income differences, and the security of gains become increasingly important, in addition to income. Long before the economics of happiness was established, James Duesenberry (1949) noted the impact of changing aspirations on income satisfaction and its potential effects on consumption and savings rates. Any number of happiness studies have since confirmed the effects of rising aspirations, and have also noted their potential role in driving excessive consumption and other perverse economic behaviours (Frank 1999).

Thus, a common interpretation of the Easterlin paradox is that humans are on a ‘hedonic treadmill’: aspirations increase along with income and, after basic needs are met, relative rather than absolute levels of income matter to well-being. Another interpretation of the paradox is the psychologists’ ‘set point’ theory of happiness, in which every individual is presumed to have a happiness level that he or she goes back to over time, even after major events such as winning the lottery or getting divorced (Easterlin 2003). The implication of this theory for policy is that nothing much can be done to increase happiness.

Individuals are remarkably adaptable, no doubt, and in the end can get used to most things, and in particular to income gains. The behavioural economics literature, for example, shows that individuals value losses more than gains (see Kahneman et al. 1999, among others). Easterlin argues that individuals adapt more in the pecuniary arena than in the non-pecuniary arena, while life changing events, such as bereavement, have lasting effects on happiness. Yet, because most policy is based on pecuniary measures of well-being, it overemphasizes the importance of income gains to well-being and underestimates that of other factors, such as health, family, and stable employment.

There is no consensus about which interpretation is most accurate. Yet numerous studies which demonstrate that happiness levels can change significantly in response to a variety of factors suggest that the research can yield insights into human well-being which provide important, if complementary, information for policymakers. Even under the rubric of set point theory, happiness levels can fall significantly in the aftermath of events like illness or unemployment. Even if levels eventually adapt upwards to a longer-term equilibrium, mitigating or preventing the unhappiness and disruption that individuals experience



Sources: Graham and Pettinato (2002);
Veenhoven (2002)

Happiness, Economics of, Fig. 1 Happiness and income per capita, selected countries, 1990s

for months, or even years, in the interim certainly seems a worthwhile objective for policy.

Selected Applications of Happiness Economics

Happiness research has been applied to a range of issues. Since a comprehensive review cannot be undertaken here, a selection of some of the issues the surveys can inform is provided. These include the relationship between income and happiness, inequality and poverty, the effects of macro-policies on individual welfare, and the effects of public policies aimed at controlling addictive substances.

Some studies have attempted to separate the effects of income from those of other endogenous factors, such as satisfaction in the workplace. Studies of unexpected lottery gains find that these isolated gains have positive effects on happiness, although it is not clear that they are of a lasting nature (Gardner and Oswald 2001). Other studies have explored the reverse direction of causality, and find that people with higher happiness levels tend to perform better in the labour market and to earn more income in the future (Diener et al. 2003; Graham et al. 2004).

A related question, and one which is still debated in economics, is how income inequality affects individual welfare. Interestingly, the results differ between developed and developing economies. Most studies of the United States and Europe find that inequality has modest or insignificant effects on happiness. The mixed results may reflect the fact that inequality can be a signal of future opportunity and mobility as much as it can be a sign of injustice (Alesina et al. 2004). In contrast, recent research on Latin America finds that inequality is negative for the well-being of the poor and positive for the rich. In a region where inequality is much higher and where public institutions and labour markets are notoriously inefficient, inequality signals persistent disadvantage or advantage rather than opportunity and mobility (Graham and Felton 2006).

Happiness surveys also facilitate the measurement of the effects of broader, non-income components of inequality, such as race, gender, and status, all of which seem to be highly significant (Graham and Felton 2006). These results find support in work in the health arena, which finds that relative social standing has significant effects on health outcomes (Marmot 2004).

Happiness research can deepen our understanding of poverty. The set point theory suggests that a destitute peasant can be very happy. While this contradicts a standard finding in the literature – namely, that poor people are less happy than wealthier people within countries – it is suggestive of the role that low expectations play in explaining persistent poverty in some cases. The procedural utilities and capabilities approaches, meanwhile, emphasize the constraints on the choices of the poor.

What is perceived to be poverty in one context may not be in another. People who are high up the income ladder can identify themselves as poor, while many of those who are below the objective poverty line do not, because of different expectations (Rojas 2004). In addition, the well-being of those who have escaped poverty is often undermined by insecurity and the risk of falling back into poverty. Income data does not reveal the vulnerability of these individuals, yet happiness data shows that it has strong negative effects on their welfare. Indeed, their reported well-being is often lower than that of the poor (Graham and Pettinato 2002).

Happiness surveys can be used to examine the effects of different macro-policy arrangements on well-being. Most studies find that inflation and unemployment have negative effects on happiness. The effects of unemployment are stronger than those of inflation, and hold above and beyond those of forgone income (Di Tella et al. 2001). The standard 'misery index', which assigns equal weight to inflation and unemployment, may be underestimating the effects of the latter on well-being (Frey and Stutzer 2002b).

Political arrangements also matter. Much of the literature finds that both trust and freedom have positive effects on happiness (Helliwell 2004; Layard 2005). Research based on variance in voting rights across cantons in Switzerland finds that there are positive effects from *participating* in direct democracy (Frey and Stutzer 2002b). Research in Latin America finds a strong positive correlation between happiness and preference for democracy (Graham and Sukhtankar 2004).

Happiness surveys can also be utilized to gauge the welfare effects of various public policies. How does a tax on addictive substances, such as tobacco and alcohol, for example, affect well-being? A recent study on cigarette taxes suggests that the negative financial effects may be outweighed by positive self-control effects (Gruber and Mullainathan 2005).

Policy Implications

Richard Layard (2005) makes a bold statement about the potential of happiness research to improve people's lives directly via changes in public policy. He highlights the extent to which people's happiness is affected by status – resulting in a rat race approach to work and to income gains, which in the end reduces well-being. He also notes the strong positive role of security in the workplace and in the home, and of the quality of social relationships and trust. He identifies direct implications for fiscal and labour market policy – in the form of taxation on excessive income gains and via re-evaluating the merits of performance-based pay.

While many economists would not agree with Layard's specific recommendations, there is nascent consensus that happiness surveys can serve as an important complementary tool for public policy. Scholars such as Diener and Seligman (2004) and Kahneman et al. (2004) advocate the creation of national well-being accounts to complement national income accounts. The nation of Bhutan, meanwhile, has introduced the concept of 'gross national happiness' to replace gross national product as a measure of national progress.

Despite the potential contributions that happiness research can make to policy, a sound note of caution is necessary in directly applying the findings, both because of the potential biases in survey data and because of the difficulties associated with analysing this kind of data in the absence of controls for unobservable personality traits. In addition, happiness surveys at times yield anomalous results which provide novel insights into human psychology – such as adaptation and coping during economic crises – but do not translate into viable policy recommendations.

One example is the finding that unemployed respondents are happier (or less unhappy) in contexts with higher unemployment rates. The positive effect that reduced stigma has on the well-being of the unemployed seems to outweigh the negative effects of a lower probability of future employment (Clark and Oswald 1994; Stutzer and Lalive, 2004; and Eggers et al. 2006). (Indeed, in Russia even *employed* respondents prefer higher regional unemployment rates. Given the dramatic nature of the late 1990s crisis, respondents may adapt their expectations downwards and are less critical of their own situation when others around them are unemployed.) One interpretation of these results for policy – raising unemployment rates – would obviously be a mistake. At the same time, the research suggests a new focus on the effects of stigma on the welfare of the unemployed.

Happiness economics also opens a field of research questions which still need to be addressed. These include the implications of well-being findings for national indicators and economic growth patterns; the effects of happiness on behaviour such as work effort, consumption, and investment; and the effects on political behaviour. In the case of the latter, surveys of unhappiness or frustration may be useful for gauging the potential for social unrest in various contexts.

In order to answer many of these questions, researchers need more and better quality well-being data, particularly panel data, which allows for the correction of unobserved personality traits and correlated measurement errors, as well as for better determining the direction of causality (for example, from contextual variables like income or health to happiness versus the other way around). These are major challenges in most happiness studies. Hopefully, the combination of better data and increased sophistication in econometric techniques will allow economists to better address these questions in the future.

See Also

- ▶ Sen, Amartya (Born 1933)
- ▶ Utilitarianism and Economic Theory
- ▶ Wage Inequality, Changes in

Acknowledgment I would like to thank Henry Aaron, Richard Easterlin, Joshua Epstein, Andrew Felton, Andrew Oswald, and George Perry for helpful comments.

Bibliography

Alesina, A., R. Di Tella, and R. MacCulloch. 2004. Inequality and happiness: Are Europeans and Americans different? *Journal of Public Economics* 88: 2009–2042.

Bertrand, M., and S. Mullainathan. 2001. Do people mean what they say? Implications for subjectivesurvey data. *American Economic Review* 91: 67–72.

Blanchflower, D., and A. Oswald. 2004. Well-being over time in Britain and the USA. *Journal of Public Economics* 88: 1359–1387.

Clark, A., and A. Oswald. 1994. Unhappiness and unemployment. *Economic Journal* 104: 648–659.

Di Tella, R., R. MacCulloch, and A. Oswald. 2001. Preferences over inflation and unemployment: Evidence from surveys of happiness. *American Economic Review* 91: 335–341.

Diener, E., and M. Seligman. 2004. Beyond money: Toward an economy of well-being. *Psychological Science in the Public Interest* 5(1): 1–31.

Diener, E., et al. 2003. The relationship between income and subjective well-being: Relative or absolute? *Social Indicators Research* 28: 195–223.

Duesenberry, J. 1949. *Income, savings, and the theory of human behavior*. Cambridge, MA: Harvard University Press.

Easterlin, R. 1974. Does economic growth improve the human lot? Some empirical evidence. In *Nations and households in economic growth*, ed. P. David and M. Reder. New York: Academic Press.

Easterlin, R. 2003. Explaining happiness. *Proceedings of the National Academy of Sciences* 100(19): 11176–11183.

Eggers, A., C. Gaddy, and C. Graham. 2006. Well-being and unemployment in Russia in the 1990s: Can society's suffering be individuals' solace? *Journal of Socio-economics* 35: 209–242.

Frank, R. 1999. *Luxury fever: Money and happiness in an era of excess*. Princeton, NJ: Princeton University Press.

Frey, B., and A. Stutzer. 2002a. *Happiness and Economics*. Princeton: Princeton University Press.

Frey, B., and A. Stutzer. 2002b. What can economists learn from happiness research? *Journal of Economic Literature* 40: 401–435.

Gardner, J., and A. Oswald. 2001. *Does money buy happiness? A longitudinal study using data on windfalls*. Mimeo. Coventry: University of Warwick.

Graham, C., and A. Felton. 2006. Inequality and happiness: Insights from Latin America. *Journal of Economic Inequality* 4: 107–122.

Graham, C., and S. Pettinato. 2002. *Happiness and hardship: Opportunity and insecurity in new market economies*. Washington, DC: Brookings Institution.

Graham, C., and S. Sukhtankar. 2004. Does economic crisis reduce support for markets and democracy in Latin America? Some evidence from surveys of public opinion and well-being. *Journal of Latin American Studies* 36: 349–377.

Graham, C., A. Eggers, and S. Sukhtankar. 2004. Does happiness pay? An initial exploration based on panel data from Russia. *Journalof Economic Behavior and Organization* 55: 319–342.

Gruber, J., and S. Mullainathan. 2005. Do cigarette taxes make smokers happier? *Advances in Economic Analysis & Policy* 5: 1412.

Helliwell, J. 2004. Well-being and social capital: does suicide pose a puzzle? Working Paper No. 10896. Cambridge, MA: NBER.

Kahneman, D., E. Diener, and N. Schwarz. 1999. *Well-being: The foundations of Hedonic Psychology.* New York: Russell Sage.

Kahneman, D., A. Krueger, D. Schkade, N. Schwarz, and A. Stone. 2004. Toward national well-being accounts. *AEA Papers and Proceedings* 94: 429–434.

Layard, R. 2005. *Happiness: Lessons from a new science*. New York: Penguin Press.

Marmot, M. 2004. *The status syndrome: How social standing affects our health and longevity.* London: Bloomsbury Press.

Hardy, Charles Oscar (1884–1948)

Robert B. Ekelund Jr.

Financial economist; born Island City, Missouri, 2 May 1884, died 30 November 1948. Hardy held posts at the University of Kansas and between 1918 and 1922 was a lecturer at the University of Chicago, where he had received the PhD in 1916. Hardy was also vice-president of the Federal Reserve Bank of Kansas City and was long associated with the Brookings Institution and with monetary policy debates of his time. As a Brookings scholar Hardy authored a number of books dealing with currency problems, focusing especially on the functioning of the gold standard. Hardy argued (1936) that the increase in the world's monetary gold stock (since 1929) led to undesirable expansions in floating credit and the potential for monetary instability. Further he thought that balance of trade shifts along with changes in long-term investments created havoc in the central bank's ability to have an impact upon domestic stability. He therefore argued for large-scale modifications in the gold standard as it was then practised. Hardy later advocated an activist fiscal policy, coordinated with monetary policy, to promote economic stabilization.

Hardy's most original and important contribution was to the theory of risk. In a 1923 paper (co-authored with Leverett S. Lyon, 1923b) Hardy analysed the functioning of futures markets in detail, carefully and correctly explaining why hedging contracts cannot be expected to provide complete protection to the user against the risk of adverse price changes. In the same year Hardy authored a pre-Knightian textbook on risk (1923a). In it Hardy features uncertainty as well as risk as elements in production and investment, crediting his colleague Frank Knight for access to preliminary versions of Knight's *Risk, Uncertainty, and Profit*.

Selected Works

1923a. *Risk and risk-bearing*. Chicago: University of Chicago Press.

1923b. (With L.S. Lyon.) The theory of hedging. *Journal of Political Economy* 31: 276–87.

1932. *Credit policies of the federal reserve system*. Washington, DC: The Brookings Institution.

1936. *Is there enough gold?* Washington, DC: The Brookings Institution.

Harris, Seymour Edwin (1897–1975)

John Kenneth Galbraith

Keywords

Federal Reserve System; Harris, S. E.; Keynesianism; New deal

JEL Classifications

B31

Harris was born in Brooklyn, New York, and graduated from Harvard University, where he also took his doctorate. His career, apart from the Second World War period and a few post-retirement years at the University of California at La Jolla, was spent at Harvard. In the Second World War, he was in charge of the pricing of exported and imported products and various liaison tasks for the Office of Price Administration. Throughout his life he undertook numerous regional and developmental assignments in New England and was one of the founders of the highly successful Massachusetts community college system.

Harris's early academic work, including a major history of the Federal Reserve System, was competent, orthodox and, as he would later view it, uninspired. Upward progress in his academic career at Harvard was also gradual and unspectacular, a circumstance related at the time to his Jewish origins. In later years he emerged as one of the most highly regarded members of the Cambridge (USA) economic and university community. He became a highly respected chairman of the Harvard economics department, and was the editor of the *Review of Economics and Statistics* and of numerous essay collections by fellow economists. He did not entirely escape criticism from his more relaxed colleagues for his prodigious work and publication schedule. President John F. Kennedy, shortly before he was killed, told of his intention of making Harris his next appointment to the Board of Governors of the Federal Reserve System.

From his earlier orthodox, even conservative, tendencies Harris was released by Keynes and the New Deal. His work came to reflect a strong commitment to Keynesian economics and policy and to the broad welfare measures of the Roosevelt, Kennedy and Johnson years. He was not a compelling writer; in his books, however, this was more than compensated for by the solid competence of his research and preparation, his strongly compassionate views on welfare issues and his very evident desire to extend knowledge on a great range of subject matter. On the economics of health care, education, social security, international monetary policy, central-bank policy, monetary history and literally a dozen other topics, he provided the basic source material from which legislators learned what could be done, what should be done and how it might be done. A full listing of his works would be among the longest in this Dictionary. Among the prominent later examples are those listed below.

Selected Works

1947. *The new economics: Keynes' influence on theory and public policy.* New York: A.A. Knopf.

1948a. *How shall we pay for education? Approaches to the economics of education.* New York: Harper.

1948b. *Saving American capitalism: A liberal economic program.* New York: A.A. Knopf.

1949. *The market for college graduates, and related aspects of education and income.* Cambridge, MA: Harvard University Press.

1952. *The economics of New England: Case study of an older area.* Cambridge, MA: Harvard University Press.

1953. *National health insurance, and alternative plans for financing health.* Foreword by Alfred Baker Lewis. New York: League for Industrial Democracy.

1955. *John Maynard Keynes, economist and policy maker.* New York: Scribner.

1962. *Higher education: Resources and finance.* New York: McGraw-Hill.

1964. *The Economics of American Medicine.* New York: Macmillan.

Harris–Todaro Hypothesis

M. Ali Khan

Abstract

The Harris–Todaro hypothesis replaces the equality of wages by the equality of 'expected' wages as the basic equilibrium condition in a segmented, but homogeneous, labour market, and in so doing generates an equilibrium level of urban unemployment when a mechanism for the determination of urban wages is specified. This article reviews work in which the Harris–Todaro hypothesis is embedded in canonical models of trade theory in order to investigate a variety of issues in development economics. These include the desirability (or the lack thereof) of foreign investment, the complications of an informal sector and the presence of clearly identifiable ethnic groups.

Keywords

Development economics; Gains from trade; Harris–Todaro hypothesis; Harris–Todaro model; Heckscher–Ohlin–Samuelson trade model; Immiserizing growth; Informal urban sector; International trade theory; Ricardo–Viner model; Rural-urban migration; Rybczynski theorem; Stolper–Samuelson theorem; Trade unions; Underemployment; Urban unemployment; Urban wage determination; Urban wage subsidies; Wage differentials

JEL Classifications

F22

The replacement of the equality of wages by the equality of 'expected' wages as the basic equilibrium condition in a segmented, but homogeneous, labour market has proved to be an idea of seminal importance in development economics. Attributed originally to Todaro (1968, 1969) and Harris and Todaro (1970), and commonly referred to as the Harris–Todaro hypothesis, the idea was very much in the air around the late 1960s, as can be seen from the contemporaneous writings of Akerlof and Stiglitz (1969), Blaug et al. (1969) and Harberger (1971), among others.

The motivation for the Harris–Todaro hypothesis lies in an attempt to explain the persistence of rural to urban migration in the presence of widespread urban unemployment, a pervasive phenomenon in many, so-called less-developed, countries (but also see Suits 1985; Partridge and Rickman 1997). It is natural to ask why such unemployment does not act as a deterrent to further migration. According to the Harris–Todaro hypothesis, the answer lies in the migrant leaving a secure rural wage w_r for a higher expected urban wage w_u^e even though the latter carries with it a non-zero probability of urban unemployment. The expected wage is computed by using the rate of urban employment as an index for the probability of finding a job. Thus

$$w_u^e = w_u \frac{L_u}{L_u + U} + 0 \frac{U}{L_u + U} = w_u \frac{1}{1+\lambda}, \qquad (1)$$

where w_u is the urban wage, L_u is the number of urban employed, U the number of urban unemployed and $\lambda = (U/L_u)$ the rate of urban unemployment. Thus, the Harris–Todaro hypothesis is precisely formulated by the equilibrium condition

$$w_r = w_u^e \Leftrightarrow w_u = w_r(1+\lambda). \qquad (2)$$

Since the Harris–Todaro hypothesis introduces a further unknown, namely, the rate of unemployment, a model in which the hypothesis is embedded must be buttressed by a theory of urban wage determination. The simplest setting is the one originally adopted by Harris–Todaro and subsequently by Bhagwati and Srinivasan (1971, 1973, 1974). This setting assumes the urban wage to be an exogenously given constant and typically rationalizes it as a consequence of government fiat.

In the 1970s, however, several theories of endogenous urban wage determination were simultaneously proposed. Foremost among these is the work of Stiglitz, who provides a microfoundation for the urban wage in terms of labour turnover (Stiglitz 1974), or in terms of biological efficiency considerations (Stiglitz 1976). One may also mention in this context the work of Calvo (1978), who sees the equilibrium urban wage as an outcome of trade union behaviour (also Quibria 1988; Chau and Khan 2001; and Calvo and Wellisz 1978, who see a higher urban wage as a consequence of costly supervision). At this stage of the development of the literature, each theory of urban wage determination led to a particular version of the Harris–Todaro model, and the common structural similarities were obscured.

In Khan (1980a), the elementary observation is made that all these variants of the Harris–Todaro model could be studied under one rubric if the Harris–Todaro hypothesis is embedded in the Heckscher–Ohlin–Samuelson (HOS) two-sector, so- called general equilibrium model (see Jones 1965; Johnson 1971), and the determination of urban wages is seen in a somewhat more abstract way, that is,

$$w_u = \Omega(w_r, \lambda, R, \tau), \tag{3}$$

where R is the rental on capital and τ a shift parameter. This led to a model whose importance lay, not so much in synthesizing the several variants of urban wage determination, but in emphasizing its points of contact with the trade theory literature. In particular, when (3) collapses to

$$w_u = w_r, \tag{4}$$

that is, when the elasticity of the omega function $\Omega(\cdot)$ with respect to w_r is unity, and those with respect to R and λ are zero, we obtain the HOS model.

This point deserves further articulation. Let a stylized economy consist solely of an urban and a rural sector, indexed by u and r respectively, and be endowed with positive amounts of labour L and capital K. Let the i^{th} sector produce a commodity i in amount X_i in accordance with a production function

$$X_i = F_i(L_i, K_i), i = u, r, \tag{5}$$

which is assumed to exhibit constant returns to scale and is twice continuously differentiable and concave. The allocation of labour and capital, L_i and K_i is determined through marginal productivity pricing. Thus, we have

$$p_r F_r^K = R = p_u F_u^K, \quad p_r F_r^L = w_r \text{ and } p_u F_u^L = w_u, \tag{6}$$

where F_i^j is the derivative of $F_i(i = u, r)$ with respect to $j(j = L_i, K_i)$. The economy is considered too small to influence the positive international prices of the two commodities, p_u and p_r. On rewriting the equilibrium condition (2) in the slightly more general form,

$$w_u = \rho w_r(1 + \lambda); \rho \text{ a shift parameter}, \tag{7}$$

(3), (5), (6) and (7), along with the material balance equations below, complete the specification of the model.

$$K_r + K_u = K \text{ and } L_r + L_u(1 + \lambda) = L. \tag{8}$$

The first point to be noticed about this model is a *decomposability property* whereby the factor prices, w_u,w_r, R and the unemployment rate λ are all independent of the endowments of labour and capital and depend solely on p_u , p_r and the shift parameters τ and ρ. This can be seen most easily if we subsume the marginal productivity conditions (6) into price-equal-unit-cost equations

$$p_i = C_i(w_i, R), i = u, r. \tag{9}$$

This allows one to decompose the model into a subsystem comprising Eqs. (7) and (3) along with (9). This basic observation leads to several interesting characteristics of the equilibria of the model. First, the market rural wage and market rental correctly measure the social opportunity cost of labour and capital if we use the international value of GNP as the relevant measure of social welfare. Second, despite the presence of a distorted labour market, there is no possibility of immiserizing growth. Third, an increase in capital (labour) increases the output of the capital- (labour-) intensive commodity provided the of the labour- (capital-) intensive commodity provided the intensities are measured in employment adjusted terms, that is

$$\frac{k_u}{1+\lambda} = \frac{K_u}{L_u(1+\lambda)} > (\text{or} <) \frac{K_r}{L_r} = k_r. \qquad (10)$$

This third property is an analogue of the Rybczynski property of the HOS model. Not surprisingly, we also obtain an analogue of the Stolper–Samuelson property whereby the effect of changes in international prices on factor returns depends on factor intensities, provided these are now measured in elasticity adjusted terms. The urban sector is said to be capital intensive in elasticity adjusted terms if

$$\begin{aligned}&\theta_{rL}(\theta_{uK}(1-e_\lambda)+\theta_{uL}e_R)\\&\quad-\theta_{uL}\theta_{rK}(e_w-e_\lambda)\\&\quad>(\text{or} <)0,\end{aligned} \qquad (11)$$

where θ_{ij} is the share of the j^{th} factor ($j = K, L$) in the i^{th} sector ($i = u, r$), and e_i is the elasticity of the $\Omega(\cdot)$ function with respect to the relevant variable. In the setting where e_w equals unity and e_R and e_λ are all zero, (10) and (11) collapse to the conventional physical and value intensities of Magee (1976) and Jones (1971) for the HOS model with proportional wage differentials. Under the further specialization that p in (7) equals unity, there is no difference between these two kinds of intensities and a perfect correspondence between the Rybczynski and Stolper–Samuelson theorems.

This reappearance of the divergence of the physical and value intensities of the wage-differential model leads us to inquire into the possibility of downward-sloping supply curves of X_r and X_u. This is indeed a possibility, and a sharp generalization is available in the result that there are perverse price–output responses in the model if and only if the employment-adjusted factor intensities do not conflict with the elasticity-adjusted intensities; see Khan (1980b) for details. Another direct consequence of the decomposability property of the model is a generalization of the Bhagwati (1968), Johnson (1971), Brecher and Alejandro (1977) paradox. This states that capital inflow in the presence of a tariff and with full repatriation of its earnings is immiserizing if and only if the imported commodity is capital intensive in employment-adjusted terms. This result is independent of the various mechanisms for the determination of urban wages; see Khan (1982a) for details, and also subsequent work by Beladi and Naqvi (1988), Grinols (1991), Chao and Yu (1994, 1995c), Chaudhuri and Mukhopadhyay (2002), Chaudhuri (2001) and Sen et al. (1997). Both of these results have a trade-theoretic flavour, and one question that has remained in the forefront of analytical work on the Harris–Todaro hypothesis relates to the effect of urban wage subsidies on urban unemployment and urban output. (As emphasized above, this question could indeed be seen as the *raison d'être*for the introduction of the hypothesis.) A seminal result here is the Corden and Findlay (1975) paradox, which draws attention to the fact that urban employment and urban output could rise if the urban wage is increased. This question has been readdressed by Neary (1981) and completely resolved in the context of endogenous urban wage determination by Khan (1980b).

So far we have focused on the comparative-static properties of the Harris–Todaro equilibrium. It is also worth emphasizing that the actual existence of the Harris–Todaro equilibrium cannot be taken for granted and must be proved. In the original Harris–Todaro model with an exogenously given rigid wage, equilibrium exists if and only if the rural sector is more capital intensive in employment-adjusted terms; see Khan (1980a) and Basu (1991) for an application of the geometric technique. Furthermore, once the 'isomorphism' with the HOS model is established and understood, one can follow Neary's (1978) lead and ask for 'reasonable' adjustment processes under which the Harris–Todaro equilibrium is locally asymptotically stable. It can be shown that an adjustment process of the Marshallian type leads to a stable equilibrium if and only if the employment-adjusted factor intensities do not conflict with the elasticity-adjusted intensities; see Khan (1980b) for details. Since the elasticity-adjusted intensities of (11) collapse to $\theta_{rL}\theta_{uK}$ in the Harris–Todaro model with a rigid wage, we have the satisfying result that the criteria for the existence of equilibrium and its stability coincide; also see Neary (1981) for this special case.

The entry on this subject in the first (1987) edition of *The New Palgrave: A Dictionary of Economics* was furnished under the title Harris–Todaro hypothesis, and the model presented above referred to as the 'generalized Harris–Todaro' (GHT) model. This is somewhat misleading in that any model in which the Harris–Todaro hypothesis is embedded has a justifiable claim to the title of a Harris–Todaro model. Indeed, unlike the case of the HOS model where capital is intersectorally mobile, the hypothesis can be embedded in the Ricardo–Viner model, a setting with three factors, or under an alternative interpretation, one where capital can be viewed as non-shiftable (for details on this and other basic constructions of classical trade theory, see, for example, Caves and Jones 1985). In many ways, this case of a two- sector model with sector-specific capital is more difficult and also more interesting; see Khan (1982a, b) and Bhatia (2002) for details. And there is at least one example in the literature where a particular Harris–Todaro model has been exported to international trade theory rather than imported from it: Jones and Marjit (1992) investigate a multi-sectoral setting of Khan (1991) by stripping it of the Harris–Todaro hypothesis.

This updated entry would be seriously incomplete if it did not note a criticism of the Harris–Todaro hypothesis centering on the urban unemployed living on a zero wage, and a corresponding generalization of the hypothesis. This criticism also dovetails into an issue that has received increasing attention from sociologists and development economists since the early 1990s: the existence of a dynamic informal urban sector, and the possibility of the urban unemployed being incorporated in it; see Portes et al. (1989) and Fields (1975, 2005b) and their references. This has led to a reformulation of (1) and (2) to

$$\begin{aligned} w_u^e &= w_u \frac{L_u}{L_u + U} + w_i \frac{U}{L_u + U} \\ &= \frac{w_u + \lambda w_i}{1 + \lambda} \Rightarrow w_r = w_u^e \\ &\Leftrightarrow w_u + \lambda w_i = w_r(1 + \lambda), \end{aligned} \tag{12}$$

where w_i is the wage in the informal sector. Again, as in the original Harris–Todaro hypothesis, this generalized hypothesis can be embedded in alternative production structures to yield a variety of models tailored to the purpose the investigator has in mind; see Chandra (1991) and Chandra and Khan (1993) for a more detailed elaboration of this point of view. The subject continues to receive attention; see Stiglitz (1982), Fields (1990, 1997), Rauch (1991), Gupta (1993, 1997a, b), Bandyopadhyay and Gupta (1995), Kar and Marjit (2001), Yabuuchi and Beladi (2001), Yabuuchi et al. (2005) and Chaudhuri (2003).

We conclude this article with a partial list of some other issues in trade and development that have been discussed in the context of urban-rural migration: gains from trade, now depending on the asymmetric nature of the model and on whether the rural or the urban commodity is being exported, as in Khan and Lin (1982), Chao and Yu (1993, 1997, 1999) and Choi and Yu (2006); underemployment or educated unemployment as in Bhagwati and Srinivasan (1977) or in Chaudhuri and Khan (1984) and Chaudhuri and Mukhopadhyay (2003); public inputs as in Chao et al. (2006); variable returns to scale as in Panagariya and Succar (1986), Beladi (1988) and Choi (1999); growth and technical progress as in Bourguignion (1990), Chao and Yu (1995a) and Chow and Zeng (2001); foreign enclaves as in Gupta and Gupta (1998); capital markets, distorted or otherwise, as in Khan and Naqvi (1983) and Chao and Yu (1992); interaction of ethnic groups as in Khan (1979, 1991) and Khan and Chaudhuri (1985); risk and uncertainty as in Beladi and Ingene (1994); environmental issues, as in Chao et al. (2000) and Chao and Yu (2003); cost-benefit analyses as in Srinivasan and Bhagwati (1975), Stiglitz (1977, 1982), Gupta (1988) and Chao and Yu (1995b); poverty and income inequality as in Moene (1992) and Rauch (1993). In summary then, the Harris–Todaro hypothesis is a versatile and useful analytic instrument for investigating a variety of questions arising in international and development economics where urban unemployment is a prominent issue.

See Also

- ▶ Development Economics
- ▶ Heckscher–Ohlin Trade Theory
- ▶ Unemployment

Bibliography

Agesa, R. 2000. The incentive for rural to urban migration: A re-examination of the Harris–Todaro model. *Applied Economics Letters* 7: 107–110.

Akerlof, G., and J.E. Stiglitz. 1969. Capital, wages and structural employment. *Economic Journal* 79: 269–281.

Bandyopadhyay, M., and M.R. Gupta. 1995. Development policies in the presence of an informal sector: A note. *Journal of Economics* 61: 301–315.

Basu, A. 1991. Locational choice for free trade zones: A comment. *Journal of Development Economics* 50: 381–387.

Beladi, H. 1988. Variable returns to scale, urban unemployment and welfare. *Southern Economic Journal* 55: 412–423.

Beladi, H., and C.A. Ingene. 1994. A general equilibrium analysis of rural-urban migration under uncertainty. *Journal of Regional Science* 34: 91–103.

Beladi, H., and N. Naqvi. 1988. Urban unemployment and non-immiserizing growth. *Journal of Development Economics* 28: 365–376.

Bhagwati, J.N. 1968. Distortions and immiserizing growth. *Review of Economic Studies* 35: 481–485.

Bhagwati, J.N., and T.N. Srinivasan. 1971. The theory of wage differentials: Production response and factor price equalization. *Journal of International Economics* 1: 19–35.

Bhagwati, J.N., and T.N. Srinivasan. 1973. The ranking of policy interventions under factor market imperfections: The case of sector-specific sticky wages and unemployment. *Sankhya,*Series B 35: 405–420.

Bhagwati, J.N., and T.N. Srinivasan. 1974. On reanalyzing the Harris–Todaro model: Policy rankings in the case of sector-specific sticky wages. *American Economic Review* 64: 502–508.

Bhagwati, J.N., and T.N. Srinivasan. 1977. Education in a job ladder model and the fairness–in–hiring rule. *Journal of Public Economics* 7: 1–22.

Bhatia, K. 2002. Specific and mobile capital, migration and unemployment in a Harris–Todaro model. *Journal of International Trade and Economic Development* 11: 207–222.

Blaug, M., P.R.G. Layard, and M. Woodhall. 1969. *The causes of graduate unemployment in India*. London: Allen Lane.

Bourguignion, F. 1990. Growth and inequality in a dual model of development: The role of demand factors. *Review of Economic Studies* 64: 502–508.

Brecher, R.A., and C.F. Diaz-Alejandro. 1977. Tariffs, foreign capital and immiserizing growth. *Journal of International Economics* 7: 317–322.

Calvo, G.A. 1978. Urban unemployment and wage determination in LDC's: Trade unions in the Harris–Todaro model. *International Economic Review* 19: 65–81.

Calvo, G.A., and S. Wellisz. 1978. Supervision, loss of control and the optimum size of the firm. *Journal of Political Economy* 86: 943–952.

Caves, R.E., and R.W. Jones. 1985. *World trade and payments*. 4th ed. Boston: Little, Brown & Co.

Chakravarty, S.R., and B. Dutta. 1990. Migration and welfare. *European Journal of Political Economy* 6: 119–138.

Chandra, V. 1991. *The informal sector in developing countries: A theoretical analysis*. Ph.D. thesis, Johns Hopkins University.

Chandra, V., and M. Ali Khan. 1993. Foreign investment in the presence of an informal sector. *Economica* 60: 79–103.

Chao, C.C., and E.S.H. Yu. 1990. Urban unemployment, terms of trade and welfare. *Southern Economic Journal* 56: 743–751.

Chao, C.C., and E.S.H. Yu. 1992. Capital markets, urban unemployment and land. *Journal of Development Economics* 38: 407–413.

Chao, C.C., and E.S.H. Yu. 1993. Content protection, urban unemployment and welfare. *Canadian Journal of Economics* 26: 481–492.

Chao, C.C., and E.S.H. Yu. 1994. Foreign capital inflows and welfare in an economy with imperfect competition. *Journal of Development Economics* 45: 141–154.

Chao, C.C., and E.S.H. Yu. 1995a. Urban growth, externality and welfare. *Regional Science and Urban Economics* 24: 565–576.

Chao, C.C., and E.S.H. Yu. 1995b. The shadow price of foreign exchange in a dual economy. *Journal of Development Economics* 46: 195–202.

Chao, C.C., and E.S.H. Yu. 1995c. International capital mobility, urban unemployment and welfare. *Southern Economic Journal* 61: 486–492.

Chao, C.C., and E.S.H. Yu. 1997. Trade liberalization in oligopolistic competition with unemployment: A general equilibrium analysis. *Canadian Journal of Economics* 30: 479–496.

Chao, C.C., and E.S.H. Yu. 1999. Export promotion, unemployment and national welfare. *International Economic Journal* 13: 17–34.

Chao, C.C., and E.S.H. Yu. 2003. Jobs, production linkages and the environment: Export promotion, unemployment and national welfare. *Journal of Economics* 79: 113–122.

Chao, C.C., J. Kerkviliet, and E.S.H. Yu. 2000. Environmental preservation, sectoral unemployment, and trade in resources. *Review of Development Economics* 4: 39–50.

Chao, C.C., J. Lafargue, and E.S.H. Yu. 2006. Public inputs, urban unemployment and welfare in a developing economy. *Asia Pacific Journal of Accounting and Economics* 13: 141–151.

Chau, N.H., and M.A. Khan. 2001. Optimal urban employment policies: Notes on Calvo and Quibria. *International Economic Review* 42: 557–568.

Chaudhuri, S. 2001. Foreign capital inflow, non-traded intermediary, urban unemployment, and welfare in a small open economy. *Pakistan Development Review* 40: 225–235.

Chaudhuri, S. 2003. How and how far to liberalize a developing economy with informal sector and factor market distortions. *Journal of International Trade and Economic Development* 12: 403–428.

Chaudhuri, T.D., and M.A. Khan. 1984. Educated unemployment, educational subsidies and growth. *Pakistan Development Review* 23: 395–409.

Chaudhuri, S., and U. Mukhopadhyay. 2002. Removal of protectionism, foreign investment and welfare in a model of informal sector. *Japan and the World Economy* 14: 101–116.

Chaudhuri, S., and U. Mukhopadhyay. 2003. Free education policy and trade liberalization: Consequences on child and adult labour markets in a small open economy. *Journal of Economic Integration* 18: 336–359.

Choi, J. 1999. Factor growth, urban unemployment and welfare under variable returns to scale. *International Economic Journal* 14: 17–34.

Choi, J., and E.S.H. Yu. 1993. Technical progress terms of trade and welfare in a mobile capital Harris–Todaro model. In *Economic theory and international trade: Essays in memoriam of J. Trout Rader*, ed. W. Neuefeind and R. Reizman. New York: Springer.

Choi, J., and E.S.H. Yu. 2006. Industrial targeting and non-shiftable capital in the Harris–Todaro model. *Review of International Economics* 14: 1–12.

Chow, Y., and J. Zeng. 2001. Foreign capital in a neoclassical model of growth. *Applied Economics Letters* 8: 613–615.

Corden, W.M., and R. Findlay. 1975. Urban unemployment, intersectional capital mobility and development policy. *Economica* 42: 59–78.

Feldmann, D.H. 1989. The trade-off between GNP and unemployment in a dual economy. *Southern Economic Journal* 56: 46–55.

Fields, G.S. 1975. Rural-urban migration, urban unemployment and job-search activity in LDCs. *Journal of Development Economics* 2: 165–187.

Fields, G.S. 1989. On-the-job search in a labor market model: Ex-ante choices and ex-post outcomes. *Journal of Development Economics* 30: 159–178.

Fields, G.S. 1990. Labour market modelling and the urban informal sector: Theory and evidence. In *The informal sector and evidence revisited*, ed. D. Turnham. Paris: OECD.

Fields, G.S. 1997. Wage floors and unemployment: A two-sector analysis. *Labour Economics* 4: 85–92.

Fields, G.S. 2005a. A welfare economic analysis of labor market policies in the Harris–Todaro model. *Journal of Development Economics* 76: 127–146.

Fields, G.S. 2005b. *A guide to multisector labor market models*, Social protection discussion paper, no. 0505. Washington, DC: World Bank.

Grinols, E.L. 1991. Unemployment and foreign capital: The relative opportunity costs of domestic labor and welfare. *Economica* 62: 59–78.

Grossman, G.M. 1983. Partially mobile capital: A general approach to two sector trade theory. *Journal of International Economics* 15: 1–17.

Gupta, M.R. 1988. Migration, welfare, inequality and the shadow-wage. *Oxford Economic Papers* 40: 477–486.

Gupta, M.R. 1993. Rural–urban migration, informal sector and development policies: A theoretical analysis. *Journal of Development Economics* 41: 137–151.

Gupta, M.R. 1995. Tax on foreign capital income and wage subsidy to the urban sector in the Harris–Todaro model. *Journal of Development Economics* 47: 469–479.

Gupta, M.R. 1997a. Foreign capital and the informal sector: Comments on Chandra and Khan. *Economica* 64: 353–363.

Gupta, M.R. 1997b. Informal sector and informal capital market in a small open less- developed economy. *Journal of Development Economics* 52: 409–428.

Gupta, K., and M.R. Gupta. 1998. Foreign enclaves and economic development: A theoretical analysis. *Journal of Economics* 67: 317–336.

Harberger, A.C. 1971. On measuring the social opportunity cost of labour. *International Labour Review* 103: 559–579.

Harris, J.R., and M. Todaro. 1970. Migration, unemployment and development: A two sector analysis. *American Economic Review* 40: 126–142.

Jha, R. and Whalley, J.R. 2003. *Migration and pollution*, Working paper, no. 20034. Department of Economics, University of Western Ontario.

Johnson, H.G. 1971. *The two-sector model of general equilibrium*, Yrjö Jahnsson Lectures. Chicago: Aldine–Atherton.

Jones, R.W. 1965. The structure of simple general equilibrium models. *Journal of Political Economy* 73: 557–572.

Jones, R.W. 1971. Distortions in factor markets and the general equilibrium model of production. *Journal of Political Economy* 79: 437–459.

Jones, R.W., and S. Marjit. 1992. International trade and endogenous production structures. In *Economic theory and international trade: Essays in memoriam of J. Trout Rader*, ed. W. Neuefeind and R. Reizman. New York: Springer.

Kar, S., and S. Marjit. 2001. Informal sector in general equilibrium: Welfare effects of trade policy reforms. *International Review of Economics and Finance* 10: 289–300.

Khan, M. Ali. 1979. A multisectoral model of a small open economy with non-shiftable capital and imperfect labor mobility. *Economic Letters* 2: 369–375.

Khan, M. Ali. 1980a. The Harris–Todaro hypothesis and the Heckscher–Ohlin–Samuelson trade model:

A synthesis. *Journal of International Economics* 10: 527–547.

Khan, M. Ali. 1980b. Dynamic stability, wage subsidies and the generalized Harris–Todaro model. *Pakistan Development Review* 19: 1–24.

Khan, M. Ali. 1982a. Social opportunity costs and immiserizing growth: Some observations on the long run returns versus and short. *Quarterly Journal of Economics* 96: 353–362.

Khan, M. Ali. 1982b. Tariffs, foreign capital and immiserizing growth with urban unemployment and specific factors of production. *Journal of Development Economics* 10: 245–256.

Khan, M. Ali. 1991. Ethnic groups and the Heckscher–-Ohlin–Samuelson trade model. *Economic Theory* 1: 355–371.

Khan, M. Ali. 1992. On measuring the social opportunity cost of labour in the presence of tariffs and an informal sector. *Pakistan Development Review* 31: 535–562.

Khan, M. Ali. 1993. Trade and development in the presence of an informal sector: A four factor model. In *Capital investment and development*, ed. K. Basu, M. Majumdar, and T. Mitra. Oxford: Basil Blackwell.

Khan, M. Ali, and T.D. Chaudhuri. 1985. Development policies in LDCs with several ethnic groups – A theoretical analysis. *Zeitschrift fur Nationalökonomie* 45: 1–19.

Khan, M. Ali, and P. Lin. 1982. Sub-optimal tariff policy and gains from trade with urban unemployment. *Pakistan Development Review* 21: 105–126.

Khan, M. Ali, and S.N.H. Naqvi. 1983. Capital markets and urban unemployment. *Journal of International Economics* 15: 367–385.

Magee, S.P. 1976. *International trade and distortions in factor markets*. New York/Basle: Marcel-Dekker.

Marjit, S. 1991. Agro-based industry and rural–urban migration: A case for an urban employment subsidy. *Journal of Development Economics* 35: 393–398.

Marjit, S. 2003. Economic reform and informal wage – A general equilibrium analysis. *Journal of Development Economics* 72: 371–378.

Marjit, S., and H. Beladi. 2003. Possibility or impossibility of paradoxes in the small country Harris–Todaro framework. *Journal of Development Economics* 72: 379–385.

Moene, K.O. 1988. A reformulation of the Harris–Todaro mechanism with endogenous wages. *Economics Letters* 27: 387–390.

Moene, K.O. 1992. Poverty and land ownership. *American Economic Review* 82: 52–64.

Neary, J.P. 1978. Dynamic stability and the theory of factor market distortions. *American Economic Review* 68: 672–682.

Neary, J.P. 1981. On the Harris–Todaro model with intersectoral capital mobility. *Economica* 48: 219–234.

Panagariya, A., and P. Succar. 1986. The Harris–Todaro model and economies of scale. *Southern Economic Journal* 52: 986–998.

Partridge, M.D., and D.S. Rickman. 1997. Has the wage-curve nullified the Harris–Todaro model? Further US evidence. *Economics Letters* 54: 277–282.

Portes, A., et al., eds. 1989. *The informal economy: Studies in advanced and less developed countries*. Baltimore: Johns Hopkins University Press.

Quibria, M.G. 1988. The Harris–Todaro model, trade unions and the informal sector: A note on Calvo. *International Economic Review* 29: 557–563.

Rauch, J. 1991. Modeling the informal sector formally. *Journal of Development Economics* 35: 33–47.

Rauch, J. 1993. Economic development, urban underemployment and income inequality. *Canadian Journal of Economics* 26: 901–918.

Sato, Y. 2004. Migration, frictional unemployment, and welfare-improving labor policies. *Journal of Regional Science* 44: 773–793.

Sen, P., A. Ghosh, and A. Barman. 1997. The possibility of welfare gains with capital inflows in a small tariff-ridden economy. *Economica* 64: 345–352.

Srinivasan, T.N., and J. Bhagwati. 1975. Alternative policy rankings in a large open economy with sector-specific minimum wages. *Journal of Economic Theory* 11: 356–371.

Srinivasan, T.N., and J. Bhagwati. 1978. Shadow prices for project selection in the presences of distortions: Effective rates of protection and domestic resource costs. *Journal of Political Economy* 86: 91–116.

Stiglitz, J.E. 1974. Alternative theories of wage determination and unemployment in LDC's: The labor-turnover model. *Quarterly Journal of Economics* 88: 194–227.

Stiglitz, J.E. 1976. The efficiency wage hypothesis, surplus labor, and the distribution of income in the LDCs. *Oxford Economic Papers* 28: 185–207.

Stiglitz, J.E. 1977. Some further remarks on cost–benefit analysis. In *Project evaluation*, ed. H. Schwartz and R. Berney. Washington, DC: Inter-American Development Bank.

Stiglitz, J.E. 1982. The structure of labor markets and shadow prices in LDCs. In *Migration and the labor market in developing countries*, ed. R.H. Sabot. Boulder: Westview Press.

Suits, D.B. 1985. US farm migration: An application of the Harris–Todaro model. *Economic Development and Cultural Change* 34: 815–828.

Todaro, M.P. 1968. An analysis of industrialization: Employment and unemployment in LDCs. *Yale Economic Essays* 8: 329–492.

Todaro, M.P. 1969. A model of labor migration and urban unemployment in less developed countries. *American Economic Review* 59: 138–148.

Yabuuchi, S., and H. Beladi. 2001. Urban unemployment, informal sector and development policies. *Journal of Economics* 74: 301–314.

Yabuuchi, S., H. Beladi, and G. Wei. 2005. Foreign investment, urban unemployment, and informal sector. *Journal of Economic Integration* 20: 123–138.

Harrod, Roy Forbes (1900–1978)

Walter Eltis

Keywords

Acceleration principle; Cost-push inflation; Domar, E. D.; Dynamic theory; Effective demand; Employment multiplier; Firm, theory of; Harrod, R. F.; Harrod–Domar growth model; Harrod-neutral technical progress; Imperfect competition; Inflation; International finance institutions; Keynesianism; Knife-edge; Marginal revenue curve; Market share; Monopolistic competition; Natural rate of growth; Neoclassical growth theory; Warranted rate of growth

JEL Classifications

B31

Roy Harrod was born in February 1900 and died in 1978. His father, Henry Dawes Harrod, was a businessman and author of two historical monographs. His mother, Frances (née Forbes-Robertson) was a novelist, and sister of the notable Shakespearean actor-manager, Sir Johnson Forbes-Robertson. Henry Harrod's business failed in 1907, but Roy won a scholarship to St Paul's School in 1911 and a King's Scholarship to Westminster in 1913. He became Head of his House, and in 1918 won a scholarship in history to New College, Oxford, his father's college. He enlisted in September 1918 and was commissioned in the Royal Field Artillery, but the war ended before his training was completed.

He went up to Oxford in early 1919 and first read Literae Humaniores (Classical Literature, Ancient History and Philosophy). He might well have devoted his career to academic philosophy, and he valued his publications in that subject more highly than his seminal contributions to economics. He has remarked that significant economic problems have only attracted the attention of profound thinkers for about 200 years, and interest in them might well disappear in another 200. In contrast, deep thought has been devoted to the great philosophical problems (such as the validity of inductive methods of thought) for more than 2,000 years and new contributions will be read for so long as civilized life remains. But his philosophy tutor at New College, H.W.B. Joseph, deterred him from devoting his life to that subject, by reacting extremely negatively to his essays. Harrod has left an account of a seminar on Einstein's theory of relativity in Oxford in 1922 where Joseph drew attention to a few terminological problems and believed this had undermined the theory. Einstein's theory of relativity survived, but Harrod was persuaded not to pursue a career in academic philosophy. In later years he published in the distinguished philosophical journal, *Mind*, and his *Foundations of Inductive Logic* (1956a) has received serious critical attention from philosophers as distinguished as A.J. Ayer (1970), but his main scholarly work was not to be in philosophy.

He followed his first class honours in Literae Humaniores in 1922 with a first class in modern history just one year later, and in 1923 Christ Church, Oxford, elected him to a Tutorial Fellowship (confusingly described as a studentship in that college) to teach the novel subject, economics, which was to be part of Oxford's new Honour School of Politics, Philosophy and Economics.

Harrod was allowed two terms away from Oxford so that he could learn enough economics to teach it, and it was suggested that he might spend this time in Europe, but he first went to Cambridge where he attended a wide range of lectures and wrote weekly essays on money and international trade for John Maynard Keynes. He was equally fortunate when he returned to Oxford, for while he was critically discussing the economics essays of Christ Church's undergraduates he was himself writing weekly microeconomic essays for the Drummond Professor of Political Economy, Francis Ysidro Edgeworth.

In addition to his new academic work Harrod took a notable part in the administration of his college (where he was Senior Censor in 1929–31, the most responsible office a student of Christ Church can be called upon to discharge), and also the university where he was elected to

Oxford's governing body (the Hebdomadal Council) in 1929 before he was 30. In the university and in Christ Church, he fought powerful campaigns on behalf of Professor Lindemann (subsequently Lord Cherwell) who held Oxford's Chair of Experimental Philosophy (Physics), and became principal scientific adviser to Winston Churchill's wartime government and a member of his post-war cabinet.

By 1930 his economics had developed to the point where he was able to publish his first important and original contribution, 'Notes on Supply', in which he was the first 20th-century economist to derive the marginal revenue curve. This should have appeared in 1928 to produce a claim for international priority, but Keynes, the editor of the *Economic Journal,* sent the article to Frank Ramsey, who first believed there were difficulties with the argument. He subsequently appreciated that his objections rested on a misunderstanding, but Harrod's new contribution was less startling in 1930 than it would have been in 1928. He followed this initial contribution to the imperfect competition literature with an important article, 'Doctrines of Imperfect Competition' (1934), in which he summarized the essential elements of the new theories of Edward Chamberlin and Joan Robinson.

During the 1930s Harrod frequently stayed with Keynes and he was increasingly drawn into the group of brilliant young economists which included Richard Kahn and Joan Robinson, who were helping him develop the new theories which culminated in *The General Theory of Employment, Interest and Money.* Harrod had written a number of important and influential articles in the press advocating new reflationary policies in the early 1930s, and these together with his extension of Kahn's employment multiplier to international trade in his *International Economics* (1933b) prompted Joseph A. Schumpeter to write in 1946 in his obituary article on Keynes, 'Mr Harrod may have been moving independently toward a goal not far from that of Keynes, though he unselfishly joined the latter's standard after it had been raised'.

Shortly after the *General Theory* appeared, Harrod published *The Trade Cycle* (1936a) in which he developed some of the dynamic implications of the new theory of effective demand. The conditions where output would grow were a central theme in Adam Smith's *The Nature and Causes of the Wealth of Nations,* and it had been much analysed in the great 19th-century contributions of Malthus, Ricardo, Mill and Marx, but the long-term dynamic implications of immediate changes to particular economic variables received virtually no attention in the neoclassical work that followed the marginal revolution. In the *General Theory* Keynes mostly went no further than to work through completely the immediate effects *on a formerly stationary economy* of a variety of disturbances such as an excess of the saving which would occur at full employment over the investment businessmen considered it prudent to undertake. Harrod went a vital step further and showed what could be expected to occur if saving was *permanently high* in relation to *the long-term opportunity to invest.* In 1939 he followed *The Trade Cycle* with 'An Essay in Dynamic Theory' (1939c), and after the war he developed his growth theory further in the book, *Towards a Dynamic Economics* (1948a). Important articles followed including a 'Second Essay in Dynamic Theory' (1960a), and 'Are Monetary and Fiscal Policies Enough?' (1964a). It is almost certainly because of Harrod's rediscovery of growth theory in the 1930s and his notable contributions to it that Assar Lindbeck, the Chairman of the Nobel Prize Committee, chose to state that he was among those who would have been awarded a Nobel Prize in economics if he had lived a little longer. The nature of Harrod's original contribution and the gradual evolution of his theory from 1939 to 1964 are set out in the second part of this article. The detailed technical characteristics of Harrod's growth model are the subject of Eltis (1987).

In the Second World War Harrod's friendship with Lindemann and his increasing distinction as an economist led to an invitation to join the Statistical Department of the Admiralty (S Branch) which Churchill set up when he again became First Lord in 1939. This moved to Downing Street when Churchill became Prime Minister in 1940, but Harrod did not have a particular talent for detailed statistical work and he developed an increasing interest in the international financial institutions, the International Monetary Fund and the World Bank, which would need to be set up as soon as the war was won, and from 1942 onwards he pursued this work in Christ Church. In the immediate post-war years he took a strong interest in national politics, and stood for Parliament unsuccessfully as a Liberal in the general election of 1945 and for a time he was a member of that party's Shadow Cabinet. He had served on Labour Party committees before the war, and in the 1950s with Churchill's support he unsuccessfully sought adoption as a Conservative parliamentary candidate: his economic advice was warmly welcomed by Harold Macmillan, Conservative Prime Minister in 1957–63. Harrod received the honour of knighthood in 1959 in recognition of his public standing and his notable academic achievements in the pre-war and post-war decades.

He had succeeded Keynes as editor of the *Economic Journal* in 1945, and in partnership with Austin Robinson (who looked after the book reviews) he sustained its reputation and quality until his retirement from the editorship in 1966.

His own post-war academic work included important contributions in three areas. In addition to the continuing development and refinement of his pre-war work on dynamic theory, he published extensively on the theory of the firm and on international monetary theory which had been his particular concern during the war.

The Oxford Economists' Research Group had begun to meet prominent British industrialists before the war. A group of Oxford economists which generally included Harrod invited individual industrialists to dine in Oxford, and after dinner they were questioned extensively on the considerations which actually influenced their decisions. This led to the publication of a number of much cited articles and the book, *Oxford Studies in the Price Mechanism* (Wilson and Andrews 1951) to which Harrod himself did not contribute. Propositions which emanated from these dinners included the notion that businessmen took little account of the rate of interest in their investment decisions, and that they did not seek to profit maximize, but priced instead by adding a margin they considered satisfactory to their average or 'full' costs of production. In his important articles, 'Price and Cost in Entrepreneurs' Policy' (1939b) and 'Theories of Imperfect Competition Revised' (1952a), Harrod set out a theoretical account of how firms price in which industrialists follow something like these procedures. Their object is especially to achieve a high market share, and by setting prices low enough to deter new entry they actually succeed in maximizing their long-run profits and avoid the excess capacity that Chamberlin and Joan Robinson had considered an inevitable consequence of monopolistic or imperfect competition. This attempt to reconcile the 'rules of thumb' that the businessmen revealed with the propositions of traditional theory was more highly regarded outside Oxford than some of the books and articles in the new tradition.

His work on the world's international monetary problems occupied a good deal of his time and attention in the post-war decades. Keynes himself had considered the breakdown in international monetary relations a crucial element in the collapse of effective demand in so many countries in the 1930s, and he devoted much of the last years of his life to the creation of new institutions which would avoid a repetition of these disasters. Harrod believed he was continuing this vital work when he devoted much thought and energy to these questions. He arrived at the conclusion that there was bound to be some inflation in a world which was successfully pursuing Keynesian policies, and that the liquidity base of the world's financial system was bound to become inadequate if the price of gold failed to rise with other prices. He believed that underlying world liquidity which rested on gold in the last resort must be allowed to rise in line with the international demand for money. He therefore came to focus on the price of gold, and in his book, *Reforming the World's Money* (1965), he proposed that a substantial increase in the price of gold would be needed if subsequent international monetary crises were to be avoided. Harry Johnson (1970) has summarized his contribution to this debate.

Harrod took a great interest in actual developments in the United Kingdom economy, and published seven books and collections of articles in the first two postwar decades which were directly concerned with the policies Britain should follow. There was in addition an immense range of articles in the academic journals, the bank reviews and the press on these questions, not to mention monthly stockbrokers' letters for Phillips and Drew. Harrod argued strongly and powerfully that nothing was to be gained by running the economy below full employment, which meant an unemployment rate of less than two per cent in the 1950s and the 1960s. In the late 1950s he was deeply concerned that the removal of import controls would render it increasingly difficult for Britain to pursue such Keynesian policies, and he was a vigorous opponent of European Common Market entry. He attached more significance than some distinguished Keynesians to holding down inflation but he published statistics in *Towards a New Economic Policy* (1967a) to show that in Britain this had tended to be faster when the economy was in recession than when output was allowed to expand. He argued therefore that deflationary policies could play no useful role in policies to control the rate of cost inflation, which he considered the essential element in inflation in Britain. Policy swung sharply away from this Keynesian tradition in the last years of his life, and he wrote a final letter to *The Times* on 21 July 1976 in which he praised the economics of Tony Benn and Peter Shore for their opposition to the Labour government's public expenditure cuts, for, 'To cut public spending when there is an undesirably high rate of unemployment is crazy'.

His advocacy of import controls and his adverse reaction to deflationary policies at all times might suggest that he was an economist of the Left, but his willingness to support each of the British political parties at various times underlines how his approach to economic and social problems cannot be typecast. The lines of policy he supported always followed directly from his understanding of the significance of the major interrelationships, and it was his belief that Keynesian theory (which he had so notably helped to refine and develop) provided the appropriate tools for the analysis of Britain's economic problems that led him towards the expansionist policies he so consistently advocated. But further theoretical and empirical relationships which he believed were equally well founded led him to advocate a series of social policies to which very right-wing labels can be attached.

Just before the 1959 election his article, 'Why I Shall Vote Conservative', in *The Sunday Times,* put forward the startlingly unfashionable argument that only the Conservatives would allow more money to go to the better off who had most to contribute to the future of Britain. Harrod's strong belief in the importance of the *quality* of the country's population stock (which, he held, mattered no less than the physical capital stock) lay behind this article. Harrod thought the quality of the population would be bound to deteriorate if the middle classes continued to have fewer children than the poor. He was a strong believer in the inheritance of every kind of ability, and a provocative conversational conclusion he drew was that in an ideal world one-third of Christ Church's much sought-after undergraduate places should be sold to the rich. Their children often had insufficient academic ability to perform well in examinations, but they had inherited abilities of other kinds which would take them to the highest positions, so they should go to Oxford first. Harrod's reasoning on the inheritance of ability and its implications is set out in detail in the Memorandum he submitted to the Royal Commission on Population in 1944. There he suggested that a difficulty in finding servants was one reason why the middle classes had fewer children. Among his suggestions to remedy this state of affairs was that Diplomas in Domestic Service should be established, and that it should become common practice for servants to have latch-keys and the same rights as their mistresses to enjoy social lives with no questions asked. His Memorandum reads strangely nowadays when it is widely regarded as unacceptable that any practical conclusions may be drawn from the proposition that human abilities are inherited. Harrod never hesitated to carry his arguments to their limits, and he always went where his reasoning took him, irrespective of the predictable reactions of others.

The unselfconsciousness of both his academic and his public writing comes out especially in his two biographical volumes, the official life of Keynes (commissioned by the executors) which he published in 1951 and *The Prof* (1959a), his personal sketch of Lord Cherwell. As well as providing magnificent accounts of their subjects from the standpoint of one who had known them intimately (and who profoundly understood the economic problems Keynes wrestled with), these books contain extensive autobiographical passages which will enable later generations to know more of Harrod than any biographer can begin to convey.

He ceased to lecture in Oxford in 1967 upon reaching the statutory retirement age of 67, but as a Visiting Professor he continued to teach in several distinguished North American Universities. He died in his Norfolk home in 1978 eleven years after his Oxford work came to an end.

Harrod's Revival of Growth Theory and His Contribution to Keynesian Macroeconomics

Harrod was intimately involved in the origins and development of Keynesian economics. As the galley proofs of the *General Theory* emerged from the printers from June 1935 onwards, copies were sent to Harrod, to Kahn and to Joan Robinson and with their assistance, Keynes rewrote extensively for final publication. Harrod helped to clarify the relationship between Keynes's new theory of the rate of interest and the then ruling neoclassical theory where this depended upon the intersection of ex ante saving and investment schedules. In the course of their correspondence, Harrod showed Keynes how well he understood the essence of the *General Theory* by setting out its novelty and its principal elements in ten lines on 30 August 1935: Your view, as I understand it is broadly this:-

$$\text{Volume of investment determined by}\left\{\begin{array}{l}\text{marginal efficiency of}\\ \text{capital schedule}\\ \text{rate of interest}\end{array}\right.$$

$$\text{Rate of investment determined by}\left\{\begin{array}{l}\text{liquidity preference}\\ \text{schedule}\\ \text{quantity of money}\end{array}\right.$$

$$\text{Volume of employment determined by}\left\{\begin{array}{l}\text{volume of investment}\\ \text{multiplier}\end{array}\right.$$

$$\text{Value of multiplier determined by}\{\text{propensity to save}$$

Keynes responded, 'I absolve you completely of misunderstanding my theory. It could not be stated better than on the first page of your letter.'

Almost immediately after the appearance of the *General Theory,* Harrod published *The Trade Cycle* (1936a) which contained for the first time in the Keynesian literature the concept of an economy growing at a steady rate. Keynes wrote of it to Joan Robinson on 25 March 1937, 'I think he has got hold of some good and important ideas. But, if I am right, there is one fatal mistake', and to Harrod himself on March 31, 'I think that your theory in the form in which you finally enunciate it is not correct, being fatally affected by a logical slip in the argument.' Harrod replied devastatingly on April 6th, 'There is no slip . . . The fact is that you in your criticism are still thinking of once over changes and that is what I regard as a static problem. My technique relates to steady growth.' Harrod's slip was in fact the first step towards the reinstatement of growth theory into mainstream economic analysis.

Harrod convinced Keynes, who on 12 April congratulated him for 'having invented so interesting a theory', but with the reservation, 'I should doubt whether any reader who has not talked or corresponded with you could be aware that the whole of the last half of the book was intended to be in relation to a moving base of steady progress.' Keynes added that it was vital that Harrod carry his ideas further and restate them more comprehensibly.

Harrod made important progress in the next 15 months, and on 3 August 1938 he sent Keynes a preliminary draft of the article, 'An Essay in Dynamic Theory', and wrote in his accompanying letter,

> my re-statement of the dynamic theory ... is, I think, a great improvement on my book . . . I have been throwing out hints in a number of places of the possibility of formulating a simple law of growth and I want to substantiate the claim. It is largely based on the ideas of the general theory of employment; but I think it gets us a step forward.

A lengthy correspondence then developed between Harrod and Keynes in which the two most original elements in Harrod's contribution which later excited much interest and controversy in the economics profession were extensively discussed.

Harrod's principal innovation was the invention of a *moving equilibrium growth path* for the economy, and he described this as the 'warranted' line of growth.

Harrod had perceived before he wrote *The Trade Cycle* that there was a fundamental contradiction between the assumptions prevalent in the microeconomic theory of the firm and industry, to which he had made notable contributions, and the new Keynesian macroeconomics. In the theory of the firm, long-term investment was zero, for firms had no motivation to undertake further investment once they were in long-period equilibrium. But the new Keynesian macroeconomics required that there be net investment by firms or the government whenever there was any net saving in the macroeconomy. A theory compatible with both macro and microeconomic equilibrium therefore required that firms invest all the time, so that they can continually absorb total net saving. Harrod's formulation of the warranted rate of growth, his novel discovery, was an attempt to set out this necessary equilibrium growth path that industrial and commercial investment decisions must all the time follow in order to achieve a complete economic equilibrium.

Harrod's moving equilibrium or warranted growth path required that saving (of *s* per cent of the national income) be continually absorbed into investment, so he asked the question: at what rate of growth will firms all the time choose to invest the *s* per cent of the national income, which equilibrium growth requires? To answer this question, he made use of the acceleration principle or 'the relation', as he called it, that firms need say C_r units of additional capital to produce an extra unit of output. It follows from these premises that the warranted rate of growth of output will be s/C_r per cent per annum. Since each rise in output by 1 unit entails that C_r extra units be invested, a rise in output by s/C_r per cent of the national income will call for an equilibrium investment of C_r times this, which is precisely *s* per cent of the national income, the ratio of *ex ante* saving in the national income. In Harrod's examples at this time, he suggested a typical *s* of 10 per cent of the national income and a C_r of 4, to produce a warranted rate of growth of 2.5 per cent.

This idea that if there is continual saving, then equilibrium entails a continual geometric growth in production came as a considerable surprise to Keynes and the other members of the 'circus'. As Harrod had already explained in April 1937,

> The static system provides an analysis of what happens where there is no increase [in output] which entails (as in Joan Robinson's long-period analysis) that saving = 0. Now I was on the lookout for a steady rate of advance, in which the rates of increase would be mutually consistent.

But Harrod's second discovery had equally radical implications. Suppose the actual growth of output is marginally above the equilibrium or warranted rate of growth. In Harrod's numerical example with *s* 10 per cent and C_r 4, it can be supposed that output actually grows 0.1 per cent faster than the warranted rate, that is by 2.6 per cent instead of 2.5 per cent. Then with 2.6 per cent output growth, the acceleration principle or relation will entail that 4 times 2.6 per cent be added to the capital stock, so that ex ante investment is 10.4 per cent of the national income. With ex ante saving limited to 10.0 per cent, the 0.1 per cent excess of actual growth over warranted growth then produces an excess in ex ante investment over ex ante saving of 0.4 per cent of the national income. Any excess in ex ante investment over ex ante saving will be associated with extra expansion of the national income according to the economics of the *General Theory.* Thus, if the actual rate of growth exceeds the warranted rate of s/C_r per cent, the tendency will be for actual growth to rise and rise, for as soon as actual growth rises from 2.6 to say 3 per cent, required investment will rise further to 4 times 3 per cent which equals 12 per cent and so exceed the 10 per cent savings ratio by a still greater margin. Conversely, when actual growth comes out at a rate just short of the warranted 2.5 per cent, ex ante investment will be below the 10 per cent savings ratio, which will cause the rate of growth to decline. This second discovery, which became known as Harrod's knife- edge, was therefore that any rate of growth in excess of the equilibrium or warranted path he had discovered would set off a continual acceleration of growth, while any shortfall would set off deceleration. He wrote to Keynes of this discovery on 7 September 1938:

> If in static theory producers produce too little, they will be well satisfied with the price they get and feel happy; but this is not taken to be the *right* amount of output; they will be stimulated to produce more. The equilibrium output is taken to be that which *just* satisfies them and induces them to go on as before. Similarly the warranted rate [of growth] is that which just satisfies them and leaves them going on as before. The difference between the warranted rate and the old equilibrium (i.e. the difference between dynamic and static theory) is, in my view, that if they produce above the warranted rate, they will be more than satisfied and be stimulated, and conversely, while in the case of equilibrium in static conditions the opposite happens. The 'field' round the [static] equilibrium contains centripetal, that round the warranted centrifugal forces.

It took Keynes time to absorb Harrod's startling discovery. On 19 September he proposed a counter-example in which C_r was merely one-tenth, while s was also one-tenth. With this counter-example, a deviation of output by a small amount from the warranted path, say by δ x, which would raise planned investment above the level at which it would otherwise be by $C_r\ \delta\ x$ would merely raise this by 0.10 δ x, which would equal the rise in planned saving of $s\ \delta\ x,$ which would also come to 0.10 $\delta\ x,$ so there would be no tendency towards an explosive growth in effective demand. This would grow explosively if C_r was one-ninth (in which case planned investment would rise by 0.11 $\delta\ x$ and saving by only 0.10 δ x) but the further growth of output would be damped if C_r was merely one-eleventh, so, Keynes insisted, 'neutral, stable or unstable equilibrium' are equally likely.

Harrod protested on 22 September, 'it is absurd to suppose extra capital required [C_r] only $\frac{1}{10}$of annual output, when the capital required in association with the pre-existent level of incomes in England today is 4 or 5 times annual output'. The probability that C_r would exceed s so that ex ante investment would rise by more than ex ante saving in order to produce instability was therefore overwhelming.

But several qualifications emerged. In comparing the increase in ex ante investment to the increase in ex ante saving following a small deviation of output from the warranted rate:

1. The relevant marginal capital coefficient (C_r) which determines how much planned investment will rise is the net new requirement of *induced* investment. In so far as investment decisions are autonomous of short-term fluctuations in output, the relevant C_r will be lower than the economy's overall capital output ratio.
2. The relevant coefficient which determines the increase in planned saving is the *marginal* and not the average propensity to save. Planned saving will rise more where output deviates upward from the warranted rate, the greater is the marginal propensity to save in relation to the average propensity.

The circumstances that could produce a stable upward deviation of growth from the warranted rate and the avoidance of Harrod's knife-edge are therefore a very high marginal propensity to save in combination with a situation where most investment is autonomous so that the induced investment coefficient, C_r, is considerably less than 1. In 'An Essay in Dynamic Theory', Harrod covered this possibility with the caveat, 'when long-range capital outlay is taken into account . . . the attainment of a neutral or stable equilibrium of advance may not be altogether improbable in certain phases of the cycle'. The possibility he had in mind here is that in the early stages of a cyclical recovery there may be so much excess industrial capacity that C_r will be quite low for a time, and therefore quite possibly lower than the marginal propensity to save. But in general any deviation of growth from the warranted line of advance would raise ex ante investment by a greater margin than ex ante saving with the result that the rate of growth would deviate further.

In addition to establishing the existence of the warranted line of advance and its instability, Harrod had to define the equilibrium investment behaviour by businesses which would actually lead to expansion at the requisite rate. In his 1939 article he omitted to offer any behavioural rule but simply asserted that the warranted rate was 'that rate of growth which, if it occurs, will leave all parties satisfied that they have produced neither more nor less than the right amount'. That is no more than a description of equilibrium growth, and much the same can be said of his definition of the warranted rate in *Towards a Dynamic Economics* (1948a) as 'that over-all rate of advance which, if executed, will leave entrepreneurs in a state of mind in which they are prepared to carry on a similar advance'. It was only in the article 'Supplement on Dynamic Theory' (1952b) that Harrod arrived at a behavioural assumption that matched his algebraic formulation of the warranted rate:

> Let the representative entrepreneur on each occasion of giving an order repeat the amount contained in his order for the last equivalent period, adding thereto an order for an amount by which he judges his existing stock to be deficient, if he judges it to be deficient, or subtracting therefrom the amount by which he judges his stock to be redundant, if he does so judge it.

With that assumption an economy which once achieves growth at the warranted rate will sustain it, while any upward or downward deviations will lead to still greater deviations wherever C_r exceeds the marginal propensity to save.

But it emerged by 1964, when Harrod published 'Are Monetary and Fiscal Policies Enough?', that even that assumption fails to define growth at the warranted rate, for it must also be assumed that the representative entrepreneur will expand at a rate of precisely s/C_r when he judges his capital to be neither deficient nor redundant. This requires an expectation by the representative entrepreneur that his market will grow at a rate of precisely s/C_r. Hence the full requirement for growth along Harrod's warranted equilibrium path is that entrepreneurs expect growth at this rate and expand and continue to expand at that rate so long as their capital stock continues to grow in line with their market so that it is neither deficient nor redundant. They will of course increase their rate of expansion if their capital should prove deficient, and curtail it if part of their stock becomes redundant.

The warranted rate of growth and its instability were Harrod's great innovations. From 1939 onwards he contrasted this equilibrium rate with the natural rate of growth, 'the rate of advance which the increase of population and technological improvements allow', which was entirely independent of the warranted rate. Harrod defined the rate of technical progress more precisely in 1948 as the increase in labour productivity 'which, at a constant rate of interest, does not disturb the value of the capital coefficient'. This then entered the language of economics as Harrod- neutral technical progress, which, together with growth in the labour force, determines the natural rate of growth, that is, the rate at which output can actually be increased in the long run. This raised few theoretical problems in 1939, and there was nothing novel in the proposition that long-term growth must depend on the rate of increase of the labour force and technical progress. Keynes himself had said as much several years earlier in 'Economic Possibilities for our Grandchildren' (1930). But the contrast between this natural rate and Harrod's innovatory warranted rate offered entirely new insights.

If the warranted rate exceeds the feasible natural rate, the achievement of equilibrium growth must be impractical because the economy cannot continue to grow faster than the natural rate. It must deviate downwards from the warranted rate towards the natural rate far more than it deviates upwards with the result that 'we must expect the economy to be prevailingly depressed'. If the natural rate is greater, output will tend to deviate upwards towards the natural rate with the result that the economy should enjoy 'a recurrent tendency towards boom conditions'.

Keynes's own reaction to the dichotomy between the warranted and natural rates was characteristically (his letter to Harrod on 26 September 1938) that the warranted rate always exceeded the natural:

> In actual conditions ... I suspect the difficulty is, not that a rate in excess of the warranted is unstable, but that the warranted rate itself is so high that with private risk-taking no one dares to attain it . . .
>
> I doubt if, in fact, the warranted rate –let alone an unstable excess beyond the warranted has ever been reached in USA and UK since the war, except perhaps in 1920 in UK and 1928 in USA. With a stationary population, peace and unequal incomes, the warranted rate sets a pace which a private risk-taking economy cannot normally reach and can never maintain.

That is characteristic Keynes, but Harrod had persuaded him to express his familiar analysis in the language of his new theory of growth. In the immediate post-war decades when full employment and creeping inflation prevailed, it was widely argued that the natural rate had come to exceed the warranted. The richness of Harrod's model is demonstrated by its ability to illuminate both kinds of situation.

Evsey Domar's growth model which has a good deal in common with Harrod's was published seven years after 'An Essay in Dynamic Theory', and a considerable literature emerged in the next 15 years on the stability conditions and other important features of what came to be known as the Harrod–Domar growth model. This is elegantly summarized by Frank Hahn and Robin Matthews in their celebrated 1964 survey article.

The development of neoclassical growth theory in the 1950s led to an increasing realization that the warranted and natural growth rates could be equated by an appropriate rate of interest. If the warranted rate was excessive so that oversaving led to slump conditions, a lower interest rate which raised C_r sufficiently would bring it down to the natural rate. Conversely the inflationary pressures that resulted from an insufficient warranted rate would be eliminated if higher interest rates reduced C_r sufficiently. If the real rate of interest and C_r responded in this helpful way, s/C_r, the warranted rate could always be brought into equality with the natural rate.

Harrod's response included his 'Second Essay in Dynamic Theory' (1960a), a title which underlines its significance. He proposed that there was an optimum real rate of interest r_n which would maximize utility, with a value of G_p/e, G_p being the economy's long-term rate of growth of labour productivity and e the elasticity of the total utility derived from real per capita incomes with respect to increases in these. If a one per cent increase in real per capita incomes raises per capita utility 0.5 per cent, e will be 0.5, and r_n the optimum rate of interest which maximizes utility will be $Gp/0.5$, viz. twice the rate of growth of labour productivity. If the marginal utility of income does not fall at all as real per capita incomes rise, per capita utility will grow one per cent when incomes rise one per cent so that e is unity, and r_n equals G_p. The more steeply the marginal utility of incomes fall, the more e will fall below unity, and the more the optimum real rate of interest, G_p/e, will exceed the rate of growth of labour productivity.

If a society actually seeks to establish the optimum rate of interest determined in this kind of way, the value of C_r will depend upon this optimum rate of interest, so it will not also be possible to use the rate of interest to equate the natural and warranted rates of growth in the manner the neoclassical growth models of, for instance, Robert Solow (1956) and Trevor Swan (1956) propose. There will therefore still be difficulties because the warranted rate of growth with real interest rates at their optimum level will not in general be equal to the natural rate. Therefore, as Harrod suggested in the final articles he published in 1960 and 1964, governments will have to run persistent budget deficits or surpluses if they are to avoid the difficulties inherent in discrepancies between the natural and the warranted rates of growth.

So Harrod remained a convinced Keynesian who continued to believe that a long-term imbalance between saving, the main determinant of the warranted rate, and investment opportunity would call for persistent government intervention. When that approach to economic policy again becomes fashionable, economists may learn a good deal from Harrod's later articles which have not yet received the same attention from the economics profession as his seminal work in the 1930s and the 1940s.

Selected Works

The 'Bibliography of the Works of Sir Roy Harrod', in *Induction, Growth and Trade: Essays in Honour of Sir Roy Harrod,* ed. W.A. Eltis, M. FG. Scott and J.N. Wolfe, Oxford: Oxford University Press, 1970, includes all the articles he published in books, journals and magazines from 1928 to 1969, and some of his most influential newspaper articles. The present list of selected works is confined to his books and academic articles in books and journals.

1930a. Notes on supply. *Economic Journal* 40: 232–241.

1930b. Progressive taxation and equal sacrifice. *Economic Journal* 40: 704–707.

1931. The law of decreasing costs. *Economic Journal* 41: 566–576. Addendum: 42: 490–492.

1933a. A further note on decreasing costs. *Economic Journal* 43: 337–341.

1933b. *International economics.* Cambridge: Cambridge University Press. 1st revised ed., 1939; 2nd revised ed., 1957; 3rd revised ed., mainly rewritten, 1974.

1934a. Professor Pigou's theory of unemployment. *Economic Journal* 44: 19–32.

1934b. Doctrines of imperfect competition. *Quarterly Journal of Economics* 48: 442–470.

1934c. The equilibrium of duopoly. *Economic Journal* 44: 335–337.

1934d. The expansion of credit in an advancing economy. *Economica* NS 1: 287–299. Rejoinders:1: 476–478; 2 (1935): 82–84.

1936a. *The trade cycle: An essay.* Oxford: Oxford University Press.

1936b. Utilitarianism revised. *Mind* 45: 137–156.

1936c. Imperfect competition and the trade cycle. *Review of Economics and Statistics* 18: 84–88.

1937. Mr. Keynes and traditional theory. *Econometrica* 5: 74–86.

1938. Scope and method of economics. *Economic Journal* 48: 383–412.

1939a. Modern population trends. *Manchester School of Economics and Social Studies* 10(1): 1–20. Rejoinder: April (1940): 47–58.

1939b. Price and cost in entrepreneurs' policy. *Oxford Economic Papers* 2: 1–11.

1939c. An essay in dynamic theory. *Economic Journal* 49: 14–33. Errata: 49: 377.

1939d. Value and capital by J.R. Hicks. *Economic Journal* 49: 294–300.

1942. Memory. *Mind* 51: 47–68.

1943. Full employment and security of livelihood. *Economic Journal* 53: 321–342.

1945. *Memorandum to the Royal Commission on equal pay for men and women.* Appendix IX in the Fourth Volume of Memoranda of Evidence. London: HMSO.

1946a. *A page of British folly.* London: Macmillan.

1946b. Price flexibility and employment. By Oscar Lange. *Economic Journal* 56: 102–107.

1946c. Professor Hayek on individualism. *Economic Journal* 56: 435–442.

1947a. A comment on R. Triffin's 'National Central Banking and The International Economy'. *Review of Economic Studies* 14(2): 95–97.

1947b. *Are these hardships necessary?* London: Rupert Hart-Davis.

1948a. *Towards a dynamic economics: Some recent developments of economic theory and their application to policy.* London: Macmillan.

1948b. The economic consequences of atomic energy. In *The atomic age*, ed. Sir Halley Stewart Lectures. London: Allen & Unwin.

1948c. The fall in consumption. *Bulletin of the Oxford University Institute of Statistics* 10: 162–167. Rejoinder: 10: 235–244; 10: 290–293.

1950. Memoranda (Submitted in August and December 1944). *Papers of the Royal Commission on population*, vol. 5. London: HMSO.

1951a. *The life of John Maynard Keynes.* London: Macmillan.

1951b. Notes on trade cycle theory. *Economic Journal* 61: 261–275.

1951c. *And so it goes on: Further thoughts on present mismanagement*. London: Rupert Hart-Davis.

1952a. Theory of imperfect competition revised. In *Economic essays,* ed. R.F. Harrod. London: Macmillan.

1952b. Supplement on dynamic theory. In *Economic essays,* ed. R.F. Harrod. London: Macmillan.

1952c. Currency appreciation as an anti-inflationary device: Comment. *Quarterly Journal of Economics* 66: 102–116.

1952d. *Economic essays.* London: Macmillan.

1952e. *The pound sterling.* Princeton Essays in International Finance No. 13. Princeton: Princeton University Press.

1953a. *The dollar.* London: Macmillan. 2nd ed. with new introduction, New York: The Norton Library, 1963.

1953b. Imbalance of international payments. *International Monetary Fund Staff Papers* 3: 1–46.

1953c. Foreign exchange rates: A comment. *Economic Journal* 63: 294–298.

1953d. Sir Hubert Henderson, 1890–1952. *Oxford Economic Papers* NS 5(Suppl): 59–64.

1953e. Full capacity vs. full employment growth: A comment on Pilvin. *Quarterly Journal of Economics* 6: 553–559.

1955. Les relations entre l'investissement et la population. *Revue économique* 6: 356–367.

1956a. *Foundations of inductive logic.* London: Macmillan.

1956b. The British boom, 1954–55. *Economic Journal* 66: 1–16.

1956c. Walras: A re-appraisal. *Economic Journal* 66: 307–316.

1957a. The common market in perspective. *Bulletin of the Oxford Institute of Statistics* 19: 51–55.

1957b. Review of international economic policy by J.E. Meade. *Economic Journal* 67: 290–295.

1957c. Clive Bell on Keynes. *Economic Journal* 67: 692–629.

1958a. *The pound sterling, 1951–58.* Princeton Essays in International Finance. Princeton: Princeton University Press.

1958b. The role of gold today. *South African Journal of Economics* 26: 3–13. Rejoinder: 27, March, 16–22.

1958c. *Policy against inflation.* London: Macmillan.

1958d. Questions for a stabilization policy in primary producing countries. *Kyklos* 11(2): 207–211.

1958e. Factor-price relations under free trade. *Economic Journal* 68: 245–255.

1959a. *The prof: A personal memoir of Lord Cherwell.* London: Macmillan.

1959b. Domar and dynamic economics. *Economic Journal* 69: 451–464.

1959c. Inflation and investment in underdeveloped countries. In *Economi, Politik, Samhälle: en bok Tillagnad Bertil Ohlin,* ed. J. Bergvall. Stockholm: Bokförlaget Folk och Samhälle.

1960a. Second essay in dynamic theory. *Economic Journal* 70: 277–293. Comment: 70: 851. Rejoinder: 72: 1009–1010.

1960b. New arguments for induction: Reply to Professor Popper. *British Journal for the Philosophy of Science* 10(40): 309–312.

1960c. Keynes' attitude to compulsory military service. *Economic Journal* 70: 166–167.

1960d. Evidence submitted to the Radcliffe Committee on the Working of the Monetary System, May 1958. *Principal memoranda of evidence,* vol. 3. London: HMSO.

1961a. The dollar problem and the gold question. In *The dollar in crisis,* ed. S.E. Harris. New York: Harcourt, Brace and World.

1961b. *Topical comment: Essays in dynamic economics applied.* London/New York: Macmillan/St Martin's Press.

1961c. The general structure of inductive argument. *Proceedings of the Aristotelian Society,* 1960–61: 41–56.

1961d. Real balances: A further comment. *Economic Journal* 71: 165–166.

1961e. A plan for increasing liquidity: A critique. *Economica* NS 28: 195–202.

1961f. The 'neutrality' of improvements. *Economic Journal* 71: 300–304.

1961g. Review of Sraffa's production of commodities by means of commodities. *Economic Journal* 71: 783–787.

1962a. Economic development and Asian regional cooperation. *Pakistan Development Review* 2: 1–22.

1962b. Dynamic theory and planning. *Kyklos* 15(3): 68–79.

1963a. *The British economy.* New York: McGraw Hill.

1963b. Themes in dynamic theory. *Economic Journal* 73: 401–21. Corrigendum: December 1963, 792.

1963c. Desirable international movements of capital in relation to growth of borrowers and lenders and growth of markets. In *International trade theory in a developing world,* ed. R.-F. Harrod. Hague/London: Macmillan.

1963d. Liquidity. In *World monetary reform,* ed. H.C. Grubel. Stanford: Stanford University Press.

1964a. Are monetary and fiscal policies enough? *Economic Journal* 74: 903–915.

1964b. *Plan to increase international monetary liquidity.* Brussels: European League for Economic Co-operation.

1964c. Comparative analysis of policy instruments. In *Inflation and growth in Latin America*, ed. W. Baer and I. Kerstenetzky. Homewood: Richard Irvin.

1964d. Retrospect on Keynes. In *Keynes' general theory,* ed. R. Lekachman. New York: Macmillan.

1965. *Reforming the world's money.* London/New York: Macmillan/St Martin's Press.

1966a. International liquidity. *Scottish Journal of Political Economy* 13: 189–204.

1966b. Optimum investment for growth. *In Problems of economic dynamics and planning: Essays in honour of Michael Kalecki.* Oxford: Pergamon Press.

1967a. *Towards a new economic policy.* Manchester: Manchester University Press.

1967b. Increasing returns. In *Monopolistic competition theory*: *Studies in impact: Essays in honour of Edward H. Chamberlin*, ed. R.E. Kuenne. New York: Wiley.

1967c. Methods of securing equilibrium. *Kyklos* 20(1): 24–33.

1967d. World reserves and international liquidity. *South African Journal of Economics* 35: 91–103.

1967e. Assessing the trade returns. *Economic Journal* 77: 499–511.

1968. What is a model? In *Value, capital and growth: Papers in honour of Sir John Hicks*, ed. J.N. Wolfe. Edinburgh: Edinburgh University Press.

1969. *Money.* London/New York: Macmillan/St Martin's Press.

1970a. *Sociology, morals and mystery.* Chichele Lectures, All Souls College. Oxford/London: Macmillan.

1970b. Reassessment of Keynes's views on money. *Journal of Political Economy* 78: 617–625.

1970c. Replacements, net investment, amortisation funds. *Economic Journal* 80: 24–31.

1972. Imperfect competition, aggregate demand and inflation. *Economic Journal* 82: 392–401.

1973. *Economic dynamics.* London/New York: Macmillan/St Martin's Press.

Bibliography

Ayer, A.J. 1970. Has Harrod answered Hume? In *Induction, growth and trade: Essays in honour of Sir Roy Harrod*, ed. W.A. Eltis, M.F.G. Scott, and J.N. Wolfe. Oxford: Oxford University Press.

Blake, R. 1970. A personal memoir. In *Induction, growth and trade: Essays in honour of Sir Roy Harrod*, ed. W.A. Eltis, M.F.G. Scott, and J.N. Wolfe. Oxford: Oxford University Press.

Domar, E. 1946. Capital expansion, rate of growth, and employment. *Econometrica* 14: 137–147.

Domar, E. 1947. Expansion and employment. *American Economic Review* 37: 34–55.

Eltis, W. 1987. Harrod–Domar growth model. In *The new Palgrave: A dictionary of economics*, ed. J. Eatwell, M. Milgate, and P. Newman, vol. 2. London: Macmillan.

Hahn, F.H., and R.C.O. Matthews. 1964. The theory of economic growth: A survey. *Economic Journal* 74: 779–902.

Johnson, H.G. 1970. Roy Harrod on the price of gold. In *Induction, growth and trade: Essays in honour of Sir Roy Harrod*, ed. W.A. Eltis, M.F.G. Scott, and J.N. Wolfe. Oxford: Oxford University Press.

Keynes, J.M. 1930. Economic possibilities for our grandchildren. In *The collected writings of John Maynard Keynes, Vol. IX*: *Essays in persuasion*. London: Macmillan, 1972.

Keynes, J.M. 1973. *The general theory and after* (Correspondence and Articles). Vols XIII and XIV of *The Collected Writings of John Maynard Keynes*. London: Macmillan.

Lindbeck, A. 1985. The prize in economic science in memory of Alfred Nobel. *Journal of Economic Literature* 23: 37–56.

Phelps-Brown, H. 1980. Sir Roy Harrod: A biographical memoir. *Economic Journal* 90: 1–33.

Schumpeter, J.A. 1946. John Maynard Keynes 1883–1946. *American Economic Review* 36: 495–518.

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

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

Wilson, T., and P.W.S. Andrews. 1951. *Oxford studies in the price mechanism*. Oxford: Oxford University Press.

Bibliographic Addendum

Besomi, D. 1999. *The making of Harrod's dynamics*. London/New York: Macmillan/St. Martin's Press.

Rampa, G., L. Stella, and A. Thirlwall, eds. 1998. *Economic dynamics, trade and growth: Essays on Harrodian themes*. London/New York: Macmillan/St. Martin's Press.

Young, W. 1989. *Harrod and his trade cycle group: The origins and development of the growth research programme*. London/New York: Macmillan/St. Martin's Press.

Harrod–Domar Growth Model

Walter Eltis

The Keynesian revolution led Roy Harrod (1939) and Evsey Domar (1946 and 1947) to work out the implications of permanent full employment. In *The General Theory of Employment, Interest and Money* (1936) Keynes himself showed how full employment could be reached, but he made no attempt to work out the long-term conditions which must be satisfied before an economy can continue to produce at that level. Harrod's and Domar's analyses of this problem show that long-term full employment requires that two fundamental conditions be satisfied.

First, the economy must invest full employment saving every year. If saving is s_f per cent of the full employment national income, and investment falls short of this, then as Keynes showed, effective demand is bound to be insufficient for full employment.

Second, for continuous full employment, the rate of growth of output must equal the growth of the physical labour force, plus the rate of increase in labour productivity. If there are *n* per cent more workers every year, and each produces *a* per cent more output, then continuous full employment requires that production grow $(n + a)$ per cent a year. There will be no need to make use of *n* per cent more workers if output grows less than this, so all the extra workers who wish to join the labour force will not find employment.

Harrod and Domar both discovered a truism which allows formulae for *g*, the rate of growth, to be derived from these fundamental conditions. *g* can be defined as $\delta Y/Y$, where δY is 'increase in output' and *Y* the level of output. $\delta Y/Y$ is identically equal to $\delta K/Y$ divided by $\delta K/\delta Y$, where $\delta K/Y$ is 'increase in capital/output', that is, 'investment/output', while $\delta K/\delta Y$ is 'increase in capital/increase in output' or the *marginal* capital-output ratio. There is therefore the truism that:

$$g \equiv \text{Investment}/\text{output}(I/Y) \div \text{the capital} - \text{output ratio}(\mathrm{C}).$$

This can be combined with two basic full employment conditions. The result is presented first in the manner suggested by Harrod (whose model was published seven years prior to Domar's).

The condition that for full employment the share of investment must equal the full employment savings ratio, s_f, means that in the above formula, it is necessary that:

$$g = s_f (\textit{which has to equal } I/Y) \textit{ divided by } C.$$

There will be one particular level of *C*, the marginal capital–output ratio, which profit maximizing entrepreneurs consider ideal, for which Harrod used the symbol, C_r, and when this is substituted for *C* in the above expression, one

Harrod–Domar Growth Model, Table 1 A table to illustrate growth at the warranted rate $s_f = 12$ and $C_r = 4$

Year	Capital stock	National income	Desired capital	Investment
	$K = K_{-1} + I_{-1}$	Y	$C_r.Y$	$I = s.Y$
1	400.00	100.00	400.00	12.00
2	412.00	103.00	412.00	12.36
3	424.36	106.09	424.36	12.73

necessary condition for continuous equilibrium growth at full employment is arrived at:

$$g = s_f/C_r$$

A second condition which needs to be satisfied if there is to be continuous full employment is that the economy's rate of growth must equal $(n + a)$, the rate of growth of the physical labour force plus labour productivity. Hence, if there is to be continuous full employment growth, it is necessary that:

$$g = s_f/C_r = n + a$$

So growth has to equal both s_f/C_r and $(n + a)$. Harrod called the first of these the 'warranted' rate of growth for which he used the symbol g_w and the second the 'natural' rate for which he wrote g_n. An economy will only be able to achieve continuous full employment if its rate of growth is equal to both g_w and g_n. Since in Harrod's account, s_f and C_r which determine the 'warranted' rate, and $(n + a)$ which determines the natural rate, are exogenously given and independent, g_w and g_n will only be equal by chance. It follows that actual economies will find it virtually impossible to achieve continuous full employment, a Keynesian result which follows naturally from Harrod's Keynesian assumptions.

In the version Domar published in 1946 and 1947 which he sent to the printers before he was aware of Harrod's 1939 article, 'the rate of growth required for a full employment equilibrium' (Harrod's g_n) is described as r, the economy's long-term saving ratio (s_f) is α, and the annual output produced by a unit of capital in the long term (l/C_r) is σ. Domar's equivalent to Harrod's condition for long term full employment equilibrium that g_n must equal $s_f/C_r$is (Harrod 1959) the identical proposition that r must equal $\alpha\sigma$. Harrod's symbols are more often used than Domar's because g, s, and C are more readily thought of as the growth rate, the savings ratio and the capital-output ratio than, r, α and l/σ.

Harrod and Domar were both then unaware of the work of Fel'dman, who had produced a growth model quite similar to theirs in the Soviet Union in 1928. Domar published an account of Fel'dman's model, 'A Soviet Model of Growth', in his theirs in the Soviet Union in 1928. Domar published an account of Fel'dman's model, 'A Soviet Model of Growth', in his *Essays in the Theory of Economic Growth* (1957), a collection of papers in which his own model of growth and its implications for public policy are fully developed.

The consequences of the all but inevitable failure to achieve Harrod's and Domar's conditions provide illuminating insights into the long term development of real economies which often fail to achieve full employment over considerable periods. Harrod's first condition is that g, the economy's actual rate of growth must equal the 'warranted' rate, s_f/C_r. The meaning of this condition is that equilibrium growth entails that full employment saving be continuously invested, as in Table 1, where a full employment savings ratio (s_f) of 12 per cent, and a required capital-output ratio (C_r) of 4 are assumed, so that the warranted rate is exactly 3 per cent. The real national income is 100 in the first year, and the initial capital stock is exactly the one required, namely four times this or 400.

Investment which is always 12 per cent of the national income is added to the capital stock of the previous year, and the national income (which grows at exactly the warranted rate of 3 per cent) is always exactly one-quarter the capital stock, so the 'desired capital stock' (which is C_r times the national income) is always in line with the actual stock. This means that if the economy grows at precisely the 'warranted' rate (3 per cent), entrepreneurs will be satisfied that they have undertaken the commercially correct rate of investment.

Harrod–Domar Growth Model, Table 2 Growth where the actual rate (g) is 1 per cent less than the warranted rate (gw)

Year	Capital stock	National income	Desired capital	Investment
	$K = K_{-1} + I_{-1}$	Y	$C_r Y$	$I = s.Y$
1	400.00	100.00	400.00	12.00
2	412.00	102.00	408.00	12.24
3	424.24	104.04	416.16	12.48
4	436.72	106.12	424.48	12.73

In 1939 Harrod defined the 'warranted' rate of growth as 'that rate of growth which, if it occurs, will leave all parties satisfied that they have produced neither more nor less than the right amount', which is precisely the situation in the table where the actual capital stock always equals the desired stock.

Table 2 illustrates what goes wrong when g, the actual rate of growth is less than g_w. It is assumed that g is only 2 per cent, while with s_f 12 per cent and C_f 4 as before, g_w is still 3 per cent.

Here, where the rate of growth is slightly less than the warranted rate the capital stock actually increases *faster* than the one entrepreneurs consider ideal. This margin of excess capital grows continuously, year after year, so the time is bound to come where entrepreneurs will respond by cutting investment. According to Harrod (1952) the rate at which firms invest to expand will be determined as follows:

> Let the representative entrepreneur on each occasion of giving an order repeat the amount contained in his order for the last equivalent period, adding thereto an order for an amount by which he judges his existing stock to be deficient, if he judges it to be deficient, or subtracting therefrom the amount by which he judges his stock to be redundant, if he does so judge it (p. 284).

In the conditions set out in Table 2 where g_w exceeds g, part of the capital stock of the representative entrepreneur gradually becomes redundant, so investment and therefore effective demand and growth will begin to fall. Thus Harrod arrived at the extremely uncomfortable conclusion that if actual growth is less than the 'warranted' rate, it will come to fall still further below this. It can be shown similarly that if g exceeds g_w for any reason, the economy will become increasingly short of capital with the result that g will rise further and further above g_w.

There are propositions in microeconomic theory which claim to demonstrate that if there is a surplus of any particular commodity, then the rate at which it is supplied will fall off with the result that market forces respond in the direction required to remove the surplus. The economy is therefore expected to respond to a shortage or surplus of an individual commodity in the manner required to remove it; but according to Harrod's instability theorem, at the macroeconomic level, any chance deviation of actual growth below the warranted rate will lead to excess capacity, and as this grows, investment and hence effective demand will be curtailed, which will lead to the creation of still more excess capacity. The response of the macro-economy to excess capital will therefore be the opposite of that required to remove the excess, with the result that economies are inherently unstable at the macro level.

Domar arrived at a similar result by directly contrasting the rate of growth of effective demand to the growth of productive capacity. In his formulation (but using Harrod's symbols) the growth in demand equals the increase in investment (δI) times the multiplier ($1/S$) while the growth of productive capacity equals total investment (I) divided by the long term capital-output ratio (C_r), with the result that where the growth of demand equals the growth of capacity:

$$\delta I/I = s/C_r,$$

A slight upward deviation of investment from this critical rate of growth (which corresponds to Harrod's 'warranted' rate) will raise $\delta I/I$ (which equals the growth of demand) relative to s/C_r, the growth of capacity, and this can be expected to lead to further increases in investment. Thus as in Harrod's argument, any chance deviation in *the rate of growth of investment* from the critical s/C_r

growth rate of productive capacity can be expected to lead to further deviations in the same direction.

The difficulties capitalist economies must overcome to achieve continuous expansion at full employment are still greater because in order to grow all the time at the 'warranted' rate and so escape the instability inherent in any departure of g from s_f/C_r, the 'warranted' rate itself must equal the natural rate, but there is no reason why s_f/C_r should equal $(n + a)$.

Suppose the conditions assumed in the above tables ($s_f = 12$ per cent and $C_r = 4$ so that $g_w = 3$ per cent) but that the labour force grows at only 0.5 per cent and productivity at 1.5 per cent so that g_n is just 2 per cent. Then the economy's full employment output can grow no more than 2 per cent a year, so it will be possible for the economy to achieve the 3 per cent growth rate required to prevent the emergence of continual excess capacity for a few years at most. Its actual long term growth rate is likely to approximate to the 2 per cent 'natural' rate with the result that g, the actual rate will fall short of g_w most of the time. Then years with excess capacity leading to economic depression will predominate over periods of expansion. The continual tendency towards depression will reduce average actual saving (s) below full employment saving (s_f). Then via unemployment and underproduction, the economy's actual long term savings ratio will come into line with the lower investment ratio (C_r times g_n) which physical conditions actually allow the economy to sustain.

Conversely, where g_n exceeds g_w, market forces will all the time attempt to push actual growth above the 'warranted' rate, with the result that conditions where capital is scarce and saving inadequate will predominant. In the first instance this will lead to excess demand for capital and therefore to a predominance of inflation over deflation which is what Harrod emphasized in 1948: 'we may have plenty of booms and a frequent tendency to approach full employment, the high employment will be of an inflationary and therefore unhealthy character' (p. 88). However, if investment of less than $C_r(n+a)$ causes the rate of growth of productive capacity to fall short of $(n + a)$, then there will be insufficient growth of the real capital stock to provide enough physical capital equipment to raise employment at the rate at which the physical labour force is growing (n), with the result that the economy will suffer from growing *structural* unemployment.

Harrod's theory therefore predicts that incompatibilities between long term saving and investment opportunity are all but certain to cause prolonged unemployment (which will be structural where g_n exceeds g_w and demand deficient where g_w exceeds g_n) with persistent inflation in addition wherever long term saving is inadequate for the natural rate of growth. This raises fundamental problems for public policy, and Harrod argued in 1939 that 'the difficulties may be too great to be dealt with by a mere anti-cycle policy'. He suggested that where an economy suffers from a long term tendency to over saving with the result that the 'warranted' rate exceeds the 'natural' rate, then a generous attitude to public investment is appropriate so that more will be undertaken than commercial and social considerations call for. Conversely governments should seek to generate more long term saving and to curtail long range and social investment where the 'natural' rate exceeds the 'warranted' rate.

By the later 1950s the United States and several West European economies were achieving full employment and negligible inflation which led a number of distinguished economists to develop models of economic growth which were less prone to predict secular unemployment or inflation. Robert Solow (1956) and Trevor Swan (1956) produced neoclassical growth models where market forces adjust the equilibrium capital–output ratio (C_r) so that this automatically equates g_w to g_n (which is achieved when $C_r(n+a)/s_f)$. Nicholas Kaldor (1955–6 and 1957) evolved a Keynesian model of growth and income distribution where shifts between wages and profits will adjust the savings ratio until this becomes the one required $(C_r(n+a))$ to equate g_w and g_n. A few years earlier, Alexander (1950) had questioned the inevitability of Harrod's knifeedge which sent an economy soaring upwards or downwards wherever g diverged from g_w.

The unemployment and stagflation of the 1970s and the 1980s has surprisingly failed to restore some of the former prestige of the Harrod–Domar model. In the 20th century in the leading Western economies there have been prolonged periods when more saving would have been beneficial, and others with every appearance of inadequate effective demand. The Harrod–Domar growth model is one of the few which actually predicts this, so it still deserves serious attention.

See Also

- ► Aggregate Demand and Supply Analysis
- ► Natural and Warranted Rates of Growth

Bibliography

Alexander, S.S. 1950. Mr Harrod's dynamic model. *Economic Journal* 60: 724–739.

Domar, E. 1946. Capital expansion, rate of growth, and employment. *Econometrica* 14: 137–147.

Domar, E. 1947. Expansion and employment. *American Economic Review* 37: 34–55.

Domar, E. 1957. *Essays in the theory of economic growth*. New York: Oxford University Press.

Harrod, R.F. 1939. An essay in dynamic theory. *Economic Journal* 49: 14–33.

Harrod, R.F. 1948. *Towards a dynamic economics*. London: Macmillan.

Harrod, R.F. 1952. Supplement on dynamic theory. In *Economic essays*, ed. R.F. Harrod. London: Macmillan.

Harrod, R.F. 1959. Domar and dynamic economics. *Economic Journal* 69: 451–464.

Kaldor, N. 1955–6. Alternative theories of distribution. *Review of Economic Studies* 23(2): 83–100.

Kaldor, N. 1957. A model of economic growth. *Economic Journal* 67: 591–624.

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

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

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

Harsanyi, John C. (1920–2000)

Roger B. Myerson

Abstract

John Harsanyi worked to extend the general theoretical framework of economic analysis. He established the modern basis for utilitarian ethics. He developed a general bargaining solution to that included the Nash bargaining solution and the Shapley value as special cases. He became a leading advocate of non-cooperative game theory as the general framework for analysis of social interactions among rational individuals. He developed the tracing procedure to select among multiple equilibria of games. He showed how to interpret mixed-strategy equilibria in game theory. His general model of Bayesian games with incomplete information became a cornerstone of information economics.

Keywords

Bargaining; Bayesian decision theory; Bayesian game; Cardinal utility; Cooperative game theory; Game theory; Harsanyi, J.; Incomplete information in game theory; Information economics; Interpersonal utility comparisons; Morgenstern, O.; Nash bargaining solution; Nash equilibrium; Neumann–Morgenstern utility axioms; Noncooperative game theory; Private information; Rationality; Rule utilitarianism; Schelling, T.; Shapley value; von Neumann, J.; Welfare economics

JEL Classifications

B31; C70; D60

John Harsanyi extended the theoretical framework of economic analysis with major contributions to game theory and welfare economics. His general approach to social theory was based on a fundamental assumption that people are rational decision-makers who share a basic understanding of the things that they value in the world. His personal experiences made him profoundly sceptical of theories that try to justify social systems from other assumptions, without respecting the values, the rationality, and the intelligence of all individuals in society. He understood that social institutions and policies should be evaluated by carefully analysing their impact on individuals' welfare. From his training in philosophy he appreciated the basic importance of general unified frameworks in social theory. He recognized the foundations of such a framework in Bayesian decision theory, with its compelling axiomatic characterizations. So he devoted his career to the development of a general framework for economic analysis based on these principles. His best-known contribution is the general model of Bayesian games with incomplete information, which became a cornerstone of information economics.

Harsanyi grew up in Budapest, Hungary, where his distinction as a student was marked in 1937 by his winning the first prize in Hungary's national mathematics competition. But at the university he chose to study pharmacy so that he could share his father's business, as other options were then clouded by the threat of war. He was forced into hiding by Nazi racial policies during the last months of the German occupation. After the war, he studied philosophy and earned a doctorate from the University of Budapest in 1947, but his intellectual independence led to political difficulties with the Communist regime, which forced him out of the university. In 1950 he fled Hungary and found refuge in Australia.

He began studying economics at Sydney University, earning an MA in 1953. He then held a lectureship at the University of Queensland, where he began to read game theory. In 1956, when he already had published articles on welfare economics, he enrolled as a student at Stanford University, earning his second doctorate in 1959. He then held faculty positions at the Australian National University, Wayne State University, and, from 1965, in the School of Business Administration at the University of California, Berkeley.

In his early contributions to welfare economics, Harsanyi established the modern basis for utilitarian ethics. Von Neumann and Morgenstern (1947) had shown axiomatically that a rational individual should choose among risky alternatives by maximizing the expected value of a cardinal utility function, but some economists doubted whether this cardinal utility, defined for individual risk analysis, had any relevance for social welfare analysis. Harsanyi (1953) argued that, in ethical decision-making, to avoid any dependence on our particular roles in society, we must imagine ourselves in an initial position before social roles have been assigned, when we could only anticipate getting the role of someone drawn at random from the whole population. Thus, ethical decision-making involves an essential element of risk, and we naturally get a social welfare function equal to the average utility of all members of society.

This average requires interpersonally comparable utility scales, assessed by sympathetically comparing the prospect of being in one person's position or another's. Harsanyi argued, as his 'similarity postulate', that such comparable utilities for all individuals can be generated by a common utility function, based on shared human values, once the factors that cause apparent differences among individuals' tastes are included as parameters of the function. Harsanyi (1955) showed that, even without this similarity postulate, the Neumann–Morgenstern utility axioms (applied to individual and social decision-making) and the Pareto welfare axiom (that social preferences should be consistent with any unanimity of individual preferences) together imply that social utility can be defined only as some linear function of individual utility values.

In his later work on welfare economics, Harsanyi (1977a, Chap. 4; 1977b, c), argued that ethical analysis should be used to evaluate general social rules or institutions rather than specific acts. That is, we may consider ethical rules that prescribe people's behaviour in a wide range of situations, recognizing that behaviour in other situations could be determined by self-interest according to some Nash equilibrium. Then, as rule utilitarians, we should advocate rules that yield the highest average of expected utilities for all individuals.

Harsanyi began working on cooperative game theory in the mid-1950s, when many different cooperative solution theories were being studied. But most of these theories could yield multiple solutions or no solutions for a game, or could not even be defined without some special structures like transferable utility. Harsanyi's view of the field was clarified by his insistence that a good solution concept should yield one well-defined solution to any game. At that time, there were only two cooperative solution concepts that yielded unique solutions to broad classes of games: the Shapley (1953) value for games with transferable utility, and the Nash (1950) bargaining solution for two-person games without transferable utility. Harsanyi (1956) showed that the Nash bargaining solution could be derived from an earlier theory of Zeuthen (1930). Then Harsanyi (1963) developed a general bargaining solution that included the Nash bargaining solution and the Shapley value as special cases.

In the mid-1960s, Harsanyi shifted from cooperative to non-cooperative game theory. The basic definition of non-cooperative equilibrium had been introduced by Nash (1951). But there was little further development of non-cooperative theory until Schelling (1960) analysed bargaining processes as games with multiple equilibria, where any cultural or environmental factor that focuses the players' attention on one equilibrium can become a self-fulfilling prophecy. Harsanyi (1961) argued that the distribution of power that is measured by a cooperative solution could be the focal factor that selects among the many non-cooperative equilibria of a bargaining game. But then Harsanyi began to recognize the force of Nash's early arguments for the greater generality of the non-cooperative approach, which is based on a precise specification of each player's individual decision problem, which is lacking in cooperative models. Thus Harsanyi became a leading advocate of non-cooperative game theory.

Harsanyi understood that the non-cooperative approach could not become a standard methodology for applied economic analysis without some refinements of Nash's equilibrium concept, because it can yield very large sets of equilibria for many games. So he began a search for theoretical criteria to select among multiple equilibria, which culminated in his book with Selten (1988). Their selection theory is based on Harsanyi's (1975) tracing procedure, which can select a unique equilibrium from a given initial hypothesis about the players' strategic behaviour. For each number t between 0 and 1, we define a t-auxiliary game that differs from the original game in that each player thinks that the other players have probability $1 - t$ of behaving according to the initial hypothesis; otherwise, with probability t, they choose their strategies rationally. The tracing procedure finds a continuous path of equilibria for these auxiliary games, starting from the trivial 0-auxiliary game and ending at a unique equilibrium of the original game when $t = 1$.

Harsanyi's work on incomplete information in games began (1962) with the problems of extending Nash's bargaining solution to situations where players do not know each others' payoffs. In this work, he began to recognize the problems of modelling players' beliefs about each others' beliefs in a game. Harsanyi (1967–8) confronted these modelling problems at the most general and fundamental level, showing how the basic definition of normal-form games should be modified to analyse situations where individuals have different information.

The early development of game theory was based on von Neumann's (1928) argument that any dynamic game in extensive form can be represented by a conceptually simpler one-stage game in normal form. In this normal-form game, each player chooses a strategy that is a complete contingent plan of action, specifying what the player would do at each stage of the dynamic game as a function depending on any information that the player might learn during the game. In normal-form analysis, we assume that the players choose their strategies simultaneously and independently at the start of the game, before anyone gets any private information, and thereafter their behaviour in the dynamic game can be determined mechanically by their strategies. Thus, questions about the players' private information are suppressed in normal-form analysis.

Harsanyi (1967–8) showed how to correct this deficiency by developing a more general game model that allows players to have different initial information, without losing the analytical simplicity of the normal form. Each player's private information at the start of the game is represented by a random variable that is called the player's *type*. Harsanyi defined a *Bayesian game* to be a mathematical model that specifies (a) the set of players, (b) the set of feasible actions for each player, (c) the set of possible types for each player,

(d) each player's expected payoff for every possible combination of all players' actions and types, and (e), for each possible type of each player, a probability distribution over the other players' possible types, which describes what each type of each player would believe about the others' types.

The beliefs in a Bayesian game are said to be *consistent* if the players' type- contingent beliefs can all be derived by Bayes's rule from some common prior distribution over types. Although not analytically essential, this assumption of consistent beliefs has been regularly used in applied economic analysis, because it allows that differences in players' beliefs may be explained by different previous experiences.

To represent dynamic extensive-form games by games in Bayesian form, each player's action in a Bayesian game may be interpreted as a plan that describes what the player would do in any situation after the beginning of the game, as a function of what the player may learn during the game. A player's strategy, in von Neumann's original sense, would then be a function that specifies a feasible action for each of the player's possible types. But each player is assumed to know his type already when the game begins, and so Harsanyi worked to avoid the fiction of strategic decision-making by players who have not yet learned their types. It would be better, he argued, to imagine that a player's different possible types correspond to different agents, one of whom will be randomly selected to be active in the game. The point is that each player's optimal decisions will maximize his conditional expected payoff given his actual type, and there is no significance to any expected value that is not conditioned on such type information.

Harsanyi emphasized that games must be analysed from the perspective of someone who only knows the information common to all players, which is summarized in the Bayesian game model. Game-theoretic analysis requires us to deny ourselves any knowledge of any player's actual type, so that we can appreciate the uncertainty of the other players who do not know it. The actual type of each player, being private information, must be treated as an unknown quantity or random variable in our analysis. So an equilibrium of a Bayesian game specifies a feasible action for every possible type of every player, such that the specified action for each type of each player maximizes his conditional expected payoff, given his type, given his beliefs about the others' types, and given the type-contingent actions of the other players according to this equilibrium.

Applications of Bayesian games developed quickly. Harsanyi and Selten (1972) defined a generalization of Nash's bargaining solution for Bayesian games, where players have incomplete information about each other. By embedding normal-form games in the larger space of Bayesian games, Harsanyi (1973) showed how to interpret mixed-strategy equilibria in non-cooperative game theory. Such equilibria had seemed to imply paradoxically that rational players should base their decisions on randomizing devices like roulette wheels, but this apparent paradox was a consequence of the normal-form assumption that players choose strategies before they get any private information. By letting each player have some minor private information that changes payoffs only slightly, Harsanyi could transform any mixed- strategy equilibrium into a Bayesian equilibrium where each type chooses an optimal action without randomization.

Harsanyi's Bayesian games have become the standard economic model for analysing transactions among individuals who have different information. Before 1967, the lack of a general framework for informational problems had inhibited economic inquiry about markets where people do not share the same information. The unity and scope of modern information economics were found in Harsanyi's framework.

See Also

- ▶ Expected Utility Hypothesis
- ▶ Interpersonal Utility Comparisons (New Developments)
- ▶ Nash Program

- ▶ Savage's Subjective Expected Utility Model
- ▶ Shapley Value
- ▶ Utilitarianism and Economic Theory
- ▶ von Neumann, John (1903–1957)

Selected Works

1953. Cardinal utility in welfare economics and the theory of risk-taking. *Journal of Political Economy* 61: 434–435.

1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy* 63: 309–321.

1956. Approaches to the bargaining problem before and after the theory of games. *Econometrica* 24: 144–157.

1961. On the rationality postulates underlying the theory of cooperative games. *Journal of Conflict Resolution* 5: 179–196.

1962. Bargaining in ignorance of the opponent's utility function. *Journal of Conflict Resolution* 6: 29–38.

1963. A simplified bargaining model for the n-person cooperative game. *International Economic Review* 4: 194–220.

1967–8. Games with incomplete information played by Bayesian players. *Management Science* 14: 159–182, 320–334, 486–502.

1972. (With R. Selten.) A generalized Nash solution for two-person bargaining games with incomplete information. *Management Science* 18(5): 80–106.

1973. Games with randomly disturbed payoffs, a new rationale for mixed strategy equilibrium points. *International Journal of Game Theory* 2: 1–23.

1975. The tracing procedure, a Bayesian approach to defining a solution for n-person noncooperative games. *International Journal of Game Theory* 4: 61–94.

1977a. *Rational behavior and bargaining equilibrium in games and social situations*. New York: Cambridge University Press.

1977b. Rule utilitarianism and decision theory. *Erkenntnis* 11: 25–53.

1977c. Morality and the theory of rational behavior. *Social Research* 44: 623–56.

1988. (With R. Selten.) *A general theory of equilibrium selection in games*. Cambridge, MA: MIT Press.

Bibliography

Harsanyi, J.C. 2001. Memorial issue. *Games and Economic Behavior* 36(1): 1–56.

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

Nash, J. 1951. Non-cooperativegames. *Annals of Mathematics* 54: 286–295.

Schelling, T. 1960. *The strategy of conflict*. Cambridge, MA: Harvard University Press.

Shapley, L. 1953. A value for n-person games. In *Contributions to the theory of games*, ed. H. Kuhn and A. Tucker, vol. 2. Princeton: Princeton University Press.

von Neumann, J. 1928. Zur Theories der Gesellschaftsspiele. *Mathematische Annalen* 100: 295–320.

von Neumann, J., and O. Morgenstern. 1947. *Theory of games and economic behavior*. 2nd ed. Princeton: Princeton University Press.

Weymark, J. 1995. John Harsanyi's contributions to social choice and welfare economics. *Social Choice and Welfare* 12: 313–318.

Zeuthen, F. 1930. *Problems of monopoly and economic warfare*. London: Routledge.

Hart, Albert Gailord (1909–1997)

Peter Earl

Keywords
Decision theory; Expectations; Hart, A. G.; Uncertainty

JEL Classifications
B31

Born in Oak Park, Illinois, Hart received his BA from Harvard in 1930 and his Ph.D. from the University of Chicago in 1936. Most of his career – from 1946 until his retirement in 1979 – was spent as Professor of Economics at Columbia University. Much of his noteworthy work concerned the implications of uncertainty for policymakers, but he should also be remembered as having worked with Kaldor and Tinbergen (1964) to produce an ingenious proposal for a commodity reserve currency: this would serve to improve international liquidity simultaneously with providing a means of protecting incomes of primary producers against shrinkage in times of depression.

Hart's work on uncertainty included a monograph (1940), one notable feature of which was an attempt to analyse how decision makers can judge their success or failure, and thence reformulate their expectations, in the light of partial knowledge of performance distributions. From 1936 onwards, he emphasized the rationality, in situations of uncertainty, of choosing flexible production technologies which, though they might not be perfectly adapted to any specific output rate, would not be disastrously expensive to run over a range of outputs. This idea, which was also promoted by his Chicago contemporary Stigler (1939), led Hart to be critical of much writing on decision theory. He felt it misleading to theorize as if firms assign probabilities to rival hypothetical outputs, aggregate these weighted values and then build their plans around the weighted average of probable output rates (1942). Hart was also irritated by Keynes's tendency to speak of expectations in terms of certainty equivalents, and he warned that, 'generally speaking, the business policy appropriate to a complex of uncertain anticipations is different in kind from that appropriate for any set of certain expectations' (1947, p. 422).

Hart carried this theme into work critical of deterministic macroeconomic model-building and fiscal policy formulation (1945), and into a distinctive approach to monetary theory (1948, especially part II). In the latter, he introduced the 'margin of safety' motive for holding liquid assets, arguing that the structure of economic affairs is such that risks are usually linked: a single disappointment is prone to cause many other things to go wrong in consequence. Hart's concern with surprise, flexibility, and structural linkages in many ways foreshadows themes that emerged in the 1980s in the business policy literature on scenario planning and strategic choices. However, he is not usually credited as the pioneer of this kind of thinking: having been largely ignored by mainstream writers, his ideas were sufficiently poorly known to end up being reinvented.

Selected Works

1940. *Anticipations, uncertainty and dynamic planning*. Chicago: University of Chicago Press.

1942. Risk, uncertainty and the unprofitability of compounding probabilities. In *Studies in mathematical economics and econometrics*, ed. O. Lange, F. McIntyre, and T.O. Yntema. Chicago: University of Chicago Press.

1945. 'Model-building' and fiscal policy. *American Economic Review* 35: 531–558.

1947. Keynes's analysis of expectations and uncertainty. In *The new economics: Keynes's influence on theory and public policy*, ed. S.E. Harris. New York: Knopf.

1948. *Money, debt and economic activity.* New York: Prentice-Hall.

1964. (With N. Kaldor and J. Tinbergen.) The case for an international commodity reserve currency. Paper submitted to UNCTAD, Geneva, March-June 1964. In *Essays on economic policy*, vol. 2, ed. N. Kaldor. London: Duckworth.

Bibliography

Stigler, G.J. 1939. Production and distribution in the short run. *Journal of Political Economy* 47: 305–327.

Hawkins–Simon Conditions

Hukukane Nikaido

Abstract

In a Leontief system of interindustrial input–output relationships consisting of n sectors of industry, each of which produces a single good, without joint products, under constant returns to scale, and using *n* goods as input in fixed proportions, the balance of demand for and supply of goods is represented by a system of linear equations.

JEL Classifications

L1

In a Leontief system of interindustrial input–output relationships consisting of *n* sectors of industry, each of which produces a single good, without joint products, under constant returns to scale, and using *n* goods as input in fixed proportions, the balance of demand for and supply of goods is represented by a system of linear equations

$$x_i = \sum_{j=1}^{n} a_{ij} x_j + c_i, \quad (i = 1,\ 2,\ \ldots,\ n),$$

where a_{ij} are non-negative input coefficients of the *j*th sector, x_j is the level of output of the *j*th sector and c_i is the level of final demand for the *ith* good ($i, j = 1, \ldots, n$).

With the input coefficient matrix *A* having a_{ij} in the *i*th row and the *j*th column, the output vector *x* having x_j in the *j*th component, and the final demand vector *c* having c_i in the *i*th component the system is represented in matrix form by the equation

$$x = Ax + c.$$

The system is productive enough to give positive net output over input, if x_j non-negative units of output of the *j*th sector ($j = 1, \ldots, n$) are produced to meet a bill of positive final demand *ci* ($i = 1, \ldots, n$).

The *productivity* of the system, which is equivalent to the condition that the *n*- dimensional square matrix $I - A$, where *I* is the identity matrix, have an inverse matrix $(I - A)^{-1}$ having all the elements non-negative, hinges on and is completely determined by the magnitudes of the input coefficients. A necessary and sufficient condition for such productivity, stated in terms of inequalities constraining the magnitudes of the input coefficients and referred to as the Hawkins-Simon conditions, after the names of its discoverers (Hawkins and Simon 1949), is that all the principal minor determinants of the matrix *I-A* be positive. This is equivalent to the seemingly weaker conditions that the *n* principal minor determinants located in the ascending order on the upper left corner of the matrix *I-A* be positive

$$\Delta_k = \begin{vmatrix} 1 - a_{11} & -a_{12} & \ldots & -a_{1k} \\ -a_{21} & 1 - a_{22} & \ldots & -a_{2k} \\ \ldots & \ldots & \ldots & \ldots \\ -a_{k1} & -a_{k2} & \cdots & 1 - a_{kk} \end{vmatrix} > 0, (k = 1, \ldots,\ n).$$

As a mathematical result the equivalence of the Hawkins–Simon conditions to productivity is very easy to prove, as can readily be shown by transforming the equation $(I - A)\ x = c$ through Gaussian elimination to a triangular form

$$b_{11}x_1 + b_{12}x_2 + \ldots + b_{1n}x_n = d_1 b_{22}x_2 + \ldots$$
$$+b_{2n}x_n = d_2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$b_{nn}x_n = d_n,$$

where

$$b_{ij} \leqq 0 (i < j), d_i \geqq 0 (i = 1, \ldots, n)$$

and

$$\triangle_k = b_{11}b_{22}\ldots b_{kk} (k = 1, \ldots, n)$$

Since the Hawkins-Simon conditions ensure the productivity of the system, they are a primary prerequisite for the Leontief system, and enlarged systems involving it as a built-in subsystem, to be well-behaved. They also make the Leontief system dynamically well-behaved. In the multiplier process over discrete time,

$$x_i(t+1) = \sum_{j=1}^{n} a_{ij} x_j(t) + c_i, \quad (i = 1,\ \ldots,\ n)$$

the solution converges to the equilibrium output levels supplying net output equal to the final demand $c_i (i\ =\ 1,\ \ldots,\ n)$, if and only if the Hawkins–Simon conditions are satisfied. This stability is equivalent to the convergence of the matrix geometric progression

$$I + A + A^2 + \ . \ . \ . \ + A^t + \dots$$

to the inverse matrix $(I - A)^{-1}$. In the multiplier process over continuous time,

$$\mathrm{d}x_i/\mathrm{d}t = \alpha_i\left(\sum_{j=1}^{n} a_{ij}x_j + c_i - x_i\right), \quad (i = 1, \ \dots, \ n)$$

the Hawkins–Simon conditions are necessary and sufficient as well for the convergence of the solution to the same equilibrium output levels, which is equivalent to the condition that the real parts of all the eigenvalues of the matrix *A-I* be negative.

See Also

- ▶ Linear Models
- ▶ Perron–Frobenius Theorem

Bibliography

Hawkins, D., and H.A. Simon. 1949. Note: Some conditions of macroeconomic stability. *Econometrica* 17: 245–248.

Hawley, Frederick Barnard (1843–1929)

Mauro Boianovsky

Abstract

Frederick Barnard Hawley (1843–1929) advanced the ‘risk theory of profit’: profit is the reward entrepreneurs get to relieve the other productive factors from risk in competitive conditions. The normal rate of profit is determined by the expectation of profit that just covers the marginal entrepreneur’s subjective valuation of risk. The current rate of profit will converge to its normal value because of the operation of the ‘readjustment period’, when income contraction brings about a fall in aggregate supply larger than the reduction in aggregate demand. This is explained by Hawley’s concept of the consumption function.

Keywords

Consumers’ expenditure; Entrepreneurship; Hawley, F.; Keynes, J.M.; Knight, F.; Paradox of thrift; Patinkin, D.; Readjustment period; Risk aversion; Risk theory of profit; Uncertainty; Underemployment equilibria

JEL Classifications

B31

Frederick Barnard Hawley was born on 5 February 1843 in Albany (New York State), and died on 31 May 1929 in New York City. After spending his freshman year at Harvard University, he went to Williams College (Massachusetts) in 1861, where he graduated three years later. Returning to Albany, he took up the study of law, but gave it up after a year to go into the family’s lumber business. In 1876 he became a cotton broker and merchant in New York City, a position he held until his retirement in 1926.

A couple of years after his move to New York Hawley published his first articles, advancing an approach to aggregate economic fluctuations based on his new conception of the saving–investment process. Those articles were expanded in 1882 into his book *Capital and Population*. In the 1890s and early 1900s Hawley published several articles in the *Quarterly Journal of Economics*, where he put forward his ‘risk theory of profit’. Hawley’s contributions to economics are contained in his 1907 book *Enterprise and the Productive Process*, in which he put together and elaborated ideas he had developed since the late 1870s. Feeling that his proposed new framework, based on the key role of the entrepreneur, had not been widely discussed, Hawley wrote 20 years later an article in the *American Economic Review* summing up his theoretical system. He was a member of the American Economic Association from 1888 to his death, and served as its treasurer (1892–95) and vice-president (1909).

Hawley was one of the main protagonists in the long and intense controversy about the theory of profit that took place in American economics from the end of the 19th century to the beginning of the 20th century and culminated with the publication of Frank Knight's classic 1921 volume. Hawley enunciated the fundamental principle that there would be no profits in a competitive market in which the course of future events was entirely foreseen, since all factor services would be paid at rates fixed in advance, with changes in their productivity during the period of contract taken into consideration. Prices and costs would converge; there would be no 'residue'. In actual economies, subject to future unforeseen influences, the function of the entrepreneur is to relieve others of risk. The entrepreneur bargains with the workers, capitalists and landlords for the use of their services, paying them not with any share of the product itself but with stipulated amounts of purchasing power. The actual product is owned by the entrepreneur, who must assume the responsibility of the enterprise and convert the output into purchasing power at the market price (cf. J. M. Keynes's 1933 similar distinction between a cooperative economy and an entrepreneur economy). According to Hawley, the entrepreneur is the dominant active element in the productive process, combining the three subsidiary passive productive factors. Since the incomes of individuals are necessarily composite, factors must be associated with functions, not with individuals. The entrepreneur's profit is a residual, non-contractual income whose amount is determined only after the output is sold. Hawley assumes that, in order to be relieved of a risk, agents are willing to pay more than the risk, calculated according to the laws of probability, is worth, since they 'prefer a certainty to an uncertainty'. Entrepreneurs perform a service worth more to its recipients than the price they have to pay, and yet worth less to themselves than they get for it. Hence, the assumption of risk by the entrepreneur creates value by rendering a service in transferring risks from those to whom their subjective value is great to those to whom their subjective value is less, a mutually advantageous exchange of 'certain goods' for 'uncertain goods'. Profit is the reward entrepreneurs get for performing that service in competitive conditions, and, by that, a component of the prices of commodities in general.

Entrepreneurs are deemed less risk averse than other economic agents, except for 'gamblers' and 'speculators', who are not risk averse but do not take part in the productive process. The entrepreneurs' subjective value of risk – the 'irksomeness of being exposed to risk' – is higher than its actuarial value, which means that industrial risks will not be assumed without the expectation of compensation in excess of their actuarial value. Since, under Hawley's assumption, entrepreneurs are on average and in the long run correct in their estimates, they pay productive factors less than the product will probably sell for, and absorb as profits a considerable portion of the annual flow of purchasing power. The 'normal rate of profit' is defined as the expectation of profit that just covers the marginal entrepreneur's subjective valuation of risks.

Hawley's theory of profit was taken up by Knight, who, however, criticized Hawley for ignoring the distinction between (known) risk and (unknown) uncertainty and overlooking the fact that the former is insurable. Although it is true that Hawley used the words 'risk' and 'uncertainty' interchangeably, it should be noted that he did pay careful attention to the implications of insurance for his argument. In the first place, the act of insurance does not imply that the risk or its reward are extinguished, but only that the entrepreneur transfers to the insurer a corresponding part of its expected profits. Moreover, the risks of ownership – the most substantial part of risk – cannot be shifted by insurance, but only by a sale. Entrepreneurs can, to some extent, protect themselves from risks arising from price fluctuations by entering into hedging operations with speculators, as Hawley was aware from his experience as a cotton merchant. Although forward markets cannot completely eliminate the risks influencing selling prices, to the extent in which entrepreneurs hedge themselves they will be forced by competition to forgo their reward for risk bearing and lower their prices accordingly.

Whereas Hawley's theory of profit attracted some attention at the time, his contributions to macroeconomics went largely unnoticed by the profession, probably because their import became clear only after the Keynesian revolution. Fluctuations in aggregate economic activity are explained, according to Hawley, by changes in the saving–investment relation throughout the business cycle. Investment, described by the act of subjecting capital to the uncertainties inherent in actual ownership of capital goods, is naturally connected with entrepreneurs, not with capitalists or savers. The demand for new capital goods depends essentially on the entrepreneurs' profit expectations, which are subject to violent changes due to 'unforeseeable and incalculable causes' that affect the subjective valuation of risks. The treatment of savers' behaviour is based on Hawley's path-breaking concept of consumption as a function of income. He argued that expenditure changes less then income at all levels of income, since consumers keep close to the standard of living they have once adopted. More specifically, consumption plans are determined by expected income, measured by the average income of a series of years (that is, the mathematical expectation). This implies that a sudden increase of income will yield a larger percentage for saving than a gradual one of equal extent and, furthermore, that the proportion of saving out of profits is higher because it is more variable and uncertain than other sources of income.

Hawley used his new hypothesis about the consumption function to investigate the dynamics of the economy when the current rate of profit differs from its normal value. Periods of depression are characterized by a rate of profit lower than normal, associated to excess saving in the goods market. 'What can enterprisers do, by varying the character of supply, to protect themselves against this attack of the saving class upon their chances of profit?' asked Hawley (1907a, p. 224). The answer is the 'readjustment period', Hawley's main contribution to macroeconomic theory, which set him apart from the rest of the pre-Keynesian business cycle literature: a decline of output caused by excess aggregate supply will reduce supply more than demand (because of the consumption function) and bring the economy to equilibrium at less than full employment. The equilibrating effect of the contraction in aggregate income, identified by Don Patinkin (1982) as the core of the Keynesian principle of effective demand, can be found already in Hawley's writings. Some corollaries of the idea of 'readjustment period' were also pointed out by Hawley, such as the notion that an increase of the saving flow for a given investment level will bring about a contraction of income and a return of saving to its initial amount, so that in the end 'national parsimony defeats itself' – an early formulation of the 'paradox of thrift' usually associated with J. M. Keynes.

Whether he had any influence on Keynes is a moot point in the history of macroeconomics. Despite the fact that the English economist never referred to Hawley, that possibility cannot be disregarded, especially in view of the similarity of the wording of some key passages.

See Also

- ▶ Entrepreneurship
- ▶ Keynes, John Maynard (1883–1946)
- ▶ Knight, Frank Hyneman (1885–1962)

Selected Works

1879a. The ratio of capital to consumption. *National Quarterly Review* 79: 95–116.

1879b. The rationale of panics. *National Quarterly Review* 79: 277–292.

1882. *Capital and population*. New York: A. M. Kelley, 1972.

1892. The fundamental error of 'Kapital und Kapitalzins'. *Quarterly Journal of Economics* 6: 280–307.

1893. The risk theory of profit. *Quarterly Journal of Economics* 7: 459–479.

1900. Enterprise and profit. *Quarterly Journal of Economics* 15: 75–105.

1907a. *Enterprise and the productive process*. New York and London: G. P. Putnam's Sons.

1907b. Capital (Chapter 7 of *Enterprise and the productive process*). In *Business cycle theory – Selected texts 1860–1939*, vol. 7, ed. M. Boianovsky. London: Pickering and Chatto, 2005.

1927. The orientation of economics on enterprise. *American Economic Review* 17: 409–428.

Bibliography

Bigelow, K. 1932. Hawley, Frederick Barnard. In *Encyclopaedia of social sciences,* vol. 7, ed. E. Seligman. London: Macmillan.

Boianovsky, M. 1996. Anticipations of the *general theory:* The case of F. B. Hawley. *History of Political Economy* 28: 371–390.

Davis, R. 1953. Frederick B. Hawley's income theory. *Journal of Political Economy* 61: 117–126.

Dorfman, J. 1949. *The economic mind in American civilization*, vol. 3. New York: Viking.

Hopkins, W. 1933. Profit in American economic theory. *Review of Economic Studies* 1: 60–66.

Keynes, J.M. 1933. The distinction between a co-operative economy and an entrepreneur economy. In *The collected writings of John Maynard Keynes,* vol. 29, ed. D. Moggridge. London: Macmillan, 1979.

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

Knight, F. 1921. *Risk, uncertainty and profit.* New York: A. M. Kelley, 1964.

Patinkin, D. 1982. *Anticipations of the general theory? And other essays on Keynes*. Oxford: Basil Blackwell.

Williams College. 1930. *Obituary record of the society of alumni, April 3–4*. Williamstown: Williams College.

Hawtrey, Ralph George (1879–1975)

R. J. Bigg

Keywords

Cambridge School; Credit; Effective demand; Expectations; Gold standard; Hawtrey, R. G.; Investment; Keynes, J. M.; Labour supply; Marginal productivity of capital; Money demand; Multiplier analysis; Natural rate and market rate of interest; Public works; Quantity theory of money; Rate of profit; Robertson, D. H.; Robinson, J. V.; Savings; Short-hoards; Stocks; Trade cycle; Wicksell, J. G. K.

JEL Classifications

B31

Hawtrey was born in Slough, near London, and went up to Trinity College, Cambridge, from Eton in 1898. Three years later he graduated 19th Wrangler in the Mathematical Tripos. Hawtrey remained at Cambridge for a further period to read for the civil service examinations, as was quite common at that time. This latter study included some economics with lectures largely by G.P. Moriarty and J.H. Clapham. In 1903 he entered the Admiralty, but in 1904 he transferred to the Treasury, where he was to remain until retirement in 1947 (his official retirement at 65 was in 1944). Hawtrey's only academic appointments in economics were in 1928–9, when he was given special leave from the Treasury to lecture at Harvard (as a visiting professor) and after his retirement, when he was elected Price Professor of International Economics at the Royal Institute of International Affairs (1947–52). Hawtrey served as President of the Royal Economic Society between 1946 and 1948.

Hawtrey was not, therefore, directly a part of the 'Cambridge School' of economics. Marshall took no immediate part in Hawtrey's economic education which was, for the most part, acquired in the Treasury. Nonetheless he had close contacts with the Cambridge economists. Away from economics he was involved with both the Apostles and with Bloomsbury, whilst within the subject he was a visitor to Keynes's Political Economy Club at Cambridge and his major work, *Currency and Credit* (1919a) became a standard work in Cambridge in the 1920s. Furthermore, although there were differences in approach between Hawtrey and the Cambridge School in some areas, Keynes himself noted in reviewing *Currency and Credit* the similarities between Hawtrey's approach to the theory of money and that of the Cambridge School – though Keynes remarked that Hawtrey had reached his results independently (Keynes 1920).

Hawtrey was primarily a monetary economist; his major contributions related to the quantity theory and the trade cycle. He was one of the first English economists to stress the primacy of credit-money rather than metallic legal tender. Furthermore his income-based approach, like that of the Cambridge School, led to a closer integration of the theories of money and output. For Hawtrey, money income determines expenditure, expenditure determines demand and demand determines prices.

Hawtrey summarized his aims in monetary theory in the preface to *Currency and Credit*:

> Scientific treatment of the subject of currency is impossible without some form of the quantity theory . . . but the quantity theory by itself is inadequate, and it leads up to the method of treatment based on what I have called the consumers' income and the consumers' outlay – that is to say, simply the aggregates of individual incomes and individual expenditures. (1919, p. v)

Investment (the result of saving) is included in consumers' outlays, since it is spent on fixed capital. Consumers' balances are then the difference between outlays and income and thus consist only of accumulated cash balances (including money held in bank accounts). In addition there is a similar demand for money balances by traders related to their turnover. Of course individual agents may hold both consumers' and traders' balances – Hawtrey notes that the true income of traders is the profits of the business and that this is included in consumers' income.

The 'unspent margin', or total money balances, consists of the consumers' and traders' balances taken together. From this Hawtrey derives a form of the quantity theory. Hawtrey argues that traders' balances are relatively stable, and thus the operational relationships are concerned with the supply of money (in a wide sense taken to include credit) and consumers' income and outlay. It is worth noting that compared to the Cambridge income-based approach Hawtrey's places greater emphasis on the demand for nominal balances rather than real balances. It is also interesting to note that Keynes used a similar balances approach to the quantity theory in the period after 1925 leading up to the theory presented in the *Treatise on Money* (1930), where he distinguishes first between investment and cash deposits and later between income, business and savings deposits.

The demand for money is also analysed in terms of motives. Hawtrey identifies a transaction demand, a precautionary demand, and a residual demand which reflects a gradual accumulation of savings balances or what Joan Robinson has called short- hoards (Robinson 1938). Hawtrey envisages agents as saving gradually but investing only larger sums periodically. In the meantime these short-hoards act as a buffer stock. The main costs of holding money balances is the interest forgone, and thus Hawtrey points to a balancing process between costs and advantages in determining desired balances. The introduction of a banking system into the model allows agents to substitute borrowing power for money balances (Hawtrey 1919a, pp. 36–7).

Hawtrey also introduces a concept of effective demand:

> The total effective demand for commodities in the market is limited to the number of units of money of account that dealers are prepared to offer, and the number they are prepared to offer over any period of time is limited according to the number they hope to receive. (1919a, p. 3)

Later, in *Trade and Credit* (1928a) Hawtrey points to a flaw in the theory of an elastic supply of labour based on marginal utilities (or disutilities) of product and effort. He argues that whilst a difference between the marginal utility of the product and the disutility of effort may prompt an additional supply of labour 'in the simple case of a man working on his own account' (1928, p. 148), this is not the general case since: 'the decision as to the output to be undertaken is in the hands of a limited number of employers, and the workmen in the industry are passively employed by them for the customary hours at the prevailing rates of wages' (1928, p. 149). In this case output decisions are based not on the gross proceeds, but on the net profit margin.

The factor of expectations is also present in Hawtrey's analysis of fluctuations. Hawtrey suggests that during a downturn in activity money balances will be reduced more quickly than they are replenished in an upswing. This is because as income drops initially consumers will draw on their balances to maintain their outlay.

There is then a further level of adjustment as changes in consumers' outlays impacts on traders. Consider an upswing: the increase in consumers' outlays will increase the nominal receipts of traders and reduce their physical stocks. Traders, finding their balances have increased can either order more stock from manufacturers or reduce their bank indebtedness. Prices will tend to rise as traders find they are unable to replenish their stocks fast enough. For Hawtrey quantity adjustments occur *before* price adjustments, indeed often the price movements result from the quantity movements. Thus 'the rise of prices, when it occurs, is caused by the activity; it is a sign that production cannot keep pace with demand' (1928, p. 156). The role of stocks in Hawtrey's theory is pivotal, in general it is quantity signals rather than price signals which are the more effective. The existence of traders' stocks means that it is nearly always possible to meet the demand for increased consumption in the short term, which implies that at least in the short term a naive proportionality between increases in the money supply and prices does not hold. Furthermore the model opens the possibility of short-run quantity adjustments in disequilibrium. Thus, argues Hawtrey:

> It is only in times of equilibrium, when the quantity of credit and money in circulation is neither increasing nor decreasing, that the relation of prices and money values to that quantity of credit and money is determined by the individual's considered choice of the balance of purchasing power appropriate to his income. . . . In practice it seldom, perhaps never, happens that a state of equilibrium is actually reached. (1919a, p. 46)

Nonetheless Hawtrey's theory of the trade cycle is money-driven. It is the fluctuations in money and credit which stimulate and support the price and quantity movements. Hawtrey argued that the periodic nature of the trade cycle was solely due to monetary factors. Traders stocks are viewed as being highly interest elastic since they are held on borrowed funds, investment in fixed capital is also interest elastic (based on a marginal efficiency of capital analysis).

Thus an increase in the rate of interest will tend to reduce the demand for credit due to a lower demand for stocks and a reduced level of new investment. If the increased rate of interest is itself the result of a decreased supply of credit then there may also be some quantitative restrictions of borrowing. To reduce their stocks traders will stop giving new orders to manufacturers, leading to a drop in the level of output which will further diminish the demand for credit, as well as the level of income and demand. Traders may reduce prices to stimulate sales to accelerate the destocking process. There is thus a tendency to a cumulative decline in output, credit and prices until the banks find themselves with excess reserves and believe it to be profitable to reduce the interest rate and expand credit. For Hawtrey, macroeconomic disequilibrium was defined in terms of monetary disequilibrium.

The solution was also therefore monetary, and in particular the short-term rate of interest (the long-term rate of interest was seen as relatively ineffective as a means of control because of its relatively slow impact on investment). Hawtrey viewed the psychological factors in the trade cycle as secondary, arguing that no amount of good news or bad could seriously affect the cycle if monetary factors were not accommodating. He also opposed the public works solution to a slump in output along similar lines – and in this respect is associated with the 'Treasury View' (see Hawtrey 1925). In later life Hawtrey did acknowledge that public works could play a role in severe depressions, but as Haberler (1939, p. 23) points out Hawtrey viewed those occasions when cheap money would fail to stimulate a revival as generally very rare – although he accepts that this was the case in the 1930s.

For Hawtrey, investment decisions were made on a Marshallian marginal productivity of capital basis. In a perfectly competitive market, the marginal return on capital employed would be equalized across every industry. In these circumstances Hawtrey identifies the 'ratio of labour saved per annum to the labour expended on first cost' as 'a physical property of the capital in use' (1913, p. 66) and as a 'natural rate' of interest. Under stable monetary conditions and in the absence of a banking system this natural rate is equal to the market rate of interest or the profit rate, as in the standard marginal efficiency of capital analysis. But changes in monetary conditions will generate changes in prices and thus profits; hence the market rate will diverge from the natural rate in the same direction as the movement in prices.

With the addition of a banking system, the actual rate of interest will depend on the behaviour of the banks, and in particular their reserve position. Thus the interest rate will diverge from the profit rate. There is a three-way equilibrium condition, relating the physical return on capital, the profit rate and the balance position of banks, that is, $N = p = r$ where N is the natural rate, p is the profit rate and r is the interest rate. An increase in the supply of money will cause a rise in prices and the availability of credit; thus $N < p$ at the same time the banks will find themselves with excess reserves and thus interest rates will tend to be lower than otherwise to stimulate borrowing, that is, $p >$ r. This will be generally expansive, demand, investment and output will all tend to rise – but the seeds of the eventual slump are already present. The rising prices and relatively low rate of interest will encourage firms to over-invest, expecting returns greater than those actually accruing. On the downward cycle N $> p$ and $p < r$.

It is worth briefly considering the relationship between Hawtrey's natural rate and that associated with Wicksell. In his early work Wicksell took the natural rate as that prevailing if loan transactions were made in kind, but he later revised this to equate the natural rate with the rate of profits received in the form of money (see Lindahl 1939, p. 261, Lindahl also discusses a physical return on capital 'natural rate' similar to Hawtrey's). Thus Wicksell's natural rate can be seen as closer to Hawtrey's profit rate. Wicksell, like Hawtrey, also associates the natural rate with an equilibrium between savings and investment and stability in the price level.

Hawtrey does not place great stress on this natural rate analysis, concentrating more on the relationship of the profit rate and the interest rate. There are also considerable practical problems in determining Hawtrey's natural rate, particularly in imperfect capital markets (see the discussion in Lindahl 1939; Haberler 1939).

Hawtrey, like most of the inter-war Cambridge economists, had a fundamental belief in the self-adjusting nature of the economic system, even though much of the analysis of the period would suggest otherwise. Hawtrey believed that the system was continually approaching or seeking an equilibrium, though in practice the next shock would come before the adjustment process was complete. However, Hawtrey's theoretical approach was to concentrate on the processes of adjustment to monetary disequilibrium.

The income/inventories approach to the trade cycle is mirrored in Hawtrey's analysis of savings and investment. For Hawtrey, savings were directed into investment opportunities by securities dealers who acted like traders, holding stocks of securities financed by bank borrowing, intermediating between the savers and investors. In the early 1930s Hawtrey developed this analysis into a model where an imbalance between savings and investment results in an unanticipated change in physical stocks of goods as a result of changes in consumers' incomes (and outlays).

Savings are the excess of consumers' income over desired consumption and are represented by investment; an increase in money balances; or purchases of goods. Net investment is defined as the total of securities sold less those bought by securities dealers. Clearly the price of securities (and by implication the long-term rate of interest) will move to achieve an equilibrium between the net amount of investment and capital raised, but planned savings can exceed the resources seeking investment, in which case the excess must flow into additional money balances or additional consumption – or vice versa. In either case an expansion or contraction of demand is set in motion. Both Saulnier (1938) and Haberler (1939) note the similarity of this analysis with that of D.H. Robertson. This aspect of Hawtrey's theory is also reviewed by Davis (1981).

Hawtrey's disequilibrium analysis where unintended changes in stocks bring about an equality of actual savings and investment, but a further chain of adjustment if intended savings and investment are not equal, is remarkably close to the modern textbook presentation of the

Keynesian equilibrium adjustment process. It is interesting therefore to briefly examine the discussions between Keynes and Hawtrey leading up to the *General Theory.* Indeed in Hawtrey's comments on the drafts of the *Treatise* he is often more 'Keynesian' than Keynes himself! (see Keynes 1973, pp. 138–69). At this stage Keynes envisages:

1. A decline in fixed investment relatively to saving.
2. A fall of prices . . .
3. A fall of output, as a result of the effect of falling prices and accumulating stocks on the minds of entrepreneurs (Letter to R.G. Hawtrey, 28 November 1930; Keynes 1973, p. 143).

The fall in output leads to a disinvestment in working capital, and eventually to a situation where total investment and prices fall too far. Once output stops declining this leads to a slight rise in prices, and, given the low level of stocks at this point, so starts the upturn. Hawtrey, on the other hand, sees a direct effect on output from the contraction in demand at unchanged prices, and criticizes Keynes for only taking account of the reduction in prices relative to costs in his fundamental equations (Keynes 1973, pp. 151–2). Hawtrey argues that 'the change in prices when it does occur is not by itself an adequate measure of the departure from equilibrium' (Keynes 1973, p. 151). And later comments that: 'A manufacturer restricts output, not because he believes that prices are about to fall, but because he cannot secure sufficient sales at the existing price' (Letter from R.G. Hawtrey, 6 December 1930; Keynes 1973, p. 165).

Prices are reduced only gradually in an attempt to boost orders, but Hawtrey also points out that it is the level of retail prices which will determine the ultimate level of sales – and this will depend on how quickly retailers pass on the manufacturers' reductions. Both Hawtrey and Keynes realize that the decline in output will rebound on savings, but do not appear to treat this as the main equilibrating factor (as in the later Keynesian theory).

The high point of Hawtrey's official career came with the Genoa International Financial Conference in 1922. The conference was concerned with the problems relating to a general return to the international gold standard after the First World War. In particular there was concern that the quantity of gold might be insufficient for a return to the system at the old pre-war parities, other concerns centred on problems relating to fluctuations in demand for monetary gold. The result was greater interest in a joint Sterling–gold standard along the lines of the gold exchange standard operated earlier by India and other countries.

Hawtrey's main suggestions adopted by the Genoa conference related to greater cooperation between central banks to manage the demand for monetary gold and to regulate credit so as to stabilize the purchasing power of gold. However, the Genoa Resolutions were never acted on, largely as a result of US scepticism, and the failure of other central banks to participate in the planned follow-up conference (see Davis 1981).

At the Treasury Hawtrey had argued that there were two primary considerations for monetary policy: the stabilization of internal prices and the stabilization of the foreign exchanges. Given the UK's status as a financial centre he argued that exchange instability was particularly damaging and would make the covering of trade finance offered through London increasingly difficult. This predisposed him towards the gold standard as the de facto most practical means of achieving exchange stability.

Though Hawtrey was aware of possible deflationary problems associated with the return to gold, he appears to have believed that the exchange rate would return to par naturally, and that the necessary adjustments would come from American inflation rather than UK deflation (see the discussion in Moggridge 1972, pp. 71–2, 91).

Despite a long and active life, Hawtrey's main theoretical contributions to economics came largely in the interwar period. His first book, *Good and Bad Trade,* was published in 1913 and sets out a view of the trade cycle which received a more rigorous theoretical treatment in *Currency and Credit* (1919a), but which remained little changed thereafter, although the debates surrounding Keynes's *Treatise* prompted some refinements and revisions, as did the experience of the 1930s depression. The last major contemporary studies of his work were Saulnier (1938), which also reviewed the theories of D.H. Robertson, F.A. von Hayek and J.M. Keynes, and Haberler (1937, 1939). Interest in Hawtrey revived in the later 1970s following his death (for example, Davis 1977, 1981, 1983; see also Deutscher 1990). Particular attention has been given to Hawtrey's role in the development of multiplier analysis; see Dimand (1997).

In the 1920s innovative monetary theory in England was largely associated with the Cambridge School and in particular D.H. Robertson and Keynes. Hawtrey with his close Cambridge contacts contributed to this work, as the correspondence with Keynes now reprinted in the *Collected Works* shows. The three were often working along similar lines in this period and their work reflects (to varying degrees) an increasing failure of conventional theory to match the problems of the age.

Selected Works

1913. *Good and bad trade.* London: Constable.
1919a. *Currency and credit.* London: Longmans.
1919b. The gold standard. *Economic Journal* 29: 428–442.
1921. *The exchequer and the control of expenditure.* London: World of Today.
1922. The Genoa resolutions on currency. *Economic Journal* 32: 290–304.
1923. *Monetary reconstruction.* London: Longmans.
1924. Discussion on monetary reform. *Economic Journal* 34: 155–176.
1925. Public expenditure and the demand for labour. *Economica* 5: 38–48.
1926. *The economic problem.* London: Longmans.
1927. *The gold standard in theory and practice*. London: Longmans.
1928a. *Trade and credit*. London: Longmans.
1928b. *Trade depression and the way out*. London: Longmans.
1930. *Economic aspects of sovereignty*. London: Longmans.
1932. *The art of central banking*. London: Longmans.
1933. Saving and hoarding. *Economic Journal* 43: 701–708.
1934. Monetary analysis and the investment market. *Economic Journal* 44: 631–49.
1937. Alternative theories of the rate of interest: Three rejoinders. *Economic Journal* 47: 436–443.
1938. *A century of bank rate*. London: Longmans.
1939. *Capital and employment.* London: Longmans.
1944. *Economic destiny.* London: Longmans.
1946a. *Economic rebirth.* London: Longmans.
1946b. *Bretton woods: For better or worse*. London: Longmans.
1949. *Western European union. Implications for the UK.* London: Royal Institute of International Affairs.
1950. *Balance of payments and the standard of living.* London: Royal Institute of International Affairs.
1954. *Towards the rescue of sterling*. London: Longmans.
1955. *Cross purposes in wage policy.* London: Longmans.
1961. *The pound at home and abroad.* London: Longmans.
1965. *An incomes policy.* Economic papers no. 4. London: Department of Economics and Management, Woolwich Polytechnic.
1967. *Incomes and money.* London: Longmans.

Bibliography

Cambridge University. 1917. *Historical register of the University of Cambridge to 1910*. Cambridge: Cambridge University Press.
Davis, E.G. 1977. The economics of R.G. Hawtrey. Carleton economic papers 77(12), Department of Economics, Carleton University.
Davis, E.G. 1981. R.G. Hawtrey, 1879–1975. In *Pioneers of modern economics in Britain*, ed. D.P. O'Brien and J.R. Presley. London: Macmillan.
Davis, E.G. 1983. The macro-models of R.G. Hawtrey. Carleton economic papers 83(4), Department of Economics, Carleton University.
Deutscher, P. 1990. *R. G. Hawtrey and the development of macroeconomics*. Ann Arbor: University of Michigan Press.

Dimand, R. 1997. Hawtrey and the multiplier. *History of Political Economy* 29: 549–559.

Haberler, G. 1937. *Prosperity and depression*. Geneva: League of Nations. Economic & Financial 1936, II. A.24.

Haberler, G. 1939. *Prosperity and depression*. 2nd ed. Geneva: League of Nations. Economic & Financial 1939, II.A.4.

Hutchison, T.W. 1953. *A review of economic doctrines 1870–1929*. Oxford: Oxford University Press.

Keynes, J.M. 1920. Review of Hawtrey's c*urrency and credit*. *Economic Journal* 30: 362–365.

Keynes, J.M. 1923. *A tract on monetary reform.* London: Macmillan. Reprinted as *The collected writings of John Maynard Keynes,* vol. 4. London: Macmillan, 1971.

Keynes, J.M. 1930. *A treatise on money,* 2 vols. London: Macmillan. Reprinted as *The collected writings of John Maynard Keynes,* vols. 5 and 6. London: Macmillan, 1971.

Keynes, J.M. 1973. *The collected writings of John Maynard Keynes,* vol. 13. London: Macmillan.

Lindahl, E. 1939. *Studies in the theory of money and capital*. London: George Allen & Unwin. Reprinted, New York: Augustus M. Kelley, 1970.

Moggridge, D.E. 1972. *British monetary policy 1924–1931*. Cambridge: Cambridge University Press.

Moggridge, D.E., ed. 1973. *The general theory and after.* Part I: Preparation. In *The collected writings of John Maynard Keynes*, vol. 13. London: Macmillan.

Robinson, J.V. 1938. The concept of hoarding. *Economic Journal* 48: 231–236.

Rouse-Ball, W.W., and J.A. Venn, eds. 1913. *Admissions to Trinity College, Cambridge, vol. 5: 1851 to 1900.* London: Macmillan.

Saulnier, R.J. 1938. *Contemporary monetary theory,* Columbia University Studies in the Social Sciences, vol. 443. New York: Columbia University Press.

The Times. 1975. Sir Ralph Hawtrey CB (Obituary), 22 March.

Hayek, Friedrich August von (1899–1992)

Bruce Caldwell

Abstract

This article reviews the major intellectual contributions of the Austrian-born Nobel laureate Friedrich Hayek. Within economics, Hayek made contributions to many areas, among them monetary theory, trade cycle theory, and capital theory. His 'knowledge-based' critique of socialism and subsequent work on 'the knowledge problem' are widely viewed as seminal contributions to economics. Hayek also did substantial work in such fields as political theory, the methodology of the social sciences, psychology and intellectual history. Finally, his writings on spontaneous orders and his 'theory of complex phenomena' anticipated later developments in such areas as complexity theory and agent-based modelling.

Keywords

Austrian business cycle theory; Austrian economics; Average period of production; Behaviourism; Böhm-Bawerk, E. von; Capital theory; Chicago school; Common law; Complements; Complex phenomena; Constitutionalism; Constructivism; Currency competition; General equilibrium; Hayek, F. A. von; Information aggregation and prices; Intertemporal equilibrium; Keynesian revolution; Knowledge problem; Liberty; Majoritarianism; Markets; Mises, L. E. von; Money; Mont Pèlerin society; Natural rate and market rate of interest; Neutrality of money; New institutional economics; Planning; Public choice; Roundabout methods of production; Rule of law; Scientism; Social justice; Socialism; Socialist calculation debate; Special interests; Spontaneous order; Substitutes; Tacit knowledge; Wicksell, J. G. K

JEL Classifications

B31

Born on 8 May 1899, the polymath economist and social theorist Friedrich August von Hayek had the good fortune to be repeatedly in the right place at the right time, crossing paths with some of the century's most brilliant economists and thinkers. He grew up in *fin de siècle* Vienna, a place and time of extraordinary intellectual vitality. Through his maternal grandfather, Franz von Juraschek, a professor of civil law and civil servant, he gained an introduction to the academic world in Vienna, and through his father, August, a medical doctor and devoted botanist, a love of biology and the sciences as well as an acquaintance with another extended community of scholars. As a student at the University of Vienna his major professor was Friedrich von Wieser, and among his classmates were Oskar Morgenstern, Gottfried Haberler, and Fritz Machlup. After finishing his studies Hayek spent 15 months in the United States where, armed with letters of introduction from Joseph Schumpeter, he encountered most of the major American economists, both those contributing to the Marginalist School as well as the leading institutionalist and business cycle analyst Wesley Clair Mitchell. When he returned he joined the *Miseskreis,* Ludwig von Mises's study circle.

In the later 1920s he published an article in German that was read by Lionel Robbins, a newly appointed professor at the London School of Economics (LSE). This led to an invitation to present some lectures, and ultimately, in 1932, to Hayek being appointed to the Tooke Chair of Economic Science and Statistics. While at the LSE Hayek would engage in debates on the leading issues in economics with some of the discipline's most important members: John Maynard Keynes and Piero Sraffa over monetary theory, Frank Knight and Nicholas Kaldor over capital theory, Oskar Lange and Evan Durbin over socialism. He was also instrumental in bringing the philosopher of science Karl Popper to the LSE.

Hayek remained at the LSE until 1950, when he moved to the Committee on Social Thought at the University of Chicago. There he counted among his colleagues Milton Friedman, Aaron Director, and George Stigler. Retiring in 1962, Hayek had successive appointments at the University of Freiburg and the University of Salzburg, returning again to Freiburg in 1977. In 1974 he was awarded, with Gunnar Myrdal, the Bank of Sweden Nobel Prize in Economic Sciences, and in 1991 the Presidential Medal of Freedom. Hayek died in Freiburg on 23 March 1992.

If Hayek was in the right place at the right time, it was usually with the wrong ideas, at least from the perspective of most of his contemporaries. He was a sharp critic of Keynes well before the onset of the Keynesian Revolution. Though he helped introduce English-speaking economists to general equilibrium theory, he claimed that a preoccupation with static equilibrium analysis would mislead economists about the true nature of a dynamic market process. He attacked socialism when most members of the intelligentsia viewed it as a preferred middle way between an apparently failed capitalist system and totalitarianisms of the communist and fascist varieties; for Hayek such thinking was 'the muddle of the middle'. When most Western democracies were embracing some form of the welfare state, he criticized the concept of social justice that provided its philosophical foundations. While most of the social sciences were moving towards more and more specialized studies, his work was increasingly integrative and multidisciplinary. The views Hayek embraced over most of his career were almost systematically out of step.

From the perspective of the early 21st century, history would judge Hayek's legacy more kindly than did many of his contemporaries. He lived to witness the collapse of the Soviet bloc, which many took as vindication of his and Ludwig von Mises's early critique of central planning. His view that a competitive market system with freely adjusting prices is an essential mechanism for coordinating social action in a world of dispersed knowledge is taken by economists as a fundamental insight. His insistence that markets be embedded in a host of other social and political institutions for their proper functioning provides a jumping off point for such diverse movements within economics as experimental investigations of market institutions, public choice and constitutional analysis, and the new institutional economics. Philosophers of mind, evolutionary biologists, and neuroscientists have been attracted to his 'connectionist' approach for understanding the development and functioning of the brain. His theory of complex phenomena and work on spontaneous orders has clear analogues in complexity theory and agent-based computational modelling (Caldwell 2004, ch. 14). If Hayek remains a controversial figure in some quarters, even his critics acknowledge the breadth and depth of his contributions. One pundit, writing in the *New Yorker* in 2000, even went so far as to call the 20th century 'the Hayek century' (Cassidy 2000, p. 45). Considering that this was only about two decades after the British Labour politician Michael Foot had referred to him as a 'mad professor', the reputational turnabout has been substantial.

Early Work

Hayek's first trip to the United States took place in 1923–24. While there he studied new work on monetary policy and the control of the business cycle; he also witnessed the policy experiments being undertaken under the auspices of the then only recently established Federal Reserve System. Hayek subsequently wrote a paper on US monetary policy in which he criticized the goal of stabilizing the general price level (Hayek 1926). According to the Austrian theory of the cycle, relative price movements play an essential role in the unfolding of the cycle, so that any policy prescription that focused solely on aggregates was judged deficient for ignoring such movements.

Hayek spelled out the Austrian approach in more detail in his first book, published in 1929, an English translation of which appeared in 1933 as *Monetary Theory and the Trade Cycle.* There he argued for a monetary approach to the origins of the cycle. Hayek claimed, first, and contra both the American institutionalists and German historical economists, that any adequate explanation of the cycle must be *theoretical,* and, further, that it must be consistent with, and presuppose the validity of, the standard equilibrium theory of the day. This poses a problem, however, for if one accepts the results of standard equilibrium theory, where prices adjust to clear markets, a question immediately arises: how can a disproportionality between the production of capital goods and consumer goods that occurs during the boom phase of the cycle occur? For Hayek, money provided the answer. Though the use of money confers substantial benefits, most evidently to facilitate trade, and thereby to encourage specialization and growth, it is also a 'loose joint' in the system of exchange: 'Money being a commodity which, unlike all others, is incapable of finally satisfying demand, its introduction does away with the rigid interdependence and self-sufficiency of the 'closed' system of equilibrium' (Hayek 1933, p. 44).

Another significant piece in this period was Hayek's paper 'Intertemporal Price Equilibrium and Movements in the Value of Money' (Hayek 1928), which is widely acknowledged as an early important contribution to the theory of intertemporal equilibrium.

Hayek Comes to the LSE

Hayek's lectures in early 1931 at the LSE were published as *Prices and Production,* a book in which he completed the task begun in *Monetary Theory and the Trade Cycle* by tracing out the effects of monetary disturbances on the economy. Using a framework developed by Knut Wicksell (1906) and further adapted by Ludwig von Mises (1924), Hayek posited a *natural rate of interest* that, in the absence of monetary factors, would just equalize the demand for capital and the supply of savings. When households save, they forgo present for future consumption. The funds are borrowed by firms for investment in more 'roundabout' methods of production which allow firms to produce more goods in the future, thereby satisfying the desires of consumers. The natural rate of interest, then, is a relative price that coordinates a community's preferences regarding present and future consumption with the production processes that create the goods.

However, in the crisis stage of the cycle, an excess of capital goods (relative to consumers' preferences) are created. This occurs because of a divergence between *the natural* and *the market rate of interest,* caused by bank lending activity. Specifically, a lowering of the market rate of interest below the natural rate leads firms to move to more roundabout methods of production, just as they would have done had there been a reduction in the natural rate. However, in this case, because there has been no change in consumers' preferences, the lengthening of production processes is not sustainable. At some point before the completion of the transition, prices for consumer goods begin to rise, which signals to firms that they have made errors. As they begin to abandon the more roundabout methods, a cyclical downturn is initiated.

Hayek's theory carried the unfortunate policy implication that there was little that policymakers could do once an economy was in a recession. Recessions were avoidable only if one could make money 'neutral' by keeping the natural rate equal to the market rate of interest. Unfortunately, no one knows what the natural rate is; only the market rate is observable. The downturn, painful as it is, is actually the system returning to equilibrium, correcting for past errors. As such, policies that attempt to address a recession by injecting money only further encourage firms to persist in their mistaken behaviours, making the ultimate downturn even more severe.

Hayek's book had a tumultuous reception. In late 1930 John Maynard Keynes published his own analysis of the problems of a monetary economy, *A Treatise on Money* (Keynes 1930), in which he also used the Wicksellian framework. Hayek's critical review of Keynes's book drew a heated response from Keynes, who also took Hayek's *Prices and Production* to task, noting famously that 'It is an extraordinary example of how, starting with a mistake, a remorseless logician can end up in bedlam' (Keynes 1931, p. 154). For a while, as John Hicks later recounted, the burning question of the day for economists was, 'Which was right, Keynes or Hayek?' (Hicks 1967, p. 203).

Others entered the fray, and the weight of the combined criticisms ultimately led both Keynes and Hayek to revise their theories. Keynes finished first, publishing *The General Theory of Employment, Interest and Money* in 1936. Hayek's initial plan was to construct a dynamic theory of a capital-using monetary economy. He worked on the book in starts and stops for the rest of the decade, finally publishing it as *The Pure Theory of Capital* in 1941. There Hayek abandoned the simplifying Böhm- Bawerkian notion of an 'average period of production', and in its place systematically explored a variety of possible relations between inputs (both those available at a given point in time and over a continuous period) and outputs (whose availability might likewise vary over time). He examined the effects of substitutability and complementarity, of the introduction of new 'inventions', both in cases in which they are foreseen and when they are not, and of whether decisions are made by a single individual or within a competitive system. A key theme of the book is that the capital structure is constantly evolving as the market continually provides new information. In that evolution, capital is rarely either so malleable as to be instantaneously transformable, or so permanent as to be incapable of being applied in a different production process.

Hayek's book made important advances in capital theory, but he never was able to accomplish his larger goal. After seven years of labour he could only provide in the closing three chapters of the book a sketch of how to integrate his capital theory into a monetary framework. As he later once put it, once you get beyond Böhm- Bawerk's simplifying assumption of an average period of production, 'things become so damn complicated it's almost impossible to follow it' (Hayek 1994, p. 141). Meanwhile Keynes's victory in the area of macroeconomics quickly became complete.

Socialist Calculation and the Knowledge Problem

In the 1920s, the British economy went through wrenching structural adjustments, and with the depression of the 1930s many among the intelligentsia came to view socialist planning as the only acceptable alternative system. Economists, some of them colleagues of Hayek's at the LSE, began issuing proposals for how to organize such a system. In 1935, Hayek entered the discussion with the publication of *Collectivist Economic Planning,* a collection of translations of essays from an earlier debate that had been initiated by Ludwig von Mises. Hayek included his mentor's essay, in which Mises argued that rational planning was 'impossible' under socialism. His point was that a monetary economy with freely adjusting market prices reveals relative scarcities among factors of production. When the means of production are state-owned, there are no prices for factors of production, and hence no signals to help socialist managers allocate resources rationally (Mises 1920).

Some socialists (for example, Dickinson 1933) responded by invoking Paretian general equilibrium theory, which they argued disproved Mises's thesis. They noted that any economic system could be represented by a system of equations, so that the only difference between a planned and a free market system lay in who was responsible for 'solving' the equations, socialist managers or private entrepreneurs. If some of the prices that the socialist managers chose were wrong, gluts or shortages would appear, signalling them to adjust the prices up or down, just as in a free market. Through such a trial and error procedure, a socialist economy could mimic the efficiency of a competitive free market system, while avoiding its many problems: wasteful competition, the market failures that attend monopoly and externalities, and an unjust income distribution (Lange 1938).

Hayek challenged this vision in a series of contributions (Hayek 1937, 1945, 1968) to what has since come to be called 'the knowledge problem'. In 'Economics and Knowledge' (1937) he pointed out that the standard equilibrium theory of his day assumed that all agents have full and correct information. In the real world, however, different individuals have different bits of knowledge, and furthermore, some of what they believe is wrong. In that world, the key question is how it comes about that the actions of individuals ever get coordinated, a question that equilibrium analysis with its full information assumption brushes aside.

Hayek posited the market as a key coordinating institution. He described the market process as operating in a world of constant change, in which freely adjusting prices are formed as the result of decisions, typically forward-looking, of literally millions of market participants. Their decisions are based in part on the vast array of prices that they confront in the market, prices that give them information about relative scarcities. But in addition, agents act on the basis of localized knowledge, knowledge of particular circumstances of time and place, some of which is tacit – that is, they cannot say why they are acting on it. Their market activity also reflects this localized knowledge, and by acting their knowledge becomes embedded in the array of market prices. In short, market activity is both price-determined (prices shape what people do) and price-determining (what people do, based on local knowledge, determines what prices are). Market prices coordinate the specific knowledge of time and place possessed by millions of market agents. Socialist schemes that involve price fixing, as many of the proposals did, would keep the communication system from working. Hayek also doubted that trial and error price adjustment methods could ever mimic the speed of adjustment produced by markets, where errors to be corrected are simultaneously profit opportunities for alert entrepreneurs. Finally, Hayek criticized the profession's focus on standard equilibrium analysis which, by concentrating on equilibrium states, obscures the competitive process by which knowledge about relative scarcities becomes known: that theory 'starts from the assumption of a "given" supply of scarce goods. But which goods are scarce goods, or which things are goods, and how scarce or valuable they are – these are precisely the things that competition has to discover' (Hayek 1968, p. 181). In short, market competition provides a discovery procedure. Hayek developed these ideas in a series of papers, the most famous of which, 'The Use of Knowledge in Society', is still widely cited by traditional general equilibrium theorists as well as economists working in the economics of information (Hayek 1945).

The Abuse of Reason Project and the Road to Serfdom

Though Hayek felt he had launched a telling attack against socialism, few in the late 1930s were persuaded by his economic reasoning. Hayek began to realize that the attractiveness of socialism went far beyond economics. Socialists promised a society that was not only more efficient than capitalism, but also one that was more just, where individuals have more self-determination and greater political freedom, and in which scientific reasoning would be used to improve upon a host of outdated social institutions. If he were successfully to challenge these utopian visions, economic arguments were not enough. He would need to develop political, historical and ethical arguments against them as well.

During the Second World War Hayek began doing just that, in a massive piece of work that he called the 'Abuse of Reason' project. His overarching goal was to show how a number of then-popular doctrines and beliefs, doctrines with which he disagreed, had a common origin in some fundamental misconceptions about the proper methods for studying social phenomena. Central to his argument was the critique of *scientism,* which he defined as the 'slavish imitation' of the methods of the natural sciences in the study of social phenomena (Hayek 1942–44, p. 24). He criticized the objectivism, historicism and collectivism of the 'scientistic prejudice', and contrasted these with his own preferred approach, one that was subjectivist, theoretical, and individualist. In the essay 'Scientism and the Study of Society' (Hayek 1942–44) he also articulated a fundamental thesis about the limitations of our knowledge in the social sciences: that rather than make precise predictions often the best we can do is to make a pattern prediction, or alternatively to provide an explanation of the principle by which some social phenomenon came into being.

Hayek never completed the Abuse of Reason project, although sections of it were published separately during and after the war. One of these became his most famous book, *The Road to Serfdom.* As noted above, many advocates of socialism had promised that socialism would bring greater political freedom. In *The Road to Serfdom* Hayek countered that planning of the economy would soon lead to increasing political control as well. One of the virtues of a market economy is that it allows people with very different tastes to express them, and (for those with the means) to get them satisfied, through the market. In a planned economy, socialist managers must decide which goods, and in what quantities, get produced. Invariably some people will not like the decisions they make, and will protest. A change in the mix will cause others to protest. If any progress is to be made, even democratically elected socialist regimes will at some point be forced simply to make the decisions for the people. This is much easier to do if political dissension is suppressed. Hayek's claim was that, to run a fully socialized planned economy successfully, its socialist managers ultimately must secure control of the political process as well.

Hayek's book was only one of many at the time to address the issues of planning versus markets and other issues related to the shape of the post-war economic and political order. Its fame, and in some quarters notoriety, was due to its being condensed in the pages of *Reader's Digest* in April 1945, appearing just as the war in Europe was coming to an end. *Reader's Digest* then had a circulation of almost nine million, and in addition, a Book of the Month Club reprint was made available that added another million readers. As a result, Hayek's little book, and the even smaller condensed version, gained widespread attention and iconic status among both its supporters and critics.

Besides fame, the publication of the book brought with it other unintended consequences. On a publicity trip to the United States, Hayek made a number of contacts, people who shared his views regarding the merits of a liberal democratic market order. In 1947 he organized the first meeting of the Mont Pèlerin Society, which brought together like-minded people from America and Europe to discuss and debate questions concerning the appropriate economic, political, legal and social institutional framework for a free society. Participants included Milton Friedman, Aaron Director and George Stigler, who would over the course of the next decade form the Chicago School of economics.

The Sensory Order

From 1945 until he joined the faculty at Chicago, Hayek took on yet another wholly different subject, theoretical psychology. Building on a student paper he had completed in 1920, he titled the resulting book *The Sensory Order* (Hayek 1952a).

This book is probably best viewed as an outgrowth of his earlier attack on scientism. Two 'objectivist' doctrines that he criticized in the 'scientism' essay were physicalism, a view espoused by the logical positivist philosopher Otto Neurath, and behaviourist psychology. The doctrines were related: physicalism insists that all truly scientific statements make reference only to observables, and behaviourist psychology likewise insists that scientific psychology should eschew all reference to mental states and deal only with observable behaviour. By eliminating all reference to subjective states and interpretations, the objectivity of science is guaranteed.

Hayek posited two orders, the sensory order that we experience, and the underlying natural order that natural science has revealed: atoms, molecules, electromagnetic waves and the like. The question arises: why are these two orders different? Hayek's answer was that the sensory order is in fact a product of our brain. He characterized the brain as a highly complex but self-ordering, hierarchical classification system, a huge network of connections. A given stimulus triggers an extensive set of neuronal firings that gives rise to our experience of a sensation. The richness of our sensory experience is due to the sheer vastness and hierarchical nature of the classifier system. As he once noted, 'During a few minutes of intense cortical activity the number of interneuronic connections actually made (counting also those that are actuated more than once in different associational patterns) may well be as great as the total number of atoms in the solar system (that is, 10^{56})' (Hayek 1964, p. 25).

If Hayek's description was right it posed problems for behaviourists, who did not even recognize the existence of the two orders, taking the sensory order as fundamental. Furthermore, the supposedly uninterpreted sensory experience so vital to the behaviouralist was itself simply a product of our minds; it was itself an interpretation. Hayek's book went virtually unnoticed when published, but subsequent neuroscientific research broadly supports his principal claims.

Political Theory

J.M. Keynes read Hayek's *The Road to Serfdom* on a boat going to the Bretton Woods conference, later writing to Hayek that 'morally and philosophically I find myself in agreement with virtually the whole of it; and not only in agreement with it, but in a deeply moved agreement' (Keynes 1944, p. 385). Keynes went on to say, though, that.

> You admit here and there that it is a question of knowing where to draw the line. You agree that the line has to be drawn somewhere, and that the logical extreme is not possible. But you give us no guidance whatever as to where to draw it. (1944, p. 386)

Hayek evidently took the criticism to heart, for in the coming years he would make two further important contributions to political philosophy that would refine and extend the arguments made in *The Road to Serfdom.*

In *The Constitution of Liberty* Hayek defined *liberty* as a condition 'in which coercion of some by others is reduced as much as possible in society' (Hayek 1960, p. 11). This poses a dilemma, because the best way to avoid coercion is to set up a coercive power that is strong enough to suppress it. Liberal constitutionalism attempts to solve the problem by defining a private sphere of acceptable individual activity, granting the state a monopoly on coercive powers, then constitutionally limiting the power of the state to those instances where it is required to prevent coercion. The state's coercive actions are limited by the rule of law: its laws made in protection of the private sphere must be prospective, known, certain, and equally enforced (Hayek 1960, pp. 205–10). He contrasted these with laws that seek particular outcomes within the private sphere, for example, price-fixing to help certain groups, or social legislation whose intent is to create or preserve a particular pattern of redistribution. Hayek linked his discussion with his perennial concern for problems caused by dispersed knowledge by noting how liberty enables individuals to make the best use of local knowledge:

> The rationale of securing to each individual a known range within which he can decide on his actions is to enable him to make the fullest use of his knowledge, especially of his concrete and often unique knowledge of the particular circumstances of time and place. The law tells him what facts he may count on and thereby extends the range within which he can predict the consequences of his action. (Hayek 1960, pp. 156–7)

In the last third of the book Hayek outlined the specific sorts of government policies that are consistent with constitutional liberalism.

Soon after completing this book he felt the need to readdress some of the same questions, ultimately producing the trilogy *Law, Legislation and Liberty* (1973–79). There Hayek lamented how Western democracies were increasingly circumventing the constitutional constraints outlined in his earlier book. Because the ideals of constitutionalism had failed to take root, the rule of law was weakening. Governments were increasingly passing coercive legislation, typically under the guise of achieving social justice, that in reality typically served well-organized coalitions of special interests. Coercive legislation was gradually replacing the rule of law.

Hayek began by contrasting spontaneous, self-generating orders (what the Greeks called a *kosmos*) with organizations that are constructed, created orders (what the Greeks referred to as a *taxis*). Agents in organizations aim at accomplishing specific goals, and do so by following explicit commands. Grown orders tend to be much more complex. They do not aim at specifiable outcomes, and agents interact in them by following abstract rules. Hayek applied these ideas to the development of the law, or *nomos*, in which rules of just conduct eventually become codified into law. He contrasted this common law heritage with legislation, the rules for organizing government, also known as *thesis.* Under the influence of various rationalist constructivist doctrines (Hayek identifies utilitarianism and legal positivism as particularly noxious), legislation to achieve particular ends began to replace the grown law, which itself does not aim at specific outcomes but instead provides a stable ordered environment in which individuals are able to employ their knowledge to make decisions.

In developing these contrasts, Hayek argued that though the concept of justice provides the foundation for notions of just conduct and ultimately of the law itself, the idea of *social justice* only has meaning within the context of a *taxis.* Only human conduct by individuals or organizations, not states of affairs or outcomes, can be called just or unjust. One must be able to hold someone responsible to apply the term. Rationalist constructivists make a fundamental error, a category mistake, to argue that it can also be applied to the outcomes of a spontaneous process, which has no specific purpose other than to allow millions of agents to pursue their own purposes. Hayek ended his trilogy with the pessimistic view that majoritarian democratic governments operating under the errors of constructivism and the guise of achieving greater social justice were increasingly replacing grown law with legislation, most of which served powerful special interests, with dire consequences for the persistence of the grown order. In the final chapter he proposed a unique political reform that aimed at increasing the independence of legislators from the influence of special interests, thereby strengthening the ideal of liberal constitutionalism. Interestingly, about the same time Hayek (1978a) also proposed an equally provocative scheme for the competing currencies that he dubbed the denationalization of money.

His final major contribution was *The Fatal Conceit* (Hayek 1988), the conceit being socialism –for Hayek the ultimate form of rationalist constructivism. The book had its origins in the late 1970s, when he tried to arrange a debate between socialists and advocates of markets on the merits of their respective systems. Though the debate never came off, the project led him to begin work on a final wide-ranging critique of socialism and constructivism. Hayek worked on the book during the early 1980s, but when his health began to fail in 1985 the philosopher W.W. Bartley III (who was also the general editor of *The Collected Works of F.A. Hayek*) stepped in to assist him. Questions have been raised about how much of the book should be attributed to Bartley and how much to Hayek, but one fundamental Hayekian claim is that the moral rules, norms, ethical precepts and practices that have led to the development of the extended market order have emerged through a process of cultural evolution. Many of these rules go against the 'natural morality' that allowed earlier humans to function successfully in small hunter-gather groups. Furthermore, because they were seldom consciously adopted and their effects are often difficult to identify, they tend to chafe against human reason, as well. Many of our moral beliefs, then, lie between, and fit uneasily with, both our instinct and our reason. This is why humans instinctively rebel against the market order, and seek to use their reason to construct an alternative.

A theme that runs throughout Hayek's work is an emphasis on the limits of our reason, and the role of rule-following in allowing us to deal successfully in a world in which knowledge is dispersed. In field after field Hayek identified spontaneous complex orders that form as the result of agents following rules. The price system represents one such an order, and, as his work on capital theory showed, if one extends the system in time it can also serve as a mechanism for the intertemporal coordination of human action. The brain is another example of a self-organizing complex order: vast networks of neuronal firings give rise to the larger phenomenon of consciousness. Within political theory, the common law tradition (as opposed to legislation) and the requirement that we follow the rule of law and obey constitutional rules are yet another manifestation of our discovering procedures that allow us to deal more successfully with the limits of our reason.

It is unfortunate that Hayek remains in some quarters a controversial figure, but it is also probably inevitable, given that so many of his key insights were formed within a context of intense political debate, and that it is difficult to separate them from that context. Even so, one hopes that his contributions on knowledge and its limits, on the role of grown institutions in helping us to overcome our ignorance, and on the workings of hierarchical networks and spontaneous self-organizing complex orders, will continue to stimulate future research.

See Also

- ▶ Austrian Economics: Recent Work
- ▶ Friedman, Milton (1912–2006)
- ▶ Mises, Ludwig Edler von (1881–1973)
- ▶ Robbins, Lionel Charles (1898–1984)
- ▶ Socialist Calculation Debate

Selected Works

1926. Monetary policy in the United States after the recovery from the crisis of 1920. In *Good money*, Part I, ed. S. Kresge; vol. 5 of *Collected works*, 1999.

1928. Intertemporal price equilibrium and movements in the value of money. In *Good money,* Part I, ed. S. Kresge; vol. 5 of *Collected works*, 1999.

1931. *Prices and production*, 2nd ed. London: Routledge, 1935.

1933. *Monetary theory and the trade cycle.* New York: Kelley, 1966.

1935. *Collectivist economic planning: Critical studies on the possibilities of socialism.* London: Routledge.

1937. Economics and knowledge. Repr. in Hayek (1948).

1941. *The pure theory of capital.* Chicago: University of Chicago Press.

1942–44. Scientism and the study of society. Repr. in Hayek (1952b).

1944. The road to Serfdom. In *The road to Serfdom: Text, documents*, ed. B. Caldwell, vol. 2 of collected works, 2007.

1945. The use of knowledge in society. Repr. in Hayek (1948).

1948. *Individualism and economic order.* Chicago: University of Chicago Press.

1952a. *The sensory order: An inquiry into the foundations of theoretical psychology.* Chicago: University of Chicago Press.

1952b. *The counter-revolution of science: Studies on the abuse of reason.* Glencoe: Free Press.

1960. *The constitution of liberty.* Chicago: University of Chicago Press.

1964. The theory of complex phenomena. Repr. in Hayek (1967).

1967. *Studies in philosophy, politics and economics*. Chicago: University of Chicago Press.

1968. Competition as a discovery procedure. Repr. in Hayek (1978b).

1973–9. *Law, legislation and liberty,* 3 vols. Chicago: University of Chicago Press.

1978a. The denationalisation of money. In *Good money,* Part II, ed. Stephen Kresge, vol. 6 of *Collected works*, 1999.

1978b. *New studies in philosophy, politics, economics and the history of ideas*. Chicago: University of Chicago Press.

1988. The fatal conceit: The errors of socialism. In *Collected works*, vol. 1, ed. W.W. Bartley III, 1988.

1988–. *The collected works of F.A. Hayek.* Chicago/London: University of Chicago Press and Routledge.

1994. *Hayek on Hayek: An autobiographical dialogue,* ed. S. Kresge and L. Wenar. Chicago: University of Chicago Press.

Bibliography

Caldwell, B. 2004. *Hayek's challenge: An intellectual biography of F.A. Hayek*. Chicago: University of Chicago Press.

Cassidy, J. 2000. The price prophet. *New Yorker* 7: 44–51.

Dickinson, H.D. 1933. Price formation in a socialist economy. *Economic Journal* 43: 237–250.

Hicks, J.R. 1967. The Hayek story. In *Critical essays in monetary theory.* Oxford: Clarendon Press.

Keynes, J.M. 1930. *A treatise on money.* Repr. as vols. 5 and 6 of *The collected writings of John Maynard Keynes.* London: Macmillan, 1971.

Keynes, J.M. 1931. *The pure theory of money: A reply to Dr. Hayek. Repr. in Contra Keynes and Cambridge: Essays, correspondence*, ed. Bruce Caldwell, vol. 9 of *The collected works F.A. Hayek,* 1995.

Keynes, J.M. 1936. *The general theory of employment, interest, and money.* Repr. as vol. 7 of Keynes (1971–89), 1973.

Keynes, J.M. 1944. *Letter, J.M. Keynes to F.A. Hayek, June 28, 1944. Repr. in Activities 1940–46. Shaping the Post-War World: Employment and commodities,* vol. 27 of Keynes (1971–89), ed. D. Moggridge, 1980.

Keynes, J.M. 1971–89. *The collected writings of John Maynard Keynes,* 30 vols, ed. A. Robinson and D. Moggridge. London: Macmillan for the Royal Economic Society.

Lange, O. 1938. On the economic theory of socialism. In *On the economic theory of socialism*, ed. B. Lippincott. Minneapolis: University of Minnesota Press.

Mises, L.V. 1920. *Economic calculation in the socialist commonwealth.* Trans. S. Adler. In Hayek (1935).

Mises, L.V. 1924. *The theory of money and credit.* Trans. H.E. Batson. London: Jonathan Cape, 1934.

Wicksell, K. 1906. *Lectures on political economy*, vol. 2. Trans. E. Classen. London: Routledge, 1935.

Hazardous Waste, Economics of

Hilary Sigman

Abstract

Public policies for hazardous waste address both current waste management and the clean-up of a legacy of contamination. Policies for current waste management should provide incentives for waste generators that are sensitive to the varying hazards posed by waste. Although conventional regulations have difficulty accomplishing this variation, alternative incentive-based policies show promise empirically. Policies for clean-up of contamination often fail to strike an appropriate balance between hazards posed by the contamination and costs of clean-up. In addition, relying on legal liability to fund these clean-ups has consequences for the costs of clean-up and possibly for the redevelopment of contaminated land.

Keywords

Brownfield sites; Compensation; Environmental economics; Environmental equity; Environmental externalities; Environmental liability; Environmental Liability Directive (EU); Hazardous waste management; Hazardous waste regulation; Hazardous waste, economics of; Hedonic property values; Legal liability; Strict liability; Superfund (USA); Waste-end taxes

JEL Classifications

Q5; Q53; K32; H23

Hazardous waste has become a major focus of environmental regulation. In the United States, spending on hazardous waste rose from only about two per cent of the compliance cost of all federal environmental regulations in 1985 to a projected 17 per cent in 2000 (U.S. EPA 1990);

its share of environmental expenses in Europe may have been comparable in 2000 (European Commission 2000). The costs arise from the management of waste from current activities and from the clean-up of a legacy of contamination. This article addresses, first, current hazardous waste management and, second, clean-up of past contamination.

Hazardous Waste Management

A wide range of industrial processes create hazardous wastes; they are also generated by commercial activities (for example, used oil and batteries from automobile repair shops) and by households (for example, used electronics). Once generated, wastes are managed in one of three ways: disposal, which usually involves placing wastes in landfills or injecting liquid wastes into underground wells; treatment (for example, incineration or stabilization), which renders the wastes less hazardous, but rarely eliminates the need to dispose some hazardous residuals; and recycling or reuse. Most hazardous waste is managed on-site by the small number of plants that generate vast quantities of waste; however, most generators create smaller quantities and use off-site commercial waste management.

The risk posed by a waste depends not only on the nature of the hazardous chemicals but also on their concentration and mobility and on the way the waste is managed. Waste generators control these variables through their output decisions, production processes, and handling of wastes, so the challenge for public policy is to create optimal incentives in all these dimensions. An efficient policy would correct many different environmental externalities, such as air pollution from incineration and groundwater and surface water contamination from land disposal. Such a policy might use taxes on environmental releases or, where feasible, impose legal liability for harms. With a competitive waste management market, these policies would not only affect waste management but also send the correct signals to waste generators to choose how much and which sorts of waste to generate.

Actual public policy for hazardous waste in developed countries tends to regulate waste management, with relatively few direct rules about the quantities or nature of wastes generated. Regulations use a traditional command-and-control approach, often requiring specific technologies (for example, specifying the thickness of liners required for hazardous waste landfills). These approaches do not provide much flexibility in tailoring the management methods to the characteristics and risks of the waste in question. For example, the United States requires that wastes be treated (often incinerated) before land disposal. These ‘land disposal restrictions’ probably account for most of the cost of hazardous waste regulations, yet economic assessments by the U.S. Environmental Protection Agency (EPA) strongly suggest that their costs greatly exceed their benefits (Sigman 2000). The land-disposal restrictions founder on their absolute nature; although the restrictions would pass a cost–benefit test for some wastes, they also apply to many wastes that are not easily treated or pose low hazards.

Some jurisdictions also impose taxes on hazardous waste. Sometimes called ‘waste-end’ taxes, these taxes vary with the quantity of waste and may depend on how waste is managed. Sigman (1996) reports empirical evidence that generators respond to waste-end taxes by reducing waste and altering their management methods. Levinson (1999) shows that the waste-end taxes levied by the United States have altered the geography of waste management. Although this evidence demonstrates the potential of incentive-based environmental policies to motivate private decisions, current taxes in the United States do not seem efficient. Indeed, Levinson (2003) argues that states practise destructive competition, specifically an inefficient ‘race to the top’, through these taxes.

Enforcement of both command-and-control waste regulation and waste-end taxes is difficult. Unlike air and water pollutants, hazardous waste is easily transported away from its source, giving rise to the possibility of illegal disposal (known as ‘midnight dumping’ in the United States and ‘fly tipping’ in the UK). Sigman (1998a) finds that rules requiring recycling or reuse of waste raise the frequency of illegal dumping. If the elasticity of illegal disposal to legal waste management costs is high enough, public policies may be counterproductive because the environmental harm from illegal disposal can be much greater than that from legal disposal.

A response to this enforcement problem is to use a deposit-refund or similar tax–subsidy combination (Fullerton and Kinnaman 1995). The policy would tax inputs or products that give rise to waste and give a refund for legal waste management that may vary with the external costs of the waste management method. For example, it might refund a modest portion of the initial tax for land disposal and a larger share for incineration. Such a system could mimic waste-end taxes without the incentives for illegal disposal. The effective tax for illegal disposal is the forfeit of the deposit; for all other activities, the effective tax is the difference between the deposit and appropriate refund. Deposit-refunds (with the refund equal to the deposit) are common internationally for products such as used batteries, electronics, and lubricating oil (OECD/EEA 2006).

The location of facilities that manage waste also raises issues. Local communities often reject these facilities because of their perceived hazards. Economists have sought to design policies that create optimal incentives in siting facilities through compensation for host communities (for example, Minehart and Neeman 2002). In assessing ‘environmental equity’, numerous studies examine whether poor people and members of minority groups disproportionately live near hazardous waste facilities, with mixed conclusions for developed countries (Hamilton 2005). Concern about the incidence of waste management costs may also be behind the Basel Convention of 1989, which restricts international shipment of hazardous wastes between developed and developing countries.

Clean-up of Contaminated Sites

Land disposal of hazardous wastes and other activities, such as storage of toxic substances for industrial processes, have created a legacy of contaminated sites that may cause damage and require clean-up. The U.S. federal policy for clean-up of abandoned contaminated sites is the Comprehensive Emergency Response Compensation and Liability Act (CERCLA), commonly known as Superfund. Superfund clean-up is mostly funded by legal liability imposed on the generators and transporters of waste and past and present owners of contaminated land. Liable parties must either undertake clean-up themselves or reimburse the government for clean-up by paying into a dedicated trust fund, which also received some tax financing in earlier years. European countries have similar programmes; a 2004 EU Environmental Liability Directive imposed additional requirements.

The appropriate level of clean-up (or ‘how clean is clean?’) has been the subject of a long-running debate. In the early years, public policies called for sites to be rendered completely clean; however, as costs have grown, greater (or at least more explicit) balancing of benefits and costs has become common. Still, decisions often fall well short of the economist’s ideal. For example, the Superfund programme sets goals that reflect biases in risk perception and political objectives as well as costs and risks to the exposed population (Hamilton and Viscusi 1999).

Quantifying the benefits of clean-up can be difficult. A substantial literature uses hedonic property value methods to evaluate the welfare effects of proximity to contaminated sites and finds large effects. In a literature survey by the U.S. EPA (2005), the studies on average find that house prices are seven per cent (or more) lower near contaminated sites. Studies also find that discovery of a site lowers local house prices. Although these results suggest that households perceive harm from contaminated sites, the much smaller literature that looks at whether clean-ups improve prices finds disappointing results; for example, Greenstone and Gallagher (2005) conclude that Superfund clean-up had minimal effect on house prices.

Another debate concerns the desirability of funding clean-up through legal liability. This funding source has advantages and disadvantages relative to use of general government revenues or a dedicated tax. Liability may create desirable incentives for current waste management, reducing the need for *ex ante* regulation of land disposal, such as the command-and-control regulation discussed above. However, unless the required clean-up spending is optimal, these incentives may not be efficient. In addition, much liability is retroactive (applying to contamination from before the clean-up law) and thus does not directly affect active waste management.

Liability also helps privatize clean-ups because liable parties may dispatch their responsibility by undertaking government-approved clean-ups. Relative to the government, private parties may have stronger incentives to control costs, better knowledge of the contamination, and greater ability to limit disruption to current economic activity at the site. However, in an effort to lower their costs, private parties may drag their feet and use political and other means to make the government choose less extensive clean-up remedies (Sigman 1998b, 2001).

Many policymakers fear that environmental liability deters redevelopment of 'brownfields', sites with potential contamination from their past use. Brownfields are a concern because they are a source of urban blight and because firms may develop relatively pristine land as a substitute for old industrial land. A buyer of a contaminated site may find itself partially or fully liable for clean-up costs; in the United States, CERCLA lists current landowners as among the potentially responsible parties. The government may choose to pursue a new owner, for example, if it has deeper pockets than the previous owner or is a private rather than a public entity. However, it is unclear that such liability deters redevelopment because it may be capitalized into land prices, as empirical research suggests (McGrath 2002).

A number of distortions may make price adjustments insufficient to restore efficient incentives for redevelopment of brownfields. Segerson (1993) argues that a distortion can arise when sellers are judgment-proof (sheltered from liability by the option of declaring bankruptcy), so a sale may increase collective private clean-up costs by exposing a buyer's deeper pockets to the government. Other studies point to adverse selection, imperfect enforcement of liability, and the effects of joint and several liability as sources of a disincentive for redevelopment of brownfields (Boyd et al. 1996; Chang and Sigman 2005). Empirical research does suggest higher vacancy rates for urban industrial land where expected liability is higher (Sigman 2006). Numerous public policies address brownfields, for example, by providing liability protections and direct subsidies to new owners of land.

Finally, liability incurs substantial legal costs, as the government sues liable parties and liable parties sue each other and their insurance carriers. Based on surveys of liable parties, Dixon (1995) estimates that as much as 30 per cent of private spending on Superfund will be transaction costs. However, these costs may be similar to the transaction costs of tort liability generally and the excess burden of the taxes that might replace liability as a funding source.

The specific form of environmental liability has also been controversial. In the European Union (EU), a debate on strict liability preceded the adoption of the 2004 Environmental Liability Directive. Under strict liability, parties are liable for any harm whereas under an 'at fault' or negligence rule parties are liable only when they fail to exercise appropriate care. The EU settled on a mixed regime in which certain hazardous activities are subject to strict liability. An extensive literature in law and economics compares these liability regimes. For hazardous waste, empirical studies have suggested that strict liability reduces accidental toxic spills and violations of hazardous waste laws (Alberini and Austin 2002; Stafford 2003).

Rules for apportioning liability with multiple defendants have been even more controversial. In the United States, courts have interpreted Superfund to impose 'joint and several' liability, meaning that the government may sue any liable party for the entire cost of clean-up at the site, regardless of that party's contribution to the contamination; the party initially held liable may then recover cost shares from other defendants. Most European countries also rely on joint and several environmental liability, but some have begun restricting its use. Joint and several liability strengthens the government's hand and increases its ability to collect 'orphan shares', the share of costs attributable to parties that are bankrupt or substantially judgment-proof. It also affects incentives for parties to settle rather than litigate, which may be favourable depending on empirical conditions (Kornhauser and Revesz 1994; Chang and Sigman 2000), and incentives for *ex ante* precaution in managing hazardous substances (Tietenberg 1989). Although some of its effects may be desirable, joint and several liability is often decried as unfair.

See Also

- ▶ Cost–Benefit Analysis
- ▶ Environmental Economics
- ▶ Liability for Accidents
- ▶ Pigouvian Taxes
- ▶ Value of Life

Bibliography

Alberini, A., and D. Austin. 2002. Accidents waiting to happen: Liability policy and toxic pollution releases. *Review of Economics and Statistics* 84: 729–741.

Boyd, J., W. Harrington, and M. Macauley. 1996. The effects of environmental liability on industrial real estate development. *Journal of Real Estate Economics and Finance* 12: 37–58.

Chang, H., and H. Sigman. 2000. Incentives to settle under joint and several liability: An empirical analysis of Superfund litigation. *Journal of Legal Studies* 29: 205–236.

Chang, H., and H. Sigman. 2005. The effect of joint and several liability under Superfund on brownfields. Working Paper No. 11667. Cambridge, MA: NBER.

Dixon, L. 1995. The transactions costs generated by Superfund's liability approach. In *Analyzing superfund: Economics, science and law*, ed. R. Revesz and R. Steward. Washington, DC: Resources for the Future.

European Commission. 2000. *Study on investment and employment related to EU policy on air, water and waste*. EC 4739/M/11452–0. Online. Available at http://europa.eu.int/comm/environment/enveco/industry_employment/investment_and_employment.htm. Accessed 14 Apr 2006.

Fullerton, D., and T. Kinnaman. 1995. Garbage, recycling, and illicit burning or dumping. *Journal of Environmental Economics and Management* 29: 78–91.

Greenstone, M., and J. Gallagher. 2005. Does hazardous waste matter? Evidence from the housing market and the Superfund program. Working Paper No. 11790. Cambridge, MA: NBER.

Hamilton, J. 2005. Environmental equity and the siting of hazardous waste facilities in OECD countries: Evidence and policies. In *International yearbook of environmental and resource economics 2005/2006*, ed. H. Folmer and T. Tietenberg. Cheltenham: Edward Elgar.

Hamilton, J., and W. Viscusi. 1999. *Calculating risks? The spatial and political dimensions of hazardous waste policy*. Cambridge, MA: MIT Press.

Kornhauser, L., and R. Revesz. 1994. Multidefendant settlements: The impact of joint and several liability. *Journal of Legal Studies* 23: 41–76.

Levinson, A. 1999. State taxes and interstate hazardous waste shipments. *American Economic Review* 89: 666–677.

Levinson, A. 2003. Environmental regulatory competition: A status report and some new evidence. *National Tax Journal* 56: 91–106.

McGrath, D. 2002. Urban industrial land redevelopment and contamination risk. *Journal of Urban Economics* 47: 414–442.

Minehart, D., and Z. Neeman. 2002. Effective siting of waste treatment facilities. *Journal of Environmental Economics and Management* 43: 303–324.

OECD (Organisation for Economic Co-operation and Development)/EEA (European Environment Agency). 2006. Database on economic instruments and voluntary approaches used in environmental policy and natural resources management. Online. Available at http://www2.oecd.org/ecoinst/queries/index.htm. Accessed 12 Apr 2006.

Segerson, K. 1993. Liability transfers: An economic assessment of buyer and lender liability. *Journal of Environmental Economics and Management* 25: S64–S65.

Sigman, H. 1996. The effects of hazardous waste taxes on waste generation and disposal. *Journal of Environmental Economics and Management* 30: 199–217.

Sigman, H. 1998a. Midnight dumping: Public policies and illegal disposal of used oil. *RAND Journal of Economics* 29: 157–178.

Sigman, H. 1998b. Liability funding and Superfund clean-up remedies. *Journal of Environmental Economics and Management* 35: 205–224.

Sigman, H. 2000. Hazardous waste and toxic substance policies. In *Public policies for environmental protection*, 2nd ed., ed. P. Portney and R. Stavins. Washington, DC: Resources for the Future.

Sigman, H. 2001. The pace of progress at superfund sites: Policy goals and interest group influence. *Journal of Law and Economics* 44: 315–344.

Sigman, H. 2006. Environmental liability and the reuse of old industrial land. Working paper. Rutgers University.

Stafford, S. 2003. Assessing the effectiveness of state regulation and enforcement of hazardous waste. *Journal of Regulatory Economics* 23: 27–41.

Tietenberg, T. 1989. Indivisible toxic torts: The economics and joint and several liability. *Land Economics* 65: 305–319.

U.S. EPA (Environmental Protection Agency). 1990. *Environmental investments: The cost of a clean environment*. Washington, DC: EPA.

U.S. EPA. 2005. *Superfund Benefits Analysis* (Draft). Online. Available at http://www.epa.gov/superfund/news/benefits.htm. Accessed 12 Apr 2006.

Health Behaviours, Economics of

Donald S. Kenkel

Abstract

The economics of health behaviours concerns decisions about smoking, diet and exercise, drinking alcohol, and other consumer choices with important health consequences. The neoclassical model of the consumer and its extensions provide the theoretical framework for most economic research on health behaviours. More recently, behavioural economic models that incorporate insights from psychology have been proposed as alternative models of health behaviours, especially behaviours involving addiction. Empirical economic research on health behaviours is extensive, and explores both the determinants and consequences of health behaviours. Welfare economic analysis of health behaviours also provides useful, if sometimes controversial, guidance for public policy.

Keywords

Addiction; Alcohol consumption; Excise taxes; Exercise; Health economics; Health information; Health production; Obesity; Smoking; Time inconsistency

JEL Classifications

I1

The economics of health behaviours concerns decisions about smoking, diet and exercise, drinking alcohol, and other consumer choices with important health consequences. The focus on consumption decisions that are usually made outside the medical care sector distinguishes the economics of health behaviours from medical care economics. Explaining decisions like smoking that eventually kill so many consumers poses intriguing challenges for economic models of rational consumers. Health behaviours also provide rich opportunities for empirical economics. Academic interest and policy relevance of the empirical findings often go hand in hand. For example, an estimate of the price-elasticity of demand for cigarettes sheds light on the relevance of economic models of addictive behaviour, and also helps policy-makers predict the impact of cigarette excise taxes.

The epidemiologic transition helps drive academic and policy interest in health behaviours. Over the course of the twentieth century the leading causes of death shifted from infectious diseases to chronic noncommunicable diseases (NCDs), including heart disease, cancer, diabetes and chronic lung disease. Because they increase the risks of NCDs, health behaviours are estimated to cause about half of all deaths in the USA. Smoking is estimated to be the leading behavioural cause of death, followed closely by diet and physical inactivity (Mokdad et al. 2004). Health behaviours are important for global health, not just for health outcomes in high-income countries like the USA. Nearly 80% of NCD deaths globally occur in low-and middle-income countries (World Health Organization 2009).

Theoretical Framework for Understanding Health Behaviours

Consumer demand for cigarettes, healthy and unhealthy foods, and other health-related goods can be studied using the standard economic model of the consumer. For example, empirical studies estimate price-and income-elasticities of health behaviours from single equation demand functions or multi-equation demand systems (Deaton and Muellbauer 1981). However, this approach misses the key feature that distinguishes these goods from other consumer goods – the health consequences. Grossman's (1972) household production model of health is a seminal contribution to both medical care economics and the economics of health behaviours.

Grossman's model formalises the insight that medical care is not a direct source of utility like other goods. Instead, the demand for medical care is derived from the more basic demand for the commodity good health. When applied to health behaviours, Grossman's model allows for joint production, where in addition to entering as a determinant of the household production of health, a good like cigarettes provides utility directly (or jointly produces some other commodity, e.g. relaxation). In health economics it is often convenient to distinguish three categories of goods: medical services, health behaviours and goods not related to health.

While some behaviours like drunk driving have immediate health consequences, more typically consumers trade off their current-period utility or disutility from the behaviour with its impact on their future health. This trade-off stems from the technology/biology of health production functions for NCDs like heart disease and cancer. These diseases often strike later in life, but the risk of disease depends on behaviours chosen decades earlier. The disease processes point to the need for an inter-temporal model of health behaviours. In a two-period, multi-period, or continuous time inter-temporal model, health behaviours depend in part on how heavily the individual discounts the future health consequences, i.e. on the individual rate of time preference. The disease processes that link behaviours and NCDs are probabilistic rather than deterministic: exercise reduces the risk of heart disease while smoking increases the risk, but some runners die young of heart disease while some smokers live long lives. In models with uncertainty, health behaviours also depend on individual risk aversion. The conventional wisdom is that differences in time-and risk-preferences might help explain a large part of observed variation in health behaviours. However, the conventional wisdom has been difficult to confirm or refute, because many data sets on health behaviours do not include measures of time-or risk-preferences.

Becker and Murphy's (1988) model of rational addiction provides a framework for the analysis of many health behaviours including smoking, alcohol abuse, illegal drug use and obesity. In their inter-temporal model, current consumption of an addictive good not only changes future health, but also changes the utility that the consumer receives from future consumption of the addictive good. Increasing the marginal utility of future consumption is a necessary but not sufficient condition for addiction, which they define as occurring when an increase in current consumption increases future consumption. The model assumes that the consumer is rational with a consistent plan to maximise lifetime utility. Moreover, the rational addict is forward-looking and anticipates the impact of current consumption on the marginal utility of future consumption. The implication that the rational addict's current consumption responds to anticipated future prices provides an empirical test that distinguishes rational from myopic addiction (Becker et al. 1994).

The model of rational addiction predicts many features of addiction and provides new insights. Powerful complementarities across time lead to unstable steady states, corresponding to high levels of consumption by addicts and low or zero levels of consumption by abstainers. Becker and Murphy (1988, p. 682) show that their analysis 'implies the common view that present-oriented individuals are potentially more addicted to harmful goods than future-oriented individuals'. The model also implies that the long-run price-elasticity of demand for an addictive good exceeds the short-run price-elasticity. A commonly overlooked result is that because current and future consumption are complements, in response to permanent price changes 'the long-run demand for addictive goods tends to be more elastic than the demand for nonaddictive goods' (Becker and Murphy 1988, p. 695). They suggest that the conventional wisdom that demand for addictive goods is inelastic might reflect observations of temporary instead of permanent price changes.

While the theory of rational addiction has been very influential, behavioural economic models have also been proposed as alternative frameworks to study health behaviours. Gruber and Koszegi (2001) develop a model of time-inconsistent addiction. The model incorporates insights from psychological and behavioural economics research which suggest that consumers have an extra bias for the current period over the future. This results in time inconsistency: the marginal rate of substitution between periods $t + 1$ and $t + 2$ is different from the perspective of time t (when both $t + 1$ and $t + 2$ are in the future) than it will be from the perspective of time $t + 1$, when period $t + 1$ is the current period. Gruber and Koszegi modify the rational addiction model to allow for quasi-hyperbolic discounters with time inconsistency. Consumers who are sophisticated about their own preferences may adopt commitment devices to prevent time-inconsistent choices.

Bernheim and Rangel (2004) develop an alternative behavioural economics model of cue-triggered addiction. In their model the consumer operates in either a cold or hot mode of decision-making. In the cold mode, properly functioning decision processes lead consumers to choose their most preferred alternatives. In the hot mode, the consumers' decision-making processes are dysfunctional. While in the hot mode, consumers make mistakes and use the addictive good, even though in the cold mode they would decide against use. Environmental cues trigger the hot mode and mistaken usage. However, the model also assumes that consumers understand their susceptibility to cues and thus can to some extent manage their potential addiction; for example, a recovering addict might avoid places or people associated with her former use.

Empirical Research on Determinants of Health Behaviours

Empirical research on the determinants of health behaviours is driven by the common interest of academics and policy-makers. Not surprisingly, a large body of economics research focuses on the role of prices as determinants of health behaviours. Gallet and List (2003) report a meta-analysis of 523 price-elasticity estimates from 86 empirical studies of the demand for cigarettes. Wagenaar et al. (2009) report a meta-analysis of 1003 price-elasticity estimates from 112 studies of the demand for alcoholic beverages. The price-elasticity estimates are often cited by policy-makers to make the case that higher excise taxes on cigarettes and alcohol can promote public health. With increased policy interest in obesity, the price-elasticities of calorie-dense foods, such as fast foods, become policy-relevant as new taxes are considered (Chou et al. 2004). The price-elasticity of the demand for illegal drugs is extremely relevant to the effectiveness of supply-side drug policies, but economic research on illegal drug markets faces substantial data challenges (Manski et al. 2001). Health behaviours like exercise often do not involve explicit monetary prices, but time costs play a similar role in determining these health behaviour choices (Meltzer and Jena 2010).

Economic and public health research points to health information as another important determinant of health behaviours. Perhaps the most compelling lesson that information can change health behaviour comes from the history of smoking over the last half of the twentieth century. As medical research established smoking's health risks and the information was disseminated to consumers, adult smoking prevalence fell dramatically in many high-income countries. In the USA, the prevalence of adult smoking fell from nearly 50% in the 1940s to its current rate of around 20%. Behaviour-changing information flows from public policies, non-profit organisations like the American Cancer Society, and for-profit firms in the private sector. Empirical health economic research has explored public sector information initiatives ranging from the 1964 Surgeon General's Report on smoking (Schneider et al. 1981) to New York City's 2008 required calorie posting in restaurant chains (Bollinger et al. 2011). In the private sector, manufacturers have strong profit incentives to provide information about the health benefits of their products. Research suggests that producer-provided health information promoted healthier behaviours related to dietary fibre in breakfast cereals (Ippolito and Mathios 1990), the consumption of fats, saturated fats, and cholesterol (Ippolito and Mathios 1995), and pharmaceutical products that aid smoking cessation (Avery et al. 2007).

Especially for adolescents, peer influences may be important determinants of a range of health behaviours, including smoking, drinking and illegal drug use. Many discussions of peer influences implicitly assume that they are what Liebenstein (1950) calls "bandwagon effects," where an individual's marginal utility from consumption is higher when their peers also consume the good. However, based on psychological research there may also be what Liebenstein (1950) calls 'snob effects', where an individual's marginal utility from consumption is lower when certain other people, e.g. adolescents in another peer group or adults, also consume the good. To date, most empirical work in health economics focuses on the basic question of whether the associations documented between an individual's health behaviours and their peers' behaviour reflect causality. There are three reasons these correlations are not sufficient evidence of causal peer effects as defined in economic models. First, in what Manski (1993) calls the reflection problem, peer influences go both ways, so individuals' behaviours and those of their peers are simultaneously determined. Second, peer groups may be endogenously chosen based on individual preferences over risk, time and social deviancy. This creates a selection problem, where it is difficult to know whether a shared behaviour like substance use stems from the common preferences or from peer influences. Third, peers may experience unobserved common environmental factors such as family background, school and market-level influences. As an example of a study that attempts to address these three problems, Clark and Loheac (2007) use rich data from Add Health to explore peer effects on young people's choices about smoking, drinking and marijuana use. Other approaches to study peer effects rely on natural experiments (Kremer and Levy 2008; Carrell et al. 2010), field experiments (Babcock and Hartman 2010), or social policy experiments (Kling et al. 2007).

Additional lines of empirical research on the determinants of health behaviours explore gradients with various aspects of socioeconomic status, including schooling, income, social class and race/ethnicity. These gradients also attract attention from policy-makers concerned about health disparities. There are some very strong gradients. For example, in the USA in 1997–98, about 38% of men and 30% of women with less than a high school education were current smokers, compared to only 9% of men and 8% of women with graduate degrees (Schoenborn et al. 2003). As with peer influences, a central focus of empirical health economics research is on whether the gradients and associations between socioeconomic status and health behaviours reflect causation. An important concern is that there might be 'hidden third variables' that are the true causes of both schooling and health behaviours. For example, people with a low rate of time preference are more willing to forego current utility and invest more in both schooling and healthier behaviours that only pay off in the future (Farrell and Fuchs 1982). To estimate the causal treatment effects of schooling on health behaviour, recent research adopts an identification strategy and instrumental variables (IVs) developed in labour economics to estimate the earnings returns to schooling (Card 2001). Although the IV studies provide evidence on whether associations between schooling and health behaviours are causal (Grossman 2006), they provide less guidance about the specific causal mechanisms involved. Many of the studies implicitly or explicitly assume that the causal mechanism is through schooling improving health information. From descriptive data, Cutler and Lleras-Muney (2010) suggest that health information differences account for small parts of schooling–health behaviour gradients, confirming the earlier results of Kenkel (1991).

One line of academic research that attracts less policy interest is empirical studies that attempt to distinguish and test different theoretical models of health behaviours. Although a line of empirical studies report results consistent with rational addiction to cigarettes, alcohol, cocaine and coffee, the same empirical tests yields evidence of rational addiction to milk, eggs and oranges (Auld and Gootendorst 2004). It has also been difficult to test the rational addiction model versus models with time-inconsistent addictions. In an empirical study of cigarette demand, the models' predictions are so similar that Gruber and Koszegi (2001, p. 480) conclude that 'we are unable to empirically distinguish the two with our data'. The use of commitment devices provides another empirical test of rational versus behavioural addiction models. Although psychological and behavioural economic research often relies on experimental laboratory evidence, research on the use of commitment devices in the field is beginning to emerge. Gine et al. (2010, p. 229) find that a voluntary commitment contract helped some smokers to quit, and suggest that their results 'are driven by a subset of smokers, with time-inconsistent preferences, who are (partly) sophisticated about their self-control problems'.

Empirical Research on Consequences of Health Behaviours

Empirical health economics research also explores the consequences of health behaviours. The health consequences are generally seen as the domain of public health and medical research, so economic research has mainly focused on consequences in the labor market and in the medical care market. Policy makers and advocates often focus on empirical research that documents the costs of unhealthy behaviours in economic terms, for example the lost wages and medical expenditures due to smoking, obesity or illegal drug use. Estimating the extent to which a health behaviour causes, and is not merely associated with, lower wages or higher medical expenditures is once again a challenging empirical exercise. Applications of the IV method face the usual problems of finding strong and valid instruments to identify the causal effects of health behaviours (Auld 2006). Structural econometric methods provide more complete models of the channels through which health behaviours might affect wages and medical expenditures, but still face difficult econometric identification problems.

Some of the consequences of health behaviours are private costs experienced by the consumers themselves, while others are external costs that they impose on others in society. Manning et al. (1991) conduct a seminal empirical study of the external costs of smoking, heavy drinking, and sedentary lifestyles. While some of their empirical conclusions were controversial and stimulated further research, from the perspective on neoclassical welfare economics the basic distinction between private versus external costs is not controversial. However, the distinction can be blurry in empirical practice. For example, are the costs to smokers' family members or drunk drivers' passengers private or external? From the perspective of behavioural economics, the distinction becomes even more blurred, if due to time-inconsistent unhealthy decisions consumers impose 'internalties' on their future selves.

Welfare Economics of Health Behaviours

In neoclassical welfare economics, government interventions to change health behaviours improve social welfare either by correcting market failures or by making outcomes more equitable and just. One set of market failures involves negative externalities from health behaviours. The victims of secondhand smoke and drunk drivers provide dramatic examples of negative externalities that could be corrected either by excise taxes on tobacco and alcohol, or other policies such as public smoking bans and drunk driving laws. When unhealthy behaviours increase medical costs they also impose costs on other members of the insured consumers' private-or public-sector insurance pools. This can be viewed as a form of *ex ante* moral hazard, where insurance changes consumers' incentives for health behaviours. Whether *ex ante* moral hazard occurs can be shown to depend crucially upon whether the price of insurance reflects the health behaviours (Ehrlich and Becker 1972; Zweifel and Breyer 1997). When smokers are charged higher rates for their health insurance, for example as allowed in US insurance markets under the 2009 Affordable Care Act, consumers will have the correct incentives to avoid smoking because it will lower their insurance premium. Employers also have incentives to promote healthier behaviours, e.g. through worksite wellness programs, to reduce their workers' medical care costs and absenteeism.

Imperfect consumer information is another market failure that justifies interventions to promote healthier behaviours. The empirical evidence reviewed above, that new information changes health behaviours, also implies that important gaps in information have been common. Quantifying how well or how badly consumers are informed about health behaviours is difficult. In another controversial study that prompted much research, Viscusi (1990) provides evidence that smokers tend to overestimate the risk that smoking causes lung cancer.

The gradient with socioeconomic status suggests that there might be scope for interventions targeted at the health behaviours of disadvantaged groups to reduce health disparities and enhance equity. Targeted school health education and mass media campaigns could address information gaps that contribute to unhealthy behaviours among disadvantaged consumers. Enhancing public safety in poor neighborhoods might also lead to more physical activity and reduce obesity. However, other policies that change the health behaviour-environment, such as higher taxes or reduced density of alcohol outlets or fast food restaurants, impose costs on the disadvantaged consumers. From the perspective of neoclassical welfare economics, it is not clear that social welfare can be enhanced when policies to reduce health disparities impose costs on the very consumers suffering the disparities. In recent years social scientists have also realised that health disparities can be an unintended consequence of scientific progress (Link and Phelan 1995). As scientific advances provide new information, such as the link between smoking and health, it may be difficult to avoid at least temporary increases in health disparities. A more puzzling, and more troubling, pattern is when disparities persist or even widen longer after the advance, which appears to be the case for smoking.

Part of the reason that the rational addiction model is controversial is its implications for welfare economics. Gruber and Koszegi (2001) emphasise that while they are unable to empirically distinguish the rational addiction model from their time-inconsistent model, the two models lead to radically different policy prescriptions. Even when there are no externalities and consumers are perfectly informed, government policies to prevent time-inconsistent health behaviours are for the consumers' 'own good'. Bernheim and Rangel (2004) emphasise that their model of cue-triggered addiction yields yet different policy prescriptions that depend upon whether the policy affects decisions made in the cold or hot modes. Bernheim and Rangel (forthcoming) highlight the ethical issues raised when the doctrine of revealed preference – which underlies neoclassical welfare economics – is relaxed: 'If we can classify, say, the consumption of an addictive substance as contrary to an individual's interests, what about choices involving literature, religion, or sexual orientation?'. They argue against the view that any departure from the doctrine of revealed preference makes welfare economics infeasible or subjective, but a standard approach to welfare economics with non-standard decision-makers has yet to emerge. Economists are on firmer ground when they give advice about policies to improve health behaviours based on correct neoclassical market failures and the existence of health disparities.

See Also

- ▶ Health Economics
- ▶ Health Insurance, Economics of
- ▶ Health Outcomes (Economic Determinants)

Bibliography

Auld, M.C. 2006. Using observational data to identify the causal effects of health-related behaviour. In *The Elgar companion to health economics*, ed. A.M. Jones. Cheltenham: Edward Elgar.

Auld, M.C., and P. Grootendorst. 2004. An empirical analysis of milk addiction. *Journal of Health Economics* 23(6): 1117–1133.

Avery, R.J., D.S. Kenkel, D. Lillard, and A.D. Mathios. 2007. Private profits and public health: Does advertising smoking cessation products encourage smokers to quit? *Journal of Political Economy* 115(3): 447–481.

Babcock, P.S., and J. Hartman. 2010. Networks and workouts: Treatment size and status specific peer effects in a randomized field experiment. *National Bureau of Economic Research Working Paper 16581*.

Becker, G.S., and K.M. Murphy. 1988. A theory of rational addiction. *Journal of Political Economy* 96(4): 675–700.

Becker, G.S., M. Grossman, and K.M. Murphy. 1994. An empirical analysis of cigarette addiction. *American Economic Review* 81(2): 237–241.

Bernheim, B.D., and A. Rangel. forthcoming. Behavioral public economics: Welfare and policy analysis with fallible decision-makers. In *Institutions and behavioral economics*, ed. Peter Diamond and Hannu Vartianen.

Bernheim, B.D., and A. Rangel. 2004. Addiction and cue-triggered decision processes. *American Economic Review* 94(5): 1558–1590.

Bollinger, B., P. Leslie, and A. Sorensen. 2011. Calorie posting in chain restaurants. *American Economic Journal: Economic Policy* 3: 91–128.

Card, D. 2001. Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica* 69: 127–160.

Carrell, S.E., M. Hoestra, and J.E. West. 2010. Is poor fitness contagious? Evidence from randomly assigned friends. National Bureau of Economic Research Working Paper 16518.

Chou, S.-Y., M. Grossman, and H. Saffer. 2004. An economic analysis of adult obesity: Results from the Behavioural Risk Factor Surveillance System. *Journal of Health Economics* 23(3): 565–587.

Clark, A., and Loheac. 2007. 'It wasn't me, it was them' social influences on risky behaviour by adolescents. *Journal of Health Economics* 26(4): 763–784.

Cutler, D.M., and A. Lleras-Muney. 2010. Understanding differences in health behaviours by education. *Journal of Health Economics* 29(1): 1–28.

Deaton, A., and J. Muellbauer. 1981. *Economics and consumer behaviour*. Cambridge: Cambridge University Press.

Ehrlich, I., and G.S. Becker. 1972. Market insurance, self-insurance, and self-protection. *Journal of Political Economy* 80: 623–649.

Farrell, P., and V. Fuchs. 1982. Schooling and health: The cigarette connection. *Journal of Health Economics* 1: 217–230.

Gallet, C.A., and J.A. List. 2003. Cigarette demand: A meta-analysis of elasticities. *Health Economics* 12(10): 821–836.

Gine, X., D. Karlan, and J. Zinman. 2010. Put your money where your butt is: A commitment contract for smoking cessation. *American Economic Journal: Applied Economics* 2: 213–235.

Grossman, M. 1972. On the concept of health capital and the demand for health. *Journal of Political Economy* 80(2): 223–255.

Grossman, M. 2006. Education and nonmarket outcomes. In *Handbook of the economics of education*, ed. E. Hanushek and F. Welch. Amsterdam: North-Holland.

Gruber, J., and B. Koszegi. 2001. Is addiction 'rational'? Theory and evidence. *Quarterly Journal of Economics* 116(4): 1261–1305.

Ippolito, P., and A.D. Mathios. 1990. Information, advertising and health: A study of the cereal market. *Rand Journal of Economics* 21(3): 459–480.

Ippolito, P., and A.D. Mathios. 1995. Information and advertising: The case of fat consumption in the United States. *American Economic Review: Papers and Proceedings* 85(2): 91–95.

Kenkel, D.S. 1991. Health behaviour, health knowledge, and schooling. *Journal of Political Economy* 99(2): 287–305.

Kling, J., J. Liebman, and L. Katz. 2007. Experimental analysis of neighborhood effects. *Econometrica* 75(1): 83–119.

Kremer, M., and D. Levy. 2008. Peer effects and alcohol use among college students. *Journal of Economic Perspectives* 23(3): 189–206.

Liebenstein, H. 1950. Bandwagon, snob, and Veblen effects in the theory of consumer demand. *Quarterly Journal of Economics* 64: 183.

Link, B.G., and J. Phelan. 1995. Social conditions as fundamental causes of disease. *Journal of Health and Social Behaviour* (Extra Issue): 80–94.

Manning, W.G., E.B. Keeler, J.P. Newhouse, E.M. Sloss, and J. Wasserman. 1991. *The costs of poor health habits*. Cambridge, MA: Harvard University Press.

Manski, C.F. 1993. Identification of endogenous social effects: The reflection problem. *Review of Economic Studies* 60(3): 531–542.

Manski, C.F., J. Pepper, and C. Petrie. 2001. *Informing America's policy on illegal drugs: What we don't know keeps hurting us*. Washington, DC: National Academies Press.

Meltzer, D., and A.B.. Jena. 2010. The economics of intense exercise. *Journal of Health Economics* 29(3): 347–352.

Mokdad, A.H., J.S. Marks, D.F. Stroup, and J.L. Gerberding. 2004. Actual causes of death in the United States: 2000. *JAMA* 291(10): 1238–1245.

Schneider, L., K. Benjamin, and K.M. Murphy. 1981. Governmental regulation of cigarette health information. *Journal of Law and Economics* 24(3): 575–612.

Schoenborn, C.A., J.L. Vickerie, and P.M. Barnes. 2003. Cigarette smoking behaviour of adults: United States, 1997–98. Advance data from *Vital and Health Statistics* no. 330.

Viscusi, W.K. 1990. Do smokers underestimate risks? *Journal of Political Economy* 98(5): 1253–1269.

Wagenaar, A.C., M.J. Salois, and K.A. Komro. 2009. Effects of beverage alcohol price and tax levels on drinking: A meta-analysis of 1003 estimates from 112 studies. *Addiction* 104(2): 179–190.

World Health Organization. 2009. *Global health risks: Mortality and burden of disease attributable to selected major risks*. Geneva: WHO Press.

Zweifel, P., and F. Breyer. 1997. *Health economics*. New York: Oxford University Press.

Health Econometrics

Andrew M. Jones

Abstract

The term *health econometrics* has been adopted to describe the development and application of econometric methods within health economics. This article outlines the distinctive issues that arise in applying econometrics to health data and how these applications have helped to shape the broader literature.

Keywords

Econometrics; Evaluation; Health economics; Microeconometrics

JEL Classifications

C1; I1

The term *health econometrics* has been adopted as a convenient shorthand to describe the development and application of econometric methods within health economics. The challenges posed by health data have stimulated important methodological innovations.

There has been a dramatic growth in econometric studies that use health data. This has stimulated developments in econometric methodology that have spread beyond health economics. The label 'health econometrics' was adopted for a chapter in the *Handbook of Health Economics* (Jones 2000) although earlier authors had reviewed the use of econometrics in the field (Newhouse 1987; Wagstaff 1989). Feldstein (1967) was an early exponent, using data on British hospital costs, and the RAND Health Insurance Experiment (HIE) was the catalyst for much of the methodological innovation that took place in the 1970s and 1980s (Manning et al. 1987b). The scale of activity has grown substantially over subsequent decades. The European Workshops on Econometrics and Health Economics (http://www.york.ac.uk/res/herc/research/ew/), established in 1992, have provided a focus for developments and networking by researchers in the field. These have been complemented by similar meetings in North America (since 2009) and Australasia (since 2010).

The majority of applications within health econometrics have measures of healthcare or health as the outcomes of interest. The latter are often self-reported but can include clinical outcomes, anthropometric data and, increasingly, biomarkers. Other studies focus on health-related behaviours such as diet, smoking, drinking and illicit drug use, and econometric methods are used to analyse the results of contingent valuation and discrete choice experiments and in the context of cost–benefit and, more often, cost-effectiveness analysis. Studies of micro-data far outweigh those using macro time-series data. Recent years have seen greater emphasis on longitudinal datasets, such as panel and cohort studies, and large-scale administrative datasets that are often linked to each other or to social surveys (e.g. Black et al. 2007)

Jones (2000) provides a comprehensive review of the literature up to 2000. This is updated in Jones (2009), with a particular emphasis on the use of longitudinal data and on the role of study design and credible identification strategies. Reviews of specific areas of health econometrics include: methods for modelling individual health care costs (Manning 2006; Jones 2011); simulation-based estimation and inference (Contoyannis et al. 2004a); methods for evaluating treatment effects (Auld 2006; Jones and Rice 2011). Textbook treatments of health econometrics are provided at an introductory level in Jones (2007) and more advanced level in Jones et al. (2007).

The RAND Health Insurance Experiment is important in many respects, not least as an early example of a large-scale social experiment – a methodology that has seen a resurgence of interest in recent years particularly within development economics (e.g. Miguel and Kremer 2004). In terms of econometric methods the RAND study focused on modelling healthcare use and expenditure and is particularly associated with the development of the two-part, or multi-part, model (2 PM) and with bringing attention to the problem of retransformation bias.

Healthcare use or expenditure is typically characterised by large numbers of zero observations (non-users), a heavily skewed distribution with a long right-hand tail and, in the context of regression models, substantial heteroscedasticity. The RAND approach to modelling the zeros was to adopt a 2 PM in which the probability of any use and the conditional distribution of costs among users are modelled independently (Duan et al. 1983; Manning et al. 1987a; Keeler et al. 1988). At the time this provoked criticism from those who advocated the use of sample selection models (e.g. Hay and Olsen 1984). The gist of the debate reflected differences in opinion about the purpose of the modelling exercise: whether it is about reliable predictions of overall costs, conditional on the covariates, or about separately identifying the impact of covariates on each of the parts.

One response to the high degree of skewness in health care costs is to estimate regression models for transformations of costs, the most popular being the logarithm. However, the economic interest lies in predicted costs on the original scale, which requires a retransformation of the predictions from the log model. Duan (1983) provided a nonparametric smearing estimator which has to be adapted when there is heteroscedasticity in the data on the log-scale (Manning 1998; Mullahy 1998; Manning and Mullahy 2001).

Retransformation bias can be avoided by specifying nonlinear regressions. For example, using an exponential specification of the conditional mean allows the impact of regressors to be multiplicative rather than additive, as in proportional hazards models for survival data. Much recent work has adopted the Generalised Linear Models (GLM) framework (e.g. Blough et al. 1999). This uses a quasimaximum likelihood approach to estimate distributions within the linear exponential family, with conditional mean specified through a link function and the conditional variance specified as a function of the mean. The flexibility of the GLM framework has been increased by Basu and Rathouz's (2005) extended estimating equations approach that allows the form of link function and distribution to be estimated from the data rather than specified a *priori*.

The GLM framework is a two-parameter approach in the sense that the GLM distributions are fully characterised by their mean and variance. A strategy to allow for more flexible distributions exploits the analogy between the skewed and heavy tailed distributions that are required for cost data and the parametric distributions that are typically applied in survival analysis. For example, Manning et al. (2005) advocate the use of the generalised gamma distribution. This is a heavily parametric approach; semiparametric alternatives include the use of finite mixture models (e.g. Conway and Deb 2005) and the conditional density estimator suggested by Gilleskie and Mroz (2004).

Cost regressions play a role in health technology assessment and cost-effectiveness analysis (Hoch et al. 2002). They also find concrete applications in regression based algorithms for risk adjustment (Van de Ven and Ellis 2000) and in weighted capitation formulas for geographic resource allocation (Smith et al. 2001). In both cases regression models are used to predict health care costs for individuals or groups on the basis of their diagnostic, demographic and socioeconomic characteristics. As the methods are often applied to very large administrative datasets simple linear regression often performs as well as more elaborate nonlinear models.

Many of the outcomes of interest for health economists are inherently categorical or are censored or truncated. Binary outcomes are widespread and are typically modelled using probit or logit models. Ordered categorical outcomes are often encountered in measures such as self-assessed health and are typically specified as ordered probits or logits (e.g. Kerkhofs and Lindeboom 1995). Outcomes often reflect multinomial choices. These can arise naturally with the choices that have to be made by users of health care, such as the selection of an insurance plan or the choice of a health care provider, or they may reflect the hypothetical options offered within discrete choice experiments. Early studies of plan choice and of the demand for medical care typically adopted a multinomial logit specification or, to relax the independence of irrelevant alternatives assumption, nested logit models (e.g. Dor et al. 1987; Dowd et al. 1991). The nested logit applies when choices can be organised into a meaningful nesting. Developments in computational methods and simulation-based estimation have extended the repertoire of multinomial choice models to include the mixed logit and multinomial probit models (e.g. Risa Hole 2008).

Count data regression is designed for outcomes that are measured as nonnegative integers and is naturally suited to the number of visits to medical practitioners. Studies based on health data have provided a test-bed for many of the innovations in the econometrics of count data regression most notably the adoption of the negative binomial (negbin) specification as an extension of the standard Poisson model (e.g. Cameron and Trivedi 1986; Cameron et al. 1988). Like health care costs, count data often exhibit excess zeros – a higher frequency of zeros than would be predicted by a Poisson distribution. This may reflect overdispersion due to unobserved heterogeneity, which can be captured using a mixture distribution such as the negbin, but may be better handled by zero-inflated or hurdle specifications that add extra weight to the probability of observing a zero (Mullahy 1986, 1997b; Pohlmeier and Ulrich 1995). The hurdle model was extended to take account of the distinction between visits and multiple sickness spells in Santos Silva and Windmeijer (2001).

Deb and Trivedi (1997) proposed the use of a latent class, or finite mixture, specification to incorporate unobserved heterogeneity within cross-section models for doctor visits. The intuition behind this approach is that the observed data are sampled from a mixture of unobserved sub-populations, each of which can be modelled using a parametric count data model. With panel data the heterogeneity can be captured by an individual effect: for example, Van Ourti (2004) adopts a Gaussian random effects specification to model count data. Bago d'Uva (2006) brings these ideas together and models unobserved heterogeneity within a panel to estimate a latent class hurdle model. A challenge for finite mixture models is to choose the appropriate number of latent classes. Jochmann and Leon-Gonzalez (2004) adopt a semiparametric Bayesian approach to this problem using a Dirichlet process mixture in which the number of classes is estimated directly as part of a Bayesian MCMC algorithm.

Estimation of count data models is problematic when there are endogenous regressors; a problem addressed by Mullahy (1997a), Windmeijer and Santos Silva (1997) and Terza (1998), who use GMM and two-step estimators. More generally, the potential for bias due to selection on unobservables has always been a central concern for empirical research in health economics; it lies behind the experimental approach adopted in the RAND HIE and was addressed in pioneering work on health production such as Auster et al. (1969), Grossman (1972) and Rosenzweig and Schultz (1983).

In linear models for panel data time-invariant unobservables can be handled as 'fixed effects' by taking first differences or mean deviations. This underlies the widespread use of difference-in-differences methods in health economics (see Jones 2009). Nonlinear models for qualitative and limited dependent variables pose more of a challenge. Contoyannis et al. (2003) use maximum simulated likelihood estimation to estimate models that allow for both an individual effect and autocorrelated error terms within panel probit models for health problems. Contoyannis et al. (2004b) estimate dynamic models for ordered measures of self-assessed health, handling the initial conditions problem explicitly by specifying the relationship between the individual effect and observed regressors. Deb (2001) adopts a semiparametric approach, avoiding assuming a parametric distribution of the individual effect by using a finite density estimator, and Grootendorst (1997) also applies a semiparametric approach using the pantob estimator for censored data.

Multivariate models bring together equations that are related through common unobservable factors. The computational problem of specifying the joint distribution of (multiple) outcomes and (multiple) treatments can be handled by methods such as maximum simulated likelihood (MSL) and Bayesian MCMC, and using copulas. In a series of papers Pravin Trivedi and co-authors model health care expenditure or utilisation along with multinomial choices of insurance coverage allowing for unobservable factors that may influence both choice of insurance plan and use of health care. Deb and Trivedi (2006) assume a parametric distribution for the latent factors and estimate the joint distribution by MSL. Zimmer and Trivedi (2006) use copulas to model the joint distribution by binding together marginal distributions for the outcomes of interest through the copula function. The Bayesian MCMC approach provides a natural way of handling the computational problem posed by systems of limited dependent variables, especially when the latent variables are handled as missing data through the data augmentation approach (e.g. Hamilton 1999; Deb et al. 2006).

Theoretical models in health economics, such as the Grossman (1972) model of the demand for health, lend themselves to econometric specifications that involve simultaneous equation models for latent variables. Applied work in this area in the 1970s and 80s adopted structural equations modelling (SEM) implemented through linear structural relationships (LISREL) and the multiple indicators and multiple causes (MIMIC) model (e.g. Wolfe and van der Gaag 1981; Van Vliet and van der Gaag 1982; van Doorslaer 1987). The strong parametric assumptions required in these models and the need for reliable and comprehensive sets of indicators lead to a decline in their use, but there has been a resurgence of interest in recent literature on early life development and health (e.g. Heckman 2012). Arcidiacono et al. (2007) is a notable example of a structural approach, in which the econometric specification is derived explicitly from a theoretical model.

Much modern applied work in health economics sets out to identify causal mechanisms and 'treatment effects'. Developments over the past couple of decades have raised the bar in terms of the need for rigorous definition of the treatment effects of interest, usually formulated in terms of the potential outcomes framework and clearly defined counterfactual outcomes, and for designing studies so that they have credible identification strategies that can be subjected to careful checks for robustness. Attention has also focused on the fact that there is likely to be heterogeneity in treatment effects (see Auld 2006). McClellan et al. (1994) and McClellan and Newhouse (1997) were quick to adopt the notion of local average treatment effects. Heterogeneity in treatment effects, captured through the concept of the marginal treatment effect, has been taken further in applied work with health data: Aakvik et al. (2005) use a structural model of a system of equations for outcomes and treatment and Basu et al. (2007) use the method of local instrumental variables. Econometric methods for policy evaluation are discussed in more detail in Jones (2009) and Jones and Rice (2011).

See Also

- ▶ Difference-in-Difference Estimators
- ▶ Econometrics
- ▶ Economic Epidemiology
- ▶ Health Economics
- ▶ Health Insurance, Economics of
- ▶ Health Outcomes (Economic Determinants)
- ▶ Instrumental Variables
- ▶ Natural Experiments and Quasi-Natural Experiments
- ▶ Non-linear Panel Data Models
- ▶ Population Health, Economic Implications of
- ▶ Selection Bias and Self-Selection
- ▶ Stratified and Cluster Sampling
- ▶ Survey Data, Analysis of
- ▶ Treatment Effect

Bibliography

Aakvik, A., J.J. Heckman, and E.J. Vytlacil. 2005. Estimating treatment effects for discrete outcomes when responses to treatment vary: An application to Norwegian vocational rehabilitation programs. *Journal of Econometrics* 125: 15–51.

Arcidiacono, P., H. Sieg, and F. Sloan. 2007. Living rationally under the volcano? An empirical analysis of heavy drinking and smoking. *International Economic Review* 48: 37–65.

Auld, M.C. 2006. Using observational data to identify the causal effects of health-related behaviour. In *The Elgar companion to health economics*, ed. A.M. Jones. Cheltenham: Edward Elgar.

Auster, R., I. Leveson, and D. Sarachek. 1969. The production of health an exploratory study. *Journal of Human Resources* 15: 411–436.

Bago d'Uva, T. 2006. Latent class models for utilisation of health care. *Health Economics* 15: 329–343.

Basu, A., and P.J. Rathouz. 2005. Estimating marginal and incremental effects on health outcomes using flexible link and variance function models. *Biostatistics* 6: 93–109.

Basu, A., J. Heckman, S. Navarro, and S. Urzua. 2007. Use of instrumental variables in the presence of heterogeneity and self-selection: An application to treatments of breast cancer patients. *Health Economics* 16: 1133–1157.

Black, S., P. Devereux, and K. Salvanes. 2007. From the cradle to the labour market? The effect of birth weight on adult outcomes. *The Quarterly Journal of Economics* 122: 409–439.

Blough, D.K., C.W. Madden, and M.C. Hornbrook. 1999. Modeling risk using generalized linear models. *Journal of Health Economics* 18: 153–171.

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

Cameron, A.C., P.K. Trivedi, F. Milne, and J. Piggott. 1988. A microeconometric model of demand for health care and health insurance in Australia. *Review of Economic Studies* 55: 85–106.

Contoyannis, P., A.M. Jones, and N. Rice. 2003. Simulation-based inference in dynamic panel probit models: An application to health. *Empirical Economics* 28: 1–29.

Contoyannis, P., A.M. Jones, and R. Leon-Gonzalez. 2004a. Using simulation-based inference with panel data in health economics. *Health Economics* 13: 101–122.

Contoyannis, P., A.M. Jones, and N. Rice. 2004b. The dynamics of health in the British household panel survey. *Journal of Applied Econometrics* 19: 473–503.

Conway, K.S., and P. Deb. 2005. Is prenatal care really ineffective? Or is the 'devil' in the distribution? *Journal of Health Economics* 24: 489–513.

Deb, P. 2001. A discrete random effects probit model with application to the demand for preventive care. *Health Economics* 10: 371–383.

Deb, P., and P.K. Trivedi. 1997. Demand for medical care by the elderly: A finite mixture approach. *Journal of Applied Econometrics* 12: 313–336.

Deb, P., and P.K. Trivedi. 2006. Specification and simulated likelihood estimation of a non-normal treatment-outcome model with selection: Application to health care utilization. *Econometrics Journal* 9: 307–331.

Deb, P., M.K. Munkin, and P.K. Trivedi. 2006. Bayesian analysis of the two-part model with endogeneity: Application to health care expenditure. *Journal of Applied Econometrics* 21: 1081–1099.

Dor, A., P. Gertler, and J. van der Gaag. 1987. Non-price rationing and the choice of medical care providers in rural Cote D'Ivoire. *Journal of Health Economics* 6: 291–304.

Dowd, B., R. Feldman, S. Cassou, and M. Finch. 1991. Health plan choice and the utilization of health care services. *Review of Economics and Statistics* 73: 85–93.

Duan, N. 1983. Smearing estimate: A nonparametric retransformation method. *Journal of the American Statistical Association* 78: 605–610.

Duan, N., W.G. Manning, C.N. Morris, and J.P. Newhouse. 1983. A comparison of alternative models for the demand for health care. *Journal of Business and Economic Statistics* 1: 115–126.

Feldstein, M.S. 1967. *Economic analysis for health service efficiency: Econometric studies of the British national health service*. Amsterdam: North-Holland.

Gilleskie, D.B., and T.A. Mroz. 2004. A flexible approach for estimating the effects of covariates on health expenditures. *Journal of Health Economics* 23: 391–418.

Grootendorst, P.V. 1997. Health care policy evaluation using longitudinal insurance claims data: An application of the panel tobit estimator. *Health Economics* 6: 365–382.

Grossman, M. 1972. *The demand for health: A theoretical and empirical investigation*. New York: Columbia University Press for the National Bureau of Economic Research.

Hamilton, B. 1999. HMO selection and medicare costs: Bayesian MCMC estimation of a robust panel data tobit model with survival. *Health Economics* 8: 403–414.

Hay, J., and R.J. Olsen. 1984. Let them eat cake: A note on comparing alternative models of the demand for health care. *Journal of Business and Economic Statistics* 2: 279–282.

Heckman, J.J. 2012. The developmental origins of health. *Health Economics* 21: 24–29.

Hoch, J.S., A.H. Briggs, and A.R. Willan. 2002. Something old, something new, something borrowed, something blue: A framework for the marriage of health econometrics and cost-effectiveness analysis. *Health Economics* 11: 415–430.

Jochmann, M., and R. Leon-Gonzalez. 2004. Estimating the demand for health care with panel data: A semiparametric Bayesian approach. *Health Economics* 13: 1003–1014.

Jones, A.M. 2000. Health econometrics. In *Handbook of health economics*, ed. A.J. Culyer and J.P. Newhouse. Amsterdam: Elsevier.

Jones, A.M. 2007. *Applied econometrics for health economists: A practical guide*. Oxford: Radcliffe Medical Publishing.

Jones, A.M. 2009. Panel data methods and applications to health economics. In *Palgrave handbook of econometrics, vol. II: Applied econometrics*, ed. T.C. Mills and K. Patterson. Basingstoke: Palgrave Macmillan.

Jones, A.M. 2011. Models for health care. In *Oxford handbook of economic forecasting*, ed. D. Hendry and M. Clements. Oxford: Oxford University Press.

Jones, A.M., and N. Rice. 2011. Econometric evaluation of health policies. In *Oxford handbook of health economics*, ed. S. Glied and P.C. Smith. Oxford: Oxford University Press.

Jones, A.M., N. Rice, T. Bago d'Uva, and S. Balia. 2007. *Applied health economics*. London: Routledge.

Keeler, E.B., W.G. Manning, and R.B. Wells. 1988. The demand for episodes of mental health services. *Journal of Health Economics* 7: 369–392.

Kerkhofs, M., and M. Lindeboom. 1995. Subjective health measures and state dependent reporting errors. *Health Economics* 4: 221–235.

Manning, W. 1998. The logged dependent variable, heteroscedasticity, and the retransformation problem. *Journal of Health Economics* 17: 283–295.

Manning, W. 2006. Dealing with skewed data on costs and expenditure. In *The Elgar companion to health economics*, ed. A.M. Jones. Cheltenham: Edward Elgar.

Manning, W.G., and J. Mullahy. 2001. Estimating log models: To transform or not to transform? *Journal of Health Economics* 20: 461–494.

Manning, W.G., N. Duan, and W.H. Rogers. 1987a. Monte Carlo evidence on the choice between sample selection and two-part models. *Journal of Econometrics* 35: 59–82.

Manning, W., J.P. Newhouse, N. Duan, E. Keeler, A. Leibowitz, and M.S. Marquis. 1987b. Health insurance and the demand for medical care: Evidence from a randomized experiment. *American Economic Review* 77: 251–277.

Manning, W.G., A. Basu, and J. Mullahy. 2005. Generalized modelling approaches to risk adjustment of skewed outcomes data. *Journal of Health Economics* 24: 465–488.

McClellan, M., and J.P. Newhouse. 1997. The marginal cost-effectiveness of medical technology: A panel instrumental variables approach. *Journal of Econometrics* 77: 39–64.

McClellan, M., J.P. Newhouse, and B. McNeil. 1994. Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? *Journal of the American Medical Association* 272: 859–866.

Miguel, E., and M. Kremer. 2004. Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica* 72: 159–217.

Mullahy, J. 1986. Specification and testing of some modified count data models. *Journal of Econometrics* 33: 341–365.

Mullahy, J. 1997a. Instrumental variable estimation of count data models. Applications to models of cigarette smoking behaviour. *Review of Economics and Statistics* 79: 586–593.

Mullahy, J. 1997b. Heterogeneity, excess zeros, and the structure of count data models. *Journal of Applied Econometrics* 12: 337–350.

Mullahy, J. 1998. Much ado about two: Reconsidering retransformation and the two-part model in health econometrics. *Journal of Health Economics* 17: 247–281.

Newhouse, J.P. 1987. Health economics and econometrics. *American Economic Review* 77: 269–274.

Pohlmeier, W., and V. Ulrich. 1995. An econometric model of the two-part decision making process in the demand for health care. *Journal of Human Resources* 30: 339–360.

Risa Hole, A. 2008. Modelling heterogeneity in patients' preferences for the attributes of a general practitioner appointment. *Journal of Health Economics* 27: 1078–1094.

Rosenzweig, M.R., and T.P. Schultz. 1983. Estimating a household production function: Heterogeneity, the demand for health inputs, and their effects on birth weight. *Journal of Political Economy* 91: 723–746.

Santos Silva, J.M.C., and F. Windmeijer. 2001. Two-part multiple spell models for health care demand. *Journal of Econometrics* 104: 67–89.

Smith, P.C., N. Rice, and R. Carr-Hill. 2001. Capitation funding in the public sector. *Journal of the Royal Statistical Society A* 164: 217–257.

Terza, J.V. 1998. Estimating count data models with endogenous switching: Sample selection and endogenous treatment effects. *Journal of Econometrics* 84: 93–127.

Van de Ven, W., and R.P. Ellis. 2000. Risk adjustment in competitive health plan markets. In *Handbook of health economics*, ed. A.J. Culyer and J.P. Newhouse. Amsterdam: Elsevier.

van de Ven, W.P.M.M., and J. van der Gaag. 1982. Health as an unobservable: A MIMIC model for health care demand. *Journal of Health Economics* 1: 157–183.

van Doorslaer, E.K.A. 1987. *Health, knowledge and the demand for medical care*. Assen/Maasricht: Van Gorcum.

Van Ourti, T. 2004. Measuring horizontal inequity in belgian health care using a gaussian random effects two part count data model. *Health Economics* 13: 705–724.

Wagstaff, A. 1989. Econometric studies in health economics. *Journal of Health Economics* 8: 1–51.

Windmeijer, F.A.G., and J.M.C. Santos Silva. 1997. Endogeneity in count data models; An application to demand for health care. *Journal of Applied Econometrics* 12: 281–294.

Wolfe, B., and J. van der Gaag. 1981. A new health status index for children. In *Health, economics, and health economics*, ed. J. van der Gaag and M. Perlman. Amsterdam: North-Holland.

Zimmer, D.M., and P.K. Trivedi. 2006. Using trivariate copulas to model sample selection and treatment effects: Application to family health care demand. *Journal of Business and Economic Statistics* 24: 63–76.

Health Economics

Barbara Wolfe

Abstract

Health care expenditures form an ever-increasing burden in most developed countries, especially the United States, where they accounted for 16.0 per cent of GDP in 2004, up from 5.1 per cent of GDP in 1960. These cost increases alone suggest that health economics is a dynamic field of economic research, but the importance and the interest of the field are driven by broader considerations. This article delineates important market failures and research issues in health and health care, the relationship between income and health, methodological issues in the measurement of health, and quality issues in the measurement of health care.

Keywords

Absolute deprivation–health hypothesis; Income–health hypotheses; Adverse selection; Asymmetric information; Crowding out; Deadweight loss; Diffusion of technology; Disability, economics of; Gradient effects; Health capital; Health economics; Health insurance; Health maintenance organizations (HMOs); Hospital economics; Managed competition; Market failures in health care; Medicaid (USA); Medical practice licensing; Medicare (USA); Moral hazard; Mortality; Non-profit organizations; Nutrition; Payroll tax; Pharmaceutical industry; RAND Health Insurance Experiment; Tax expenditures; Well-being

JEL Classifications

I1

Health care expenditures form an ever-increasing burden in most developed countries. Between 1960 and 2002, expenditures as a percentage of GDP rose from 3.8 to 9.7 per cent in France, from 5.4 to 9.6 per cent in Canada, and from 4.9 to 11.2 per cent in Switzerland. (Cross-country comparisons are hindered to some extent by the differential inclusion of components of health care and social services in different countries.) In the United States, health care expenditures are far greater; they accounted for 16.0 per cent of GDP in 2004 (a per capita expenditure of $6,280), up from 5.1 per cent of GDP in 1960. By themselves these increases, projected to rise to 20 per cent or more of GDP by 2015, suggest that health economics is a dynamic field of economic research, but the importance of the field is driven by far more than the costs of care.

In many ways, health economics mimics the broader field of classical economics in its areas of research specialization – there are theoretical studies, micro and macro studies, industrial organization studies, public economic studies, andlabour studies, among others. But health economics has a unique quality, identified in one of the earliest papers in the field, by Kenneth Arrow (1963). That is the large role played by *market failures,* which make it likely that resources will be allocated inefficiently if market outcomes alone prevail.

Market Failures

Market failures take several forms.

Failures Related to Information

Individuals tend to be poor judges of the care they need and the quality of care they obtain. This ignorance works both *ex ante* and *ex post.* Consumers do not know whether they might benefit from medical care. They tend to lack information about the appropriate type, amount, quality, and price of care. They also lack information about the counterfactual: would alternative care, or even no care at all, be equally or more or less beneficial and cost-effective? The rapid introduction of new technology and the need to make decisions under stress tend to result in ineffective information searches.

Failures Related to the Role of Supplier Agents

Providers acting for the patient are rarely perfect agents: they do not fully understand individual patient preferences, their own earnings are influenced by their advice (conflict of interest), and their training tends to impel them to do all that is technologically possible (that is, provide care until expected marginal benefits equal zero). The usual remedy for this kind of failure is public sector interventions in the health care market. These include licensing of providers to assure a minimum level of competence; licensing of facilities and new technologies (including pharmaceuticals) to assure quality; reimbursement schemes to minimize conflicts of interest; subsidies for certain types of care (for example, those with external benefits, such as vaccinations); and subsidies for the purchase of insurance.

Failures Related to Uncertainty

The combination of uncertain need for care and the high expense associated with a major health problem leads naturally to a demand for pooling risk and insurance. But, as in most types of insurance, willingness to pay for coverage based on individual assessments of one's own (differing) risks may lead to adverse selection (see below) and incomplete coverage. For those insured, the reduction in the price of care may lead to increased demand for care, less attention to the price of care, and less attention to avoiding the need for care ('moral hazard'), making it very difficult to estimate optimal amounts of care. Insured individuals may demand care beyond the point where the marginal expected benefit is equal to the true cost of the resource, or substitute care for other health-preserving options such as exercise or diet. When insurance is combined with reimbursement schemes that pay providers for each service ('fee for service'), the likely result is demand and provision of care in excess of expected benefits. (The idea of providers recommending care to the patient, knowing that the expected benefit to the patient is less than the cost, has been termed 'physician-induced demand'. Though numerous articles have been written on the topic, proof of such behaviour remains elusive.) But reimbursement schemes that pay providers a fixed or capitated amount for all services may lead to too little care, based on a comparison of marginal cost to marginal benefits. Much recent work has focused on designing reimbursement schemes that mitigate overuse or underuse.

The Demand for Health

Michael Grossman (1972) was the first to emphasize that the outcome of interest in health economics is health, not medical care. The demand for medical care, in other words, is derived from the demand for good health. Grossman's model and all the work that derives from it are essentially extensions of the household production literature. In brief, Grossman argues that health capital is a form of human capital which changes over time because of depreciation and investment. His model begins with a utility function where health is an argument in addition to utility gained from the consumption of goods and services. Health investments include time (exercise and sleep) and medical care, subject to health endowments such as genetic traits or environmental factors that are known to the individual or family. Without investment in health, health deteriorates. Net investment in health equals gross investment minus depreciation, the rate of which is assumed to be exogenous and to increase with age, so that it becomes more and more expensive to obtain good health. There is also an education efficiency parameter. Higher health stock increases healthy time, leading to higher income (increased productivity and time to work), making income endogenous in the model. All individual choices are subject to both a time and money budget constraint. The time constraint requires that the total amount of time available in any period must be exhausted by all possible uses, which include time spent working, producing health, lost to illness, and spent in leisure activities. The income constraint explicitly includes expenditures on medical care.

Empirical estimates have used both reduced form and structural versions of this model to answer questions such as how much a tax on cigarettes will reduce low birthweight (reduced form), or how much a change in maternal smoking might influence birthweight (structural estimates of the marginal product). The models underlie some studies of unhealthy behaviours (such as alcohol abuse) and have extensive applications in environmental economics. Although there are always issues of endogeneity (for example, income, medical care) and hence of identification in empirical applications of this model, the model's improved focus on the nature of the demand for medical care and the important role of time allocation have had a major impact on the field. (Gerdtham et al. (1999), conducted one of the best empirical tests of the Grossman model using Swedish individual data.)

Income and Health: Estimating the Relationship

Consistent with the empirical evidence, the Grossman model suggests that income is positively associated with health, but the direction of causation is not clear. Those with better health earn more and hence have higher incomes, whereas those with higher incomes can invest more in better health, suggesting that the observed income–health gradient may best be modelled as a simultaneous system.

The idea that income is associated with health goes back a long way in the health economics literature, and a number of hypotheses have been advanced to explain the relationship. Samuel Preston (1975) observed that the impact of additional income on health (as measured by mortality) is greater for those with lower income than for those with higher income. This observation of diminishing marginal productivity is called the 'absolute income' hypothesis. In its simplest form, it argues that, if income is all that matters to individual health, a community with more equal income will tend to have better average health than a community with more inequality, even if the two communities have the same average income. In an international context, Angus Deaton (2002) points out that, according to the absolute income hypothesis, redistribution can improve health even if average income is not increased, and that redistribution from rich to poor countries would in principle improve worldwide average health.

A related concept is the 'absolute deprivation' or poverty hypothesis, which suggests that those with the lowest incomes face poorer health and a greater risk of mortality owing to inadequate nutrition, poor-quality health care, exposure to physical hazards, and heightened stress. According to this hypothesis, a dollar redistributed from rich to poor would improve the health of the poor and improve the average health of the entire population.

The 'relative income' hypothesis focuses on an individual's income relative to others in his or her group. If the incomes of all members but one in a group increase, that one person's health is expected to deteriorate. A related, 'relative position' hypothesis holds that one's rank (occupation or education) in society is tied to health outcomes. Research in the United States and the United Kingdom has demonstrated an association between socio-economic position and health (Mullahy et al. 2004, review the evidence for this). Referred to as a 'gradient effect', this hypothesis implies that psychosocial and other factors that remain unevenly distributed all the way up the income scale perpetuate income inequalities in health. Perceptions of being relatively deprived ('keeping up with the Joneses'), stress, and other non-material factors may play a role in perpetuating income inequalities in health at the upper income levels. The distinction between absolute and relative income effects has important policy implications; if the income of everyone were to increase or decrease, no change in health would be expected under a relative income model, but change would be expected under an absolute income model.

The hypothesis that focuses most directly on the tie between inequality in both health and income has two versions, one 'strong' and one 'weak'. According to the strong version of the income inequality hypothesis, if the average income of the society is held constant, societies with greater inequality produce worse health among their citizens. Those in the most unequal communities may fear for their lives and property whether they are poor or wealthy, or the stress of keeping up with the Joneses may reduce time allocated to producing health. The weaker version argues that those with incomes below the mean will be negatively influenced by greater income inequality, perhaps through higher residential density and the associated increases in crime and contagious disease. A related issue for research is the extent to which these observed ties between income and health can be ascribed to race or ethnicity, through the systematically differing average incomes of racial and ethnic groups in a country like the United States. But if groups also differ in diet and genetics, health may be causally linked to these other factors rather than to the observed income gradient.

In general, there have been two empirical approaches to examining the hypotheses described above. Research into the absolute deprivation, relative income, and relative position hypotheses has usually relied upon individual-level data on income and health or mortality to examine the existence and shape of the income–health relationship among individuals; research on race and ethnicity follows a similar approach. In contrast, research examining the income inequality hypothesis has employed aggregate data exclusively, at least in measuring income inequality.

Insurance

As noted above, uncertainty in the need for medical care and the potentially very large costs of care lead to a demand for health insurance. (Nyman (1999), has suggested an additional motivation for demanding health insurance: having insurance permits consumers to consume very expensive care that would otherwise be beyond their budget constraints.) But costs of operation, inability of the insurer to accurately discern risks (information asymmetry), and the potential for increased expenditures on the insured mean that insurance may not be supplied at a price reflecting an individual's actuarially fair cost. A traditional market failure may occur, in which only those with high expected medical expenditures are willing to buy coverage (adverse selection). Or there may be complete failure of the insurance market because no (risk-neutral) supplier is willing to offer coverage at a price that any individual would willingly pay. This has led to publicly provided coverage in many countries and subsidies towards the purchase of insurance in others, such as the United States. It has also led to incomplete insurance, in which deductibles, copayments, and co-insurance attempt to reduce moral hazard. (A deductible requires that some initial level of expenditures is covered directly by the consumer; a copayment is usually a fixed dollar payment per specified unit of service; and coinsurance is a fixed percentage payment. In general, although co-payments are quite common – for example, in pharmaceutical coverage –they have poorer incentive effects than coinsurance.)

The insurance market raises several interesting issues.

1. *The role of secondary insurance.* Secondary insurance may reduce the cost of public coverage if privately financed care is sufficiently substituted for publicly financed care, or may raise the cost of public coverage if it primarily pays for the cost-sharing components of publicly financed care. In the United States, for example, 'Medigap' insurance covers deductibles and co-payments required under Medicare, the system covering those aged 65 or more and the significantly disabled; it thereby increases demand for those services primarily paid for by Medicare.
2. *The efficiency and equity of subsidizing the purchase of insurance through the tax system.* The US system provides the highest subsidies to those with the highest marginal tax rates (those with high incomes) and offers little or no subsidy to those with low incomes.
3. *The incentive effects of income-conditioned eligibility for public insurance.* To be eligible for the Medicaid programme in the United States, persons have to meet state-specified eligibility requirements linked to income, assets and family structure. There is an all-or-nothing eligibility requirement, such that a person with income or assets a dollar above the cut-off is ineligible for Medicaid. Three potential consequences are less work effort (reduced earnings) by individuals who wish to become or remain eligible, increased numbers without private insurance among the near-eligible who are in effect 'insured' for costly care since they could become eligible if they needed such care,

and reduced savings among those eligible and near-eligible.

4. *The welfare loss from public subsidies for private insurance or public provision of medical care.* This is potentially higher the more services that are covered, for example, covering all new technology rather than only that which passes some benefit-cost analysis, covering all pharmaceuticals compared with only the least costly drug within a category, or covering all types of counselling rather than that tied to a diagnosis of severe mental illness.
5. *Issues tied to optimal breadth of coverage.* Such issues include, for example, whether universally provided or mandated insurance increases welfare and the optimal depth of that coverage, life-threatening emergency care but not cosmetic surgery, semi-private hospital rooms but not private rooms, treatment for cancer but not dementia.
6. *The labour market implications of using a payroll tax to subsidize or directly provide health insurance coverage*. This policy potentially increases the costs to employers of hiring additional workers, especially older workers or those who have a chronically ill family member. Most modern economies struggle with how best to design employer-based taxation to fund health care insurance.
7. *How to minimize crowd-out in countries with both public and private health insurance systems.* In the United States the issue is designing public insurance coverage to minimize incentives to turn down private coverage (see issue 3 above). In the Netherlands the issue is balancing payroll taxes against private insurance premiums so that younger and healthier earners will not find ways to join the private system in lieu of the public system.
8. *Designing policies to increase take-up of insurance among eligible populations within the constraints of equity and efficiency, thus avoiding unnecessary subsidies for those already enrolled.*
9. *How to effectively cover the treatment of mental illnesses.* A particular variant of this in the United States is the role and design of mental health parity laws.

The Demand for Medical Care

Much research has focused on empirically estimating the elasticity of demand for medical care. This question gets renewed attention whenever there are proposed changes to insurance coverage, since the cost of such policy changes lies in the elasticity of demand. (An additional or second-order cost of any expansion of insurance via public policy is the magnitude of the social welfare loss –deadweight loss – associated with the moral hazard effect. The most-cited original contribution on this is Mark Pauly 1968.) The simplest empirically estimated models (most of which use number of visits or units of care as the dependent variable) include the price of care, income, simple demographic factors, and the price of alternative goods and services. But accurately measuring the marginal cost of care is not a simple proposition for individuals with insurance coverage that includes deductibles, copayments or co-insurance, maximums per episode, and otherwise incomplete coverage. More satisfactory models include the value of time (including time spent in care, time spent in transit, and time spent waiting). Most empirical research differentiates the demand for hospital stays from physician services. A number of studies narrow the question still further, to the level of the market for hospital or physician services or for individual physicians. As expected, demand elasticities for individual physicians are quite high in large markets (suggesting a competitive market), whereas elasticities for physicians as a whole or for hospitals tend to be far lower, suggesting their considerable market power.

All non-experimental estimates of the demand for care using US populations suffer from the endogeneity of the marginal price of care because the demand for insurance is endogenous to the demand for medical care; that is, individuals or families with the highest expected medical expenses will seek out relatively generous coverage. In a large-scale experiment, the RAND Health Insurance Experiment (see Newhouse 1994), individuals were randomly assigned to various plans ranging from full coverage (free care) to care with a 95 per cent coinsurance and a maximum dollar contribution. The design of the experiment was such that researchers could more accurately assess elasticities of demand and the marginal price of care. Participants were observed for 3–5 years. The study had shortcomings (some by design): it excluded those with the highest demand (the elderly and disabled), experienced attrition especially among those with the least generous plans, and made the decision to pay participants a lump sum to be sure all families were made no worse off by participation in the experiment. The experiment convincingly established that individuals do respond to price, even for hospital care. (The experiment found that 86.7 per cent of those with full coverage used care in a given year compared with 68 per cent of those with 95 per cent coverage: medical expenditures for the year, in 1984 dollars, were $777, or 32.8 per cent, versus $534, or 27.4 per cent, respectively.)

Substantial econometric issues in estimating demand elasticities remain and have spawned much methodological research in health economics. Four main issues are: (*a*) the highly skewed nature of utilization and expenditure distributions in populations of interest (a multipart – usually two-or four-part – model in which first the probability of any use, or particular use such as outpatient care, and then the level of use conditional on any use has been frequently used in the literature. These models may not be readily identifiable, however; Duan et al. 1983.); (*b*) the episodic nature of care; (*c*) whether to use quantity of care as the outcome of interest (and, if so, how to include dimensions of the duration, extent, and quality of care) or to use expenditures (and, if so, how to measure actual expenditures rather than billed amounts); and (*d*) how to capture the nature of demand, much of which is based on ill health, which is difficult to measure accurately.

Measuring Health Outcomes

Given the prominence of market failure in health economics, accurate measures of health to capture the effectiveness of medical care are crucial. (Defining health is a major problem that lies largely outside the domain of the health economist. Perhaps the most often used concept is that promoted by the World Health Organization – a complete state of well-being. This is not very useful for the measurement of health.) Mortality is the 'health' measure most commonly discussed – and arguably most precisely measured – in this literature, but the relationship of mortality to care may be quite distant in time as well as 'coarse' or 'noisy'. Other measures of health, which may have more proximate temporal relationships to medical care, are necessary. Such non-mortality measures of health ('biological well-being') may be of many types: cellular or molecular (such as measles antibodies or titres); clinical (for example, body mass index); functional (for example, indices such as Activities of Daily Living scores); self-rated (for example, on a scale from excellent to poor); medical providers' diagnosis of particular physical or mental health diseases, or tied to activities (such as days of school missed). Many measures are straightforward while others attempt to capture 'utility' by asking individuals to compare a state of illness to days or years of total healthiness to create indices such as quality adjusted life years (QALYs). The measures selected must be appropriate for the purpose. For instance, some measures of health may be largely determined by genetic factors (such as risk for schizophrenia), others closely related to opportunity costs (for example, days of school missed because of parental work schedules). A measure appropriate for the analysis of labour force participation (for example, poor or fair health versus good or better health) may not be suitable for evaluating a targeted intervention. And, as with other empirical work, measures should detect changes (vary), measure what they intend to measure (be valid), and be free from error (be reliable).

Measurement Issues in Evaluating Policy

Although the RAND Experiment established that consumers are responsive to the price of care, their results did not establish the value or effectiveness of care in influencing health. Nor did they determine how we might socially evaluate the benefits from a change in policy (such as an increase in the proportion of people eligible for public coverage) or the safety and efficacy trade-offs involved in designing regulation of new technologies. A large ‘industry’ exists to try to determine whether individual well-being improves, deteriorates, or remains largely unchanged after some change in policy or intervention. Critical issues include how to measure and value health changes and how to account for individual differences.

These issues tie health economics to core questions in public finance centring on the evaluation of policy changes using cost–benefit analysis, cost-effectiveness analysis, and multi-attribute utility analysis. Health economics has a further question, however. Is only the social perspective relevant, or is the perspective of a more narrowly defined group (payers, patients, providers) relevant also? And which population or individuals should be included in the analysis – only those currently alive, the next generation, or those who might survive only because of the intervention? The perspective adopted may determine whether to provide an intervention or invest in a new technology. Measuring costs is also very complex; calculation is rendered difficult by skewed cost distributions, interdependent costs, and economies. In many countries the pricing of new technologies and drugs is an additional and growing concern in policy design. (A related issue is the rate of change of technology, including pharmaceuticals. Decisions made on adoption and pricing modify the incentives for private investment in developing new drugs and equipment. Challenging issues include: Who should take the risk, particularly at early stages of development? How to encourage investment in new technology for diseases that have limited markets? How to share the benefits of new technology with those with limited incomes (including those in poor countries)?

Quality Issues in Evaluating Care

Research concerning health care providers ranges from the role of licensing and malpractice to the use of quality indexes both broad (for example, hospital report cards) and narrow (for example, risk-adjusted mortality indices for interventions such as coronary artery bypass grafts or high-risk deliveries). The concept of ‘pay for performance’ is gaining credibility among both private and public payers, but the implications for quality and distribution of care and for overall expenditures need further research. Preliminary studies suggest that both practices – quality indicators and pay for performance –pose a real danger of cream skimming; the ability to obtain high-quality care may diminish for those with the most to gain from such care. The medical malpractice system aims to signal the appropriate amount and quality of care, but to work properly it requires that all who suffer significant injury through medical care bring a legal action, that rewards or settlements reflect true utility-based losses, and that rewards are not paid where the cost of prevention exceeds the full loss. Economists continue to debate whether the system leads to defensive medicine (too much care) and to evaluate the implications of tort reform for quality of care.

Supply Side Issues

Providers

Do we have enough medical providers? Restrictions on entry into the profession make this an interesting question. In order to practise medicine one must be licensed, and in order to be licensed one must have a medical degree from an accredited institution. The supply of medical schools and student enrolment are regulated in nearly all countries. As Eli Ginzberg (1989, p. 88) noted, ‘Neither the restrictive policies of the first four decades of [the 20th] century, nor the expansionary policies of the postwar era were formulated and implemented on the basis of demand and supply of physician services’. Rate of return calculations indicate, in general, that rates are high for specialists and lower for primary care doctors, but, because entry into specialties is limited by available residencies, such analysis cannot fully answer the question of supply. Current research on the adequacy of supply extends beyond numbers per capita to include primary versus specialty care, geographic distribution, design of public subsidy, and repayment schemes. The substitutability of other health professionals including nurses, nurse practitioners, and social workers for medical doctors (or psychologists for psychiatrists) is tied to two main issues: whether we have sufficient medical providers and whether competition and lower reimbursement may increase the efficiency of some types of medical care, many of them routine and predictable. A sufficient supply of nurses is from time to time a particularly acute issue. Issues of interest in the market for nurses include the monopsony power of hospitals, working conditions including work hours and locations, and childcare during working hours, which tend to fall outside the standard 8 a.m. to 5 p.m. workday.

Hospitals

Much earlier work in health economics focused on hospitals, memorably described by Baumol and Bowen (1966, p. 497, referring to non-profit organizations in general) as 'bottomless receptacles into which limitless funds can be poured'. Various alternative reimbursement schemes have been designed to limit these expenditures, but all have difficulties. Hospitals paid on the basis of *fee-for-service* have an incentive to provide care wherever marginal benefits are positive (especially if patients are fully insured.) From an efficiency perspective, this leads to too much care, because costs are not considered. Hospitals paid according to a *fee schedule* have an incentive to over-provide services for which the fee is greater than or equal to the marginal cost but to skimp on other services. Hospitals paid a *per diem* have an incentive to extend patient stays, especially since marginal costs tend to be far lower later in a stay. Hospitals paid by *diagnosis* (such as in Medicare's diagnosis- related groups or DRGs) have the incentive to serve the healthiest of those who seek care (cream skimming), avoiding those for whom expected costs are greater than expected payments (dumping). Finally, hospitals paid on the basis of *capitation,* or that are part of a fully owned health maintenance organization (HMO) with a capitation-based income, face incentives to provide cost-effective care, but perhaps not all care with positive net benefits.

Understanding hospital behaviour is important for designing policies that influence their behaviour. Many, but far from all, hospitals are non-profit, and use their non-profit status to convey the message to patients that quality is not compromised by the desire for profits. They thus aim to generate trust that reduces the need for complicated contracts between the hospital and the consumer. (Hospitals are increasingly adding for-profit components to their array of services, thus masking the difference between for-profit and non-profit institutions.) The non-profit nature of most hospitals has led to a variety of models of hospital behaviour. One model views hospitals as two organizations in one. There is first the hospital staff, which provides resources to the physicians for the care of their patients. The physician staff want sufficient resources to treat their patients without delays and prefer some excess capacity; they want the latest technology, however expensive. Hospitals provide such technology in order to compete for physicians and their patients. The result is duplication and rapid diffusion of the newest technology. A second model is a utility maximization model of hospitals, in which hospital managers get utility from the increased quantity (size or number of beds) and quality of care provided. They can expand in both dimensions more easily within non-profits and with fewer binding constraints than within for-profit hospitals. Again the result is duplication. (A related version of this is a quantity maximization model.)

Two other models of hospitals suggest (*a*) that physicians control hospitals, behaving in ways that maximize their own incomes, or *(b)* that hospitals can be thought of as physician cooperatives that act to maximize their own well-being but are inefficient if the hospital grows too large. Some research on forms of ownership and hospital behaviour suggests that competition matters far more than form of ownership where price is concerned.

Managed Competition

First put forward by Alain Enthoven (1978), this approach changed the incentives facing providers to improve efficiency and quality. The plan called for multi-specialty group practices that would provide a specified, comprehensive set of medical care services in exchange for a per capita prospective payment covering a defined period of time. Individuals could choose an HMO plan (usually a closed panel or a limited set of providers) or traditional fee-for-service; all bidders would be required to offer a plan that covered the specified set of services. Employers would offer a broad set of plans but would pay only a fixed dollar amount towards the premium; consumers would pay the full difference between that contribution and the actual premium. The lower cost and more comprehensive benefits of the prepaid plans would lead consumers to choose those plans. Information on the quality of care under these plans would be systematically collected and shared with consumers, who would have an annual open enrolment period.

Empirical evidence on the effectiveness of managed competition suggests that it generates one-time savings but that the forces driving toward new technology, the desire among consumers for point-of-service choice, and adverse selection have limited its role. Reform, analysis and experimentation with variants of managed competition continue. The issue of how to reduce the rate of increase in the cost of medical care continues to be important in all developed countries. Managed competition is but one approach. Others include: limiting the number of providers who can practise in a jurisdiction (Canada and the United Kingdom, for example); increasing the co-payments required of consumers; modifying reimbursement of providers; using waiting lists to reduce access; providing free telephone advice to improve efficiency of demand (Australia and parts of the United Kingdom); regulating insurance coverage; setting a budget for a fixed period of time; and rationing care on the basis of age. Designing these approaches and evaluating their success is a continuing challenge.

The Economics of Disability

In this area, health economists have been concerned with the proper design of public policy, in particular the efficiency of disability-based benefits, including their work and health insurance incentives. Measurement (in this case, of disability) is again a major impediment to the quality of the research. Health economists are attempting to better understand the determinants of chronic health problems, including those such as obesity, asthma, and diabetes, which are on the increase. Models from other fields of economics (such as intergenerational mobility, time use, and consumer demand models) are being applied to understand the determinants of the increase and to identify interventions to stem the trend.

This short overview suggests that problems for exploration and opportunities to influence the design of policies from prevention, through insurance design, to reimbursement and regulation are likely to expand as the costs of health care continue to rise. Health economics approaches range from theory through empirical analysis to policy reform. They can include the development of new models or improved policy analysis and design. They can be country-specific or comparative; sector-specific (hospital care) or more comprehensive; and tied to labour, public, micro theory, or international studies.

See Also

- ▶ Adverse Selection
- ▶ Crowding Out
- ▶ Health Insurance, Economics of
- ▶ Market Failure
- ▶ Tax Expenditures
- ▶ Technology
- ▶ Uncertainty

Acknowledgments Thanks to Rashi Fein, Robert Haveman, John Mullahy, Nathan Tefft and Jason Fletcher for helpful suggestions.

Bibliography

Arrow, K. 1963. Uncertainty and the welfare economics of medical care. *American Economic Review* 53: 941–973.

Baumol, W., and W. Bowen. 1966. *Performing arts – The economic dilemma*. New York: Twentieth Century Fund.

Deaton, A. 2002. Policy implications of the gradient of health and wealth. *Health Affairs* 21 (2): 13–30.

Duan, N., W. Manning, C. Morris, and J. Newhouse. 1983. A comparison of alternative models for the demand for medical care. *Journal of Business and Economic Statistics* 1: 115–126.

Enthoven, A. 1978. A consumer choice health plan: a national health insurance proposal based on regulated competition in the private sector. *New England Journal of Medicine* 298: 650–658. and 709–20.

Gerdtham, U.-G., M. Johannesson, L. Lundberg, and D. Isacson. 1999. The demand for health: Results from new measures of health capital. *European Journal of Political Economy* 15: 501–521.

Ginzberg, E. 1989. Physician supply in the year 2000. *Health Affairs* 8 (2): 84–90.

Grossman, M. 1972. On the concept of health capital and the demand for health. *Journal of Political Economy* 80: 223–255.

Mullahy, J., S. Robert, and B. Wolfe. 2004. Health, income and inequality. In *Social inequality*, ed. K. Neckerman. New York: Russell Sage Foundation.

Newhouse, J., and Health Insurance Experiment Group. 1994. *Free for all? Lessons from the RAND Health Insurance Experiment*. Cambridge, MA: Harvard University Press.

Nyman, J. 1999. The value of health insurance: The access motive. *Journal of Health Economics* 18: 141–152.

Pauly, M. 1968. The economics of moral hazard: Comment. *American Economic Review* 58: 531–537.

Preston, S. 1975. The changing relation between mortality and level of economic development. *Population Studies* 29: 231–248.

Health Insurance, Economics of

Joseph P. Newhouse

Abstract

Health care finance has been dominated by moral hazard, potential rents and the deadweight loss from financing them, and adverse selection. Public health services and insurance tend to be universal, solving the selection problem. Private health insurance markets and public schemes that offer a choice of insurance plans generally exhibit selection. Research has found strong evidence of responsiveness of demand to insurance coverage. In health insurance markets information is asymmetric among patients, providers, and insurers, and principal–agent relationships abound. Actual health insurance and health care financing institutions have adapted to these features.

Keywords

Adverse selection; Akerlof, G.; Arrow, K.; Asymmetric information; Deadweight loss; Employment-based medical insurance; Group insurance; Health insurance: private vs. public; Insurance contracts; Managed care insurance plans; Medical care, state provision of; Medicare; Moral hazard; Principal-agent relationship; Rent; Risk aversion

JEL Classifications

I11; G22; I1; H51; D12; D82; L15; L84

As the capabilities and the associated expense of medicine advanced during the 20th century, the demand for financial protection against the risk of large medical spending grew. The result of the increased demand has been widespread health insurance or direct public provision of medical care, or both, in every developed country.

Both health insurance and the public provision of medical care heavily subsidize that care at the point of service, meaning that the user bears only a fraction (usually a small one) of the cost. As a result, insurance induces moral hazard and potentially rents in factor prices as well, which in turn induces deadweight loss through the taxes needed to finance any public insurance. Both private and public insurers, however, may combat moral hazard and rents through the nature of their contracts with providers, and sometimes through command-and-control type intervention.

In addition to moral hazard and potential rents, a voluntary health insurance market exhibits adverse selection. To combat selection, voluntary insurance contracts may contain provisions for medical underwriting (exclusion of an individual from a small group insurance plan), exclusions of pre-existing conditions from coverage, or exclusion of certain services from coverage altogether. Such features may prevent willing buyers and sellers from contracting, as well as hamper the efficiency of labour markets with employment-based insurance, as employees may not leave jobs because of the inability to obtain comparable insurance (Gruber 2000).

These three features – moral hazard, potential rents and the deadweight loss from financing them, and adverse selection – have influenced countries' health care financing institutions, as first suggested in the seminal paper by Kenneth Arrow (1963). To combat selection, some countries, for example the United Kingdom and southern Europe, deliver health services through a public health service. In this case health insurance markets are supplemental to the public health service. Other countries, such as Canada and many northern European countries, combat selection by offering public or quasi-public insurance with no choice of insurance plan; private health insurance markets are again supplemental. Because public health services and public health insurance tend to be universal (though in some countries the affluent can opt out), the selection problem is solved, potentially at the expense of a poorly performing monopoly.

By contrast, private health insurance markets and public schemes that offer a choice of insurance plans generally do exhibit selection. Among developed countries the United States relies most on private health insurance markets, although because of selection American private health insurance is primarily organized through employment rather than through an individual insurance market. Again in part because of selection, many Americans without an employment connection, most notably the elderly, are insured through public insurance. But not all American employers offer insurance, not all those employees offered insurance purchase it, and not all those without a labour market connection are eligible for public insurance. As a result, the United States has a much higher proportion of its population with no health insurance than other developed countries. This group tends to receive care through subsidized direct delivery systems.

In this article I first discuss selection and the demand for insurance. Then I discuss moral hazard and the demand for medical care conditional on insurance. Finally, I discuss the nature of the contract between the insurer and the provider, such as the physician or hospital, and its relation to the provider's supply of services. The material covered in this entry is treated more extensively in Cutler and Zeckhauser (2000).

Selection

Health insurance was used as an example of selection by two of the classic papers on asymmetric information (Akerlof 1970; Rothschild and Stiglitz 1976). The models in those papers showed that an equilibrium may not exist in competitive insurance markets if insurers could not identify high risks (for example hypochondriacs and possibly the chronically ill). If a competitive market pooled high and low risks (that is both risk types buying the same policy), insurers would offer products that differentially appealed to low risks, thus breaking the pooling equilibrium. Under certain conditions a separating equilibrium (that is each risk type buying a different policy) might also be impossible.

Later papers showed that under different assumptions an equilibrium might exist (Dubey and Geanakoplos 2002; Newhouse 1996; Wilson 1977), but selection behaviour seems pervasive in individual health insurance markets, and even a separating equilibrium is a form of market failure given the almost universal nature of annual contracts in private insurance markets. That is because a low risk this year has the risk of becoming a high risk next year, for example by contracting a chronic disease. But the higher premium facing a high risk is uninsurable with annual premiums. Notice that in a family insurance context this risk includes having a high-risk child born into the family, and, when the child becomes an adult, extends also to the child (typically the child can no longer be covered under a parent's policy after a certain age).

Cochrane (1995) pointed out that lifetime contracts would solve this problem, but lifetime contracts are not observed in the private market, for several reasons. First, the rapid rate of technological change in medicine and the associated cost increase (Cutler 2004; Newhouse 1992) is a

Health Insurance, Economics of, Table 1 Cost and actuarial value of insurance policies

	Individual policy ($)		Family policy ($)	
Percentile	Premium	Actuarial value	Premium	Actuarial value
10	1,220	1,740	2,760	4,220
25	1,670	1,910	3,950	4,600
50	2,100	2,100	5,070	5,070
75	2,620	2,260	6,090	5,450
90	3,220	2,440	7,670	5,890
Difference 90–10	164%	40%	178%	40%

Source: Cutler (1994, Table 2)

non-diversifiable risk for a given cohort. Second, there are large economies to group insurance, in part because of lower marketing costs. To avert selection, however, groups must be formed primarily for reasons other than obtaining health insurance, which is why health insurance in many countries forms around the place of employment. But even here selection can be a problem, most obviously for the self-employed, who are effectively in an individual market, but also for small employers.

Empirical work has confirmed the importance of selection. I show three examples. The first is the variation in insurance premiums by generosity of the insurance. Table 1 orders insurance policies by generosity, where a higher percentile indicates more complete coverage. The column labelled 'Premium' is the premium charged for the policy; the column labelled 'Actuarial value' is the estimated spending among a standardized population for each policy, reflecting the increase in demand for services when the insurer covers a greater portion of any medical spending. (One might ask how actuarial values are known. They can come from similar policies from groups where selection is minimal, such as employees of large firms with no choice of insurance plan, or from other evidence on how the demand for care varies with insurance generosity; one form of such evidence is discussed below in conjunction with moral hazard.)

Whereas premiums between the 90th and 10th percentile plans differ by a factor of about 2.7, the actuarial value of the two plans differs only by a factor of 1.4. The difference between these values indicates that high risks are disproportionately choosing the 90th percentile plan, and low risks are disproportionately choosing the 10th percentile plan.

A second form of evidence comes from the US Medicare programme, the near- universal insurance programme for individuals over the age of 65. Medicare gives its beneficiaries choice between a traditional indemnity insurance plan that allows free choice of physician, and a prepaid plan, which can restrict choice of physician but in return charges a smaller premium or covers certain additional services. Through 2005 individuals have been allowed to change between these two types of plans monthly (under current law this is to change in 2006).

The data suggest the traditional indemnity plan has been more appealing to high-risk individuals. Although spending and use data have not been available for those in prepaid plans, one can compare use among those in the traditional plan who subsequently enrol in a prepaid plan with that of those who do not. Adjusted for age and sex and a few other covariates, spending among those switching from the traditional plan to a prepaid plan in the 12 months before they switched was 23 per cent less than among those who remained in the traditional plan (Medicare Payment Advisory Commission 2000). Because the individuals had the same insurance plan when this difference was observed, the group opting to change to the prepaid plan appeared to be in considerably better health. Similarly, adjusted for age and sex, mortality rates among those enrolled in prepaid plans were 15 per cent less than among those in the traditional plan, a difference much too large to be plausibly related to any difference in benefits or care. Consistent with selection, the mortality rate difference was largest, 21 per cent, among those who had switched to the prepaid plan within the previous 12 months, and then steadily narrowed as individuals remained in the prepaid plan, a form of regression to the mean.

Finally, selection can give rise to so-called premium death spirals. Cutler and Reber (1998) studied a natural experiment among Harvard University employees that gave rise to one such spiral. Harvard allowed its employees to choose among insurance plans of varying generosity. Initially it subsidized a constant percentage of the premium (between 75 and 85 per cent, depending on the employee's earnings), but it subsequently changed the subsidy to a lump sum. (I note in passing that the rationale for an employer subsidy is to combat selection. In effect, such a subsidy makes it attractive for low risks to pool with high risks within the employment group.)

With a percentage-of-premium subsidy, the employee bore only 15–25 per cent of the premium difference among plans, but with a lump sum subsidy the employee bore the full incremental cost of more generous plans. For the employee who only marginally favoured a more generous plan with a percentage-of-premium subsidy (that is, the better risk within the group choosing the more generous plan), it became attractive with a lump sum to choose a less generous plan (the relative price to the employee of more generous plans rose by a factor of four or more). But, as the better risks within the more generous plans opted out, the premium necessary to cover the medical cost of those remaining rose. This in turn set off another round of plan changing, which raised the premium in the more generous plans still more. At that point the most generous plan was withdrawn from the market.

Moral Hazard

Insurance creates a trade-off between risk aversion and moral hazard, or the failure to take actions that would lessen the probability of or the severity of or damage from an adverse event (Zeckhauser 1970). In the context of health insurance, the focus has been on the costliness of the event rather than on the likelihood of the event. That is because there are enough unpleasant uninsured consequences around illness and injury, such as pain and discomfort, that the extent of insurance for medical care probably changes the incentive to avoid illness or injury rather little. In fact, individuals when randomly assigned to better insurance do not change their lifestyle habits (Newhouse and the Insurance Experiment Group 1993).

The RAND Health Insurance Experiment randomized 2000 American families to health insurance plans that varied the portion of medical bills they had to pay, from nothing (all services were free to the family) to approximately a large family deductible of $1,000 in late 1970s dollars (Manning et al. 1987; Newhouse and the Insurance Experiment Group 1993). The deductible was scaled down for low-income families. The families were followed for either three or five years (the period was randomly assigned), and both their medical care use and health outcomes were observed. Over the course of the experiment families assigned to the large deductible plan used around 30 per cent fewer services than those assigned to the plan in which services were free. They made about two fewer visits to physicians during the year, and they were admitted to the hospital about 20 per cent less.

The average family's health was little changed by the additional medical services consumed when care was free to them, although those who were both sick and poor had better outcomes, primarily because of better control of hypertension (high blood pressure). There was thus ample confirmation of moral hazard. For a review of studies of moral hazard see Zweifel and Manning (2000).

The Supply of Medical Care

Insurers or health care services either contract with or employ health care providers, most notably physicians, to deliver medical services. The need for this is most apparent if the insurance policy covers all of the patient's medical costs in full; because the patient has no incentive to search for a lower-cost provider, if the insurer passively reimbursed medical bills providers could in theory bill an infinite amount. The same incentives apply if patients bear a modest fixed charge for each, say, physician visit.

The terms of the insurer's contracts with medical providers can have important effects on the services delivered. I discuss here two features of such contracts: provider networks and so called supply-side cost sharing. (For further discussion of these issues see Chalkley and Malcolmson 2000; McGuire 2000; Newhouse 2002; Pauly 2000.)

Before the 1980s the usual model of American health insurance allowed 'free' choice of provider, meaning that to a first approximation the patient's out-of-pocket payment was unaffected by the physician(s) he or she sought care from. Many non- American models still allow this. These arrangements began to change in the 1980s with the advent of managed care insurance plans, which sought to establish networks of preferred physicians and hospitals. In-network physicians contracted with the insurer; at a minimum the contract specified a discount off a usual fee. In return, patients were given financial incentives to use in-network physicians. By increasing the elasticity of demand facing a physician, networks began to reduce rents in physician fees (Cutler et al. 2000).

Many insurance plans went further than simply asking for fee discounts, and gave physicians financial incentives to reduce utilization. For example, instead of being paid a fee for each narrowly defined service (such as a visit or a laboratory test), a primary care physician might receive a capitation payment for each insured who selected that physician as a personal physician. In this case, the proximate marginal revenue the physician earned from delivering any additional services was zero, although there might have been an indirect effect of reducing services on patient retention. A less high-powered incentive than pure capitation was a 'risk corridor' around a target utilization rate. For example, the physician and the insurer might share deviations from the target rate fifty-fifty up to a certain size deviation, and above or below that amount all risk was on the insurer.

Such arrangements, also used in the British National Health Service and in Denmark, were termed 'supply-side cost sharing', in contrast to the demand-side cost sharing described above that was paid by the patient. Supply-side cost sharing created two incentives. First, and most obviously, it created an incentive for the physician to treat less intensively than with pure fee-for-service reimbursement, thereby potentially addressing moral hazard (Chalkley and Malcolmson 1998, 2000; Ellis and McGuire 1986). Second, if the capitation were only for the primary care physician's own services, the most common arrangement, it created an incentive for that physician to refer the patient to another physician, a form of unbundling. In the American context the insurer would simply pay for the services of the other physician; in the British context the patient would be referred to a salaried physician at a hospital that might have a lengthy queue.

Several pieces of evidence support the view that physicians respond to the type of contract they face. A near-universal finding of American managed care plans is that they reduce the use of services relative to traditional indemnity plans, which passively paid a fee-for-service; that is, they indemnified the patient against any incurred medical bills (Glied 2000). As described above, however, a characteristic of managed care plans is that they did not passively reimburse whatever service a physician chose to deliver.

Other data show the effects of particular contracts. A natural experiment in Denmark showed that physicians whose services were only partly at risk delivered more services than those who were fully at risk, although the increase was not sustained (Krasnik et al. 1990). A small-scale experiment in the United States randomized pediatricians to be paid by either a fee-for-service system or a salary; the salaried pediatricians earned no additional income for treating more intensively (Hickson et al. 1987). The physicians paid with a fee-for-service method delivered more than 20 per cent more services, with the difference almost entirely in well-child visits (preventive care). Most likely, mothers brought sick children in, but brought well children in only with some effort from the pediatrician. Finally, numerous studies have shown that physicians respond to the level of payment (McGuire 2000; Newhouse 2002). This is particularly relevant to insurers that administratively set prices on a take-it-or-leave-it basis and to insurers that negotiate fees for all physicians in an area.

In sum, health insurance markets violate numerous assumptions of the introductory textbook model of perfectly competitive markets with full or at least symmetric information. In particular, information is asymmetric among patients, providers, and insurers, and principal–agent relationships abound. Actual health insurance and health care financing institutions have adapted to these features.

See Also

- Agent-Based Models
- Contract Theory
- Market Competition and Selection
- Risk Sharing
- Social Insurance
- Tragedy of the Commons

Bibliography

Akerlof, G. 1970. The 'market for lemons': Qualitative uncertainty and the market mechanism. *Quarterly Journal of Economics* 74: 488–500.

Arrow, K. 1963. Uncertainty and the welfare economics of medical care. *American Economic Review* 53: 941–973.

Chalkley, M., and J. Malcolmson. 1998. Contracting for health services when patient demand does not reflect quality. *Journal of Health Economics* 17: 1–19.

Chalkley, M., and J. Malcolmson. 2000. Government purchasing of health services. In *Handbook of health economics*, 1Ath ed, ed. A. Culyer and J. Newhouse. North-Holland: Amsterdam.

Cochrane, J. 1995. Time consistent health insurance. *Journal of Political Economy* 103: 445–473.

Cutler, D. 1994. A guide to health care reform. *Journal of Economic Perspectives* 8(3): 13–29.

Cutler, D. 2004. *Your money or your life: Strong medicine for America's health care system*. New York: Oxford University Press.

Cutler, D., and S. Reber. 1998. Paying for health insurance: The tradeoff between competition and adverse selection. *Quarterly Journal of Economics* 113: 433–466.

Cutler, D., and R. Zeckhauser. 2000. The anatomy of health insurance. In *Handbook of health economics*, vol. 1A, ed. A. Culyer and J. Newhouse. Amsterdam: North- Holland.

Cutler, D., M. McClellan, and J. Newhouse. 2000. How does managed care do it? *RAND Journal of Economics* 31: 526–548.

Dubey, P., and J. Geanakoplos. 2002. Competitive pooling: Rothschild–Stiglitz reconsidered. *Quarterly Journal of Economics* 117: 1529–1570.

Ellis, R., and T. McGuire. 1986. Provider behavior under prospective reimbursement. *Journal of Health Economics* 5: 129–151.

Glied, S. 2000. Managed care. In *Handbook of health economics*, vol. 1A, ed. A. Culyer and J. Newhouse. Amsterdam: North-Holland.

Gruber, J. 2000. Health insurance and the labor market. In *Handbook of health economics*, vol. 1A, ed. A. Culyer and J. Newhouse. Amsterdam: North-Holland.

Hickson, G., W. Altmeier, and J. Perrin. 1987. Physician reimbursement by salary or fee-for-service: Effect on physician practice behavior in a randomized prospective study. *Pediatrics* 80: 344–350.

Krasnik, A., et al. 1990. Changing remuneration systems: Effects on activity in general practice. *British Medical Journal* 300: 1698–1701.

Manning, W., et al. 1987. Health insurance and the demand for medical care: Results from a randomized experiment. *American Economic Review* 77: 251–277.

McGuire, T. 2000. Physician agency. In *Handbook of health economics*, vol. 1A, ed. A. Culyer and J. Newhouse. Amsterdam: North-Holland.

Medicare Payment Advisory Commission. 2000. *Report to the congress: Improving risk adjustment in medicare*. Washington, DC: Medicare Payment Advisory Commission.

Newhouse, J. 1992. Medical care costs: How much welfare loss? *Journal of Economic Perspectives* 6(3): 3–21.

Newhouse, J. 1996. Reimbursing health plans and health providers: Selection versus efficiency in production. *Journal of Economic Literature* 34: 1236–1263.

Newhouse, J. 2002. *Pricing the priceless: A health care conundrum*. Cambridge, MA: MIT Press.

Newhouse, J., and the Insurance Experiment Group. 1993. *Free for All? Lessons from the RAND health insurance experiment*. Cambridge, MA: Harvard University Press.

Pauly, M. 2000. Insurance reimbursement. In *Handbook of health economics*, vol. 1A, ed. A. Culyer and J. Newhouse. Amsterdam: North-Holland.

Rothschild, M., and J. Stiglitz. 1976. Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. *Quarterly Journal of Economics* 90: 629–650.

Wilson, C. 1977. A model of insurance markets with incomplete information. *Journal of Economic Theory* 16: 167–207.

Zeckhauser, R. 1970. Medical insurance: A case study of the tradeoff between risk spreading and appropriate incentives. *Journal of Economic Theory* 2: 10–26.

Zweifel, P., and W. Manning. 2000. Moral hazard and consumer incentives in health care. In *Handbook of health economics*, vol. 1A, ed. A. Culyer and J. Newhouse. Amsterdam: North-Holland.

Health Outcomes (Economic Determinants)

Christopher J. Ruhm

Abstract

This article summarizes the conclusions of research examining how health is affected by income inequality and temporary changes in macroeconomic conditions. In both cases, recent analyses have raised serious doubt about the 'conventional wisdom' derived from earlier studies using less adequate analytical methods. Specifically, the latest studies question the hypothesis that inequality has a large independent effect on population health, and suggest that economic downturns improve rather than worsen physical well-being.

Keywords

Absolute income hypothesis; Counter-cyclical; Fixed effects; Health outcomes; Income inequality hypothesis; Inequality of income; Life expectancy; Macroeconomic conditions; Morbidity; Mortality; Pro-cyclical; Relative income hypothesis; Unemployment

JEL Classifications

I1

The health of individuals and populations is affected by a variety of economic factors. Other dictionary entries (health economics; population health, economic implications of) consider how health is related to income, education, infrastructural investments, prices and insurance. This article focuses on the roles of income inequality and of temporary changes in macroeconomic conditions in wealthy industrialized countries, which have been the subjects of considerable debate and empirical analysis. In both cases, the recent use of more sophisticated analytical approaches has raised serious doubt about the 'conventional wisdom' derived from earlier research using techniques less able to account for possible sources of bias.

Income and health are positively correlated. Although there are questions about the extent to which higher incomes *cause* better health (Smith 1999), most analysts believe that there is some causal effect and the discussion below presumes this is so. Similarly, *permanent* economic progress is assumed to improve most aspects of health.

Income Inequality and Health

Conceptual Issues

There is a widespread belief that average health would improve if inequality could be reduced by redistributing income from richer to poorer households, without changing its average level. This is supported by the predictions of theory and empirical evidence, since at least Preston (1975), indicating that the health benefits of income exhibit diminishing returns. A direct consequence is that the health reductions resulting from lowering incomes of the well-off are more than offset by gains to the less advantaged, improving average health. Wagstaff and van Doorslaer (2000) call this the *absolute income hypothesis* (AIH). One important implication of AIH is that income inequality will be negatively related to average health in the cross section, but that this correlation will disappear if individual incomes are controlled for.

Far more controversial is the proposition that inequality has negative effects on health, with individual (or household) income held constant. This is the *income inequality hypothesis* (IIH). Under IIH, an individual living in a country with high inequality will be in worse health than a counterpart with the same income but residing in a nation with a more equal distribution. The main mechanism for this is hypothesized to be that relative income (or position) matters – this is the relative income hypothesis (RIH). For instance, being higher in the income distribution might allow access to goods that promote health and provide individuals with more control over their lives; conversely, low status might raise stress and reduce social trust or cohesion (Wilkinson 1997).

The existence of RIH is not a sufficient condition for the income inequality hypothesis. An additional requirement is that the negative health effects of low rank exceed the gains (if any) accruing to high relative position. On the other hand, since some deleterious effects of increasing inequality (for example, loss of social capital) have adverse effects throughout the distribution, heightened inequality could worsen health of the wealthy as well as the poor. Direct tests of the relative income hypothesis are rare, partly due to the difficulty in defining an appropriate reference group. Advocates therefore often cite animal studies indicating that low-status primates have greater stress levels than their higher-status counterparts. Recently, however, some research provides direct evidence on the role of relative status. For instance, Eibner and Evans (2005) show high rates of mortality, morbidity, and body mass index, and poor self-reported health for persons whose incomes are low relative to a reference group defined by location, race, education and age.

Empirical Evidence

Much interest in the income inequality hypothesis stems from Wilkinson's (1992) influential study. His most important finding was that more equal incomes were strongly positively correlated with life expectancy in nine industrialized countries, while average incomes had little effect. Subsequent research initially focused on cross-national comparisons but, starting in mid-1990s, was increasingly conducted for geographic regions within countries, particularly the United States.

The early studies were criticized on both technical and methodological grounds. For instance, Judge et al. (1998) discuss problems with the data used by Wilkinson (1992), and provide evidence that the results are sensitive to the choice of inequality measures. Moreover, as mentioned, 'ecological studies' using aggregate data will generally reveal a negative correlation between inequality and health (with average incomes controlled for) if income has diminishing benefits, and so cannot distinguish between the absolute income and income inequality hypotheses (Gravelle 1998).

A potential solution is to use micro-data, since IIH predicts that the inequality relationship will persist after individual income is controlled for. Such research has proliferated in recent years, beginning with Fiscella and Franks (1997). These investigations do not, however, fully address the concern that cross-sectional inequality–health relationships may be confounded by omitted factors. Particularly significant, in this regard, is evidence that the association disappears or diminishes greatly when covariates are included for education, census region or per cent black (Mellor and Milyo 2002; Deaton and Lubotsky 2003). Other researchers (for example, Subramanian and Kawachi 2003) argue that one or more of these variables may be caused by inequality and so are not appropriate to control for, or present evidence that a (diminished) correlation persists after including them. Nevertheless, the sensitivity of results suggests the vexing difficulty in accounting for all relevant causal factors when using cross-sectional data.

Whereas early studies provided strong evidence favouring IIH, the conclusions of more recent research, generally using better data and more sophisticated techniques, are far more mixed. An indication of this can be obtained from the comprehensive literature review by Lynch et al. (2004). They classified 98 peer-reviewed studies according to whether they supported IIH, had mixed findings, or obtained no associations or positive estimated effects of inequality on health. Overall, 40 studies contained strongly favourable evidence, 25 had mixed findings, and 33 were not supportive. However, 24 of the 37 studies published after 2001 failed to obtain results consistent with IIH, and just five were strongly favourable. The evidence is also generally less supportive when individual rather than aggregate data are used, particularly for recent analyses. For instance, of the 18 such studies published after 2001 and reviewed by Lynch et al., 12 obtained negative findings, two had mixed results, and only four strongly supported IIH.

After carefully reviewing the literature, Angus Deaton states, 'The stories about income inequality affecting health are stronger than the evidence'

(Deaton 2003, p. 150). This conclusion seems reasonable. There may be some causal effect but it is almost certainly weaker than that suggested by the early research, and is probably confined to a limited set of health outcomes (such as homicides).

Macroeconomic Conditions and Health

Conceptual Issues

Health is conventionally believed to improve during economic expansions and deteriorate during downturns. Psycho-social determinants are usually focused upon, with recessions postulated to harm physical and mental health by increasing stress and risk taking (Brenner and Mooney 1983). However, economic factors could also matter if, for example, incomes fall or medical care becomes more expensive because health insurance becomes less available or comprehensive.

However, there are at least four reasons why health instead might improve in bad times. First, the opportunity cost of time declines, making it less expensive to undertake time-intensive health investments such as exercise and consumption of a healthy diet. Second, health is an input into the production of goods and services, implying that hazardous working conditions, job-related stress and some environmental risks (like pollution) may decrease. Third, some external sources of death may fall. For instance, traffic fatalities are likely to decrease due to reductions in driving. Fourth, migration may fall if individuals have fewer opportunities to move into areas with robust economic conditions. This could reduce social isolation, with especially beneficial effects on the young and old (Eyer 1977).

The health effects of temporary and permanent changes in economic conditions may be quite different. An important distinction is that transitory growth is usually produced through more intensive use of existing inputs, whereas lasting improvements require some combination of technical innovation or increases in productive capital that have the potential to raise all types of consumption, including good health.

Time-Series Evidence

Most empirical research, until recently, analysed aggregate time-series data for a single geographic area. Particularly influential were a series of investigations by M. Harvey Brenner providing evidence that recessions (and other sources of macroeconomic instability) raise overall mortality, specific sources of death, and other health problems. For instance, using data from England and Wales for 1936–76, Brenner (1979) found that unemployment rates (growth in per capita income) were positively (negatively) related to total and age-specific mortality. However, researchers (such as Wagstaff 1985) have pointed out serious technical flaws in Brenner's methods, and studies correcting the problems (for example, McAvinchey 1988) failed to replicate his findings.

A key issue is that any lengthy time series may contain omitted variables that are spuriously correlated with economic conditions and have a causal effect on health. Potential confounders include changes in lifestyles, the public health infrastructure or medical technologies. Given this fundamental shortcoming, it is no surprise that the results of aggregate time-series analyses are sensitive to the countries, time periods, and proxies for health examined. After reviewing 16 such studies, Ruhm concludes: 'with the exception of Brenner's analyses, the majority of the time series evidence suggests that the contemporaneous effect of economic downturns is to improve health or reduce mortality' (Ruhm 2006, p. 5). Interestingly, such 'counterintuitive' findings are not new. Research undertaken as early as the 1920s identifies a positive association between macroeconomic activity and mortality (Ogburn and Thomas 1922).

Estimates Using Pooled Data

A potential solution to the aforementioned shortcoming is to conduct 'a more refined ecological analysis . . . taking advantage of local and regional variations in the business cycle as well as in disease rates' (Kasl 1979, p. 787). Research using such strategies has become increasingly common beginning with Ruhm's (2000) analysis of state-specific mortality rates for 1972–91. The key advantage of using multiple geographic areas and periods is that time effects can be included to account for potential time-varying confounders that have common impacts across locations, and location ‘fixed effects’ can be added to control for unobserved factors that differ across geographic areas but remain constant within them. Researchers also often include area-specific time trends, which hold constant some other time-varying omitted determinants. Although most analyses have utilized aggregate data, the techniques can easily be adapted for use with individual data, and some research has begun to do so.

Mortality rates are the most commonly studied health outcome, and area unemployment rates are the typical proxy for macroeconomic conditions. Although data from the United States was first examined, analysis has recently been conducted for several European nations (for example, Neumayer 2004; Tapia Granados 2005), as well as using international data on multiple countries. Ruhm (2006) reviews seven such studies published between 2000 and 2004, and concludes that there is strong evidence of a *pro-cyclical* variation in total mortality, infant deaths and fatalities from traffic accidents, cardiovascular disease and influenza or pneumonia. The results are mixed for other types of mortality, and it is noteworthy that some studies uncover a counter-cyclical variation in suicides. This raises the possibility that people become ‘healthier but not happier’ when economic conditions deteriorate. Data restrictions have severely limited analyses of morbidities, although the majority of evidence from the few studies available (for example, Ruhm 2003) indicates a counter-cyclical variation in health.

Lifestyle changes appear to explain some of the health improvements observed during economic downturns. Most available research suggests that alcohol use and problem drinking, smoking, severe obesity and physical inactivity all decline when the economy weakens. However, direct evidence on the role of work hours is mixed and disparate results are sometimes obtained for specific population groups or countries other than the United States. Also, the improvements in physical health appear to occur despite reductions in incomes and decreased use of medical care.

See Also

- ▶ Health Economics
- ▶ Population health, economic implications of

Bibliography

Brenner, M.H. 1979. Mortality and the national economy. *The Lancet* 314: 568–573.

Brenner, M.H., and A. Mooney. 1983. Unemployment and health in the context of economic change. *Social Science Medicine* 17: 1125–1138.

Deaton, A. 2003. Health, inequality and economic development. *Journal of Economic Literature* 41: 113–158.

Deaton, A., and D. Lubotsky. 2003. Mortality, inequality and race in American cities and states. *Social Science and Medicine* 56: 1139–1153.

Eibner, C., and W.N. Evans. 2005. Relative deprivation, poor health habits, and mortality. *Journal of Human Resources* 40: 591–620.

Eyer, J. 1977. Prosperity as a cause of death. *International Journal of Health Services* 7: 125–150.

Fiscella, K., and P. Franks. 1997. Poverty or income inequality as predictor of mortality: Longitudinal cohort study. *British Medical Journal* 314: 1724–1727.

Gravelle, H. 1998. How much of the relationship between population mortality and unequal distribution of income is a statistical artifact? *British Medical Journal* 316: 382–385.

Judge, K., J.A. Mulligan, and M. Benzeval. 1998. Income inequality and population health. *Social Science and Medicine* 45: 567–579.

Kasl, S.V. 1979. Mortality and the business cycle: Some questions about research strategies when utilizing macro-social and ecological data. *American Journal of Public Health* 69: 784–788.

Lynch, J., G.D. Smith, S. Harper, M. Hillemeier, N. Ross, G.A. Kaplan, and M. Wolfson. 2004. Is income inequality a determinant of population health? Part 1. A systematic review of the literature. *Milbank Quarterly* 82: 5–99.

McAvinchey, I.D. 1988. A comparison of unemployment, income and mortality interaction for five European countries. *Applied Economics* 20: 453–471.

Mellor, J.M., and J. Milyo. 2002. Income inequality and health status in the United States: Evidence from the current population survey. *Journal of Human Resources* 37: 510–539.

Neumayer, E. 2004. Recessions lower (some) mortality rates. *Social Science & Medicine* 58: 1037–1047.

Ogburn, W.F., and D.S. Thomas. 1922. The influence of the business cycle on certain social conditions. *Journal of the American Statistical Association* 18: 324–340.

Preston, S.H. 1975. The changing relationship between mortality and the level of development. *Population Studies* 29: 249–257.

Ruhm, C.J. 2000. Are recessions good for your health? *Quarterly Journal of Economics* 115: 617–650.

Ruhm, C.J. 2003. Good times make you sick. *Journal of Health Economics* 22: 637–658.

Ruhm, C.J. 2006. Macroeconomic conditions, health and mortality. In *Elgar companion to health economics*, ed. A.M. Jones. Cheltenham: Edward Elgar.

Smith, J.P. 1999. Healthy bodies and thick wallets: The dual relationship between health and economic status. *Journal of Economic Perspectives* 13: 145–166.

Subramanian, S.V., and I. Kawachi. 2003. The association between state income inequality and worse health is not confounded by race. *International Journal of Epidemiology* 32: 1022–1028.

Tapia Granados, J. 2005. Recessions and mortality in Spain, 1980–1997. *European Journal of Population* 21: 393–422.

Wagstaff, A. 1985. Time series analysis of the relationship between unemployment and mortality: A survey of econometric critiques and replications of Brenner's studies. *Social Science and Medicine* 21: 985–996.

Wagstaff, A., and E. van Doorslaer. 2000. Income inequality and health: What does the literature tell us. *Annual Review of Public Health* 21: 543–567.

Wilkinson, R.G. 1992. Income distribution and life expectancy. *British Medical Journal* 304: 165–169.

Wilkinson, R.G. 1997. Health inequalities: Relative or absolute material standards. *British Medical Journal* 314: 591–595.

Health State Evaluation and Utility Theory

Alastair McGuire

Abstract

Valuing health outcomes is a fundamental concern in health economics. This article considers a measure of health outcomes: the Quality Adjusted Life Year (QALY). The QALY has been used extensively for two main reasons: (1) it arguably values health outcomes in a more acceptable metric than money does; and (2) it feeds more easily into the wider medical decision-making. To be an acceptable measure of health state preferences, however, the QALY requires a number of restrictive assumptions to hold. We discuss these assumptions and conclude that, if these do not hold, the QALY reverts to a measure of health state rather than to a health state preference.

Keywords

Disability Adjusted Life Years (DALY); Quality Adjusted Life Year (QALY); Social policy; Utility theory; Years of Healthy Life (HYL)

JEL Classifications

I120; I180

Valuing health outcomes is a fundamental concern in health economics. Given that extensive market failure pervades this sector, the efficient allocation of resources requires a workable definition of the valuation of health outcomes. While there is a long history associated with the valuation of (statistical) life based on willingness to pay and other Hicksian measures derived from monetary valuations, the trend in health economics has been to attempt to value health states using a different metric: the Quality Adjusted Life Year (QALY). The QALY has been used extensively for two main reasons. First, it is arguably easier to measure than monetary valuation of health states while remaining comparable across different disease areas, allowing direct assessment of resource use across any health sector. Second, as we outline below, given the assumptions required to allow the QALY to value health states, for some, the measure may merely allow the representation of a health state on a quantitative scale, allowing decision makers to attach value to resource use through identification of 'acceptable' QALY levels. So the QALY has been used extensively as it either values health outcomes in a more acceptable metric than money and/or because it feeds easily into the wider medical decision-making process.

Valuation of health states in economics builds on utility theory. Utility theory imposes structure on an individual's choice across all commodities. The individual, choosing from a large bundle of different quantities of various combinations of commodities, does so in a manner that is reflexive, transitive and subject to a behavioural restriction that satiation does not occur. Reflexiveness ensures that if bundle A is chosen over bundle B, then bundle B must never be chosen over bundle A from the same choice set. Transitivity ensures that if bundle A is chosen in preference to bundle B, and bundle B is chosen in preference to bundle C then bundle A will be chosen over bundle C. Non-satiation states that if an individual becomes satiated with any given bundle, no more of that bundle is chosen. If indifference curves are used to analyse such choices, continuity is normally also assumed such that if bundle A is preferred to B and bundle B is preferred to C, there must exist some bundle D, a weighted average of bundles A and C, such that the individual is indifferent between bundles B and D. Under these assumptions the individual is said to be rational with respect to their choices.

Given this structure a number can be attached to each bundle, and bundles with higher numbers are chosen over bundles with lower numbers. These numbers are ordinal if the same order of choice ranking is preserved when one set of numbers is replaced with another set. These choices define a utility function. Consumers are assumed to maximise utility functions.

Ordinal utility functions do not provide information on the differences between chosen bundles. Extending the preference ordering to consider choice under uncertainty provides information on the relativities between bundles. Here the utility function expresses expected utility across the sum of the utilities of the expected outcomes defined as the sum of each prospect multiplied by the probability of its occurrence:

$$\begin{aligned} U(x_1, x_2, x_3, \ldots x_n) &= p_1 U(x_1) + P_2 U(x_2) \\ &\quad + p_3 U(x_3) + \ldots \\ &\quad + p_n U(x_n) \\ &= \sum_{i=1}^{n} p_i U(x_i) \end{aligned} \tag{1}$$

This utility function was first proposed by Von Neumann and Morgenstern (1944) who stated (p. 18): ‘Let p be a real number between 0 and 1 such that A is exactly equally desirable with the combined event consisting of a chance of probability $1 - p$ for B and the remaining chance of probability p for C. Then we suggest the use of p as a numerical estimate for the ratio of the preference of A over B to that of C over B’. Defining this utility function up to a linear transformation, such that any positive affine transform of the value of p fully describes choice, a cardinal utility function is returned. As a function it provides cardinal information on the differences between chosen bundles through the fact that the preference for one bundle over another is expressed as a ratio.

This formulation of cardinal utility function is additively separable: the utility of any particular outcome, $p_i U(x_i)$, is independent of all others. Moreover, this expected utility function incorporates risk preference.

This formulation of rational choice underpins normative economics, informing policy makers on how rational individuals ought to behave when faced with resource decisions derived from economic or social policy. The expected utility function and the rational choice stemming from the set of axioms which dictate rational behaviour allow consistent, logical arguments to be made in support of policy implementation. Policy makers can use expected utility theory, most commonly through the use of cost–benefit analysis, to determine whether a given rational individual is better or worse off after some policy enactment. Higher utility, for an individual, is a criterion of improvement.

Note that expected utility need not provide information on actual choices. In reality, individuals make mistakes, may be inconsistent and need not be rational in the sense of the definition given above. In positive economics, that is where choice is observed, utility theory plays a limited role. It may provide some benchmark, but nothing more. It may define the starting assumptions when specific motivations are analysed, particularly when assumptions are required to initiate any experiment in behaviour. However, it need play no role whatsoever in such experiments. As Little (2002) put it, some economists ‘even try to test utility theory: but that should perhaps count as psychology rather than economics’.

That individuals do not necessarily reflect expected utility theory in choices observed under experimental study need not negate the use of this theory for policy analysis. Given rational choices made under idealised information holding, expected utility provides a benchmark for the assessment of policy on welfare improvement at the individual level.

Expected utility theory is applied to health policy through the application of multiattribute utility (Keeney and Raiffa 1976; Weinstein and Stason 1977; Zeckhauser and Shepard 1976). It is assumed that a treatment intervention leads to an uncertain outcome, but the outcome is a permanent change in an individual's health state over the rest of their life. This presumes that chronic health states are dealt with; transitory states would by definition not be permanent. The outcome is normally taken to be a product of both a set of possible health states, Q, and a set of possible life durations, T, where Q and T are the different attributes (or components) of the outcome. The outcome set, (Q,T), is therefore a product set of all possible combinations of the attributes; hence the term *multiattribute*. Expressing in terms of the expected utility of health (*H*), drawing on 1 above, we have:

$$EU_H = \sum_{i=1}^{n} p_i U\left((QT)_i\right) \tag{2}$$

This can be transformed into a common multi-attribute health scale, the Quality Adjusted Life Year (QALY), by imposing the specific form as Pliskin et al. (1980) do:

$$EU_H = \sum_{i=1}^{n} p_i U(T_i{}^* v(Q_i)) \tag{3}$$

where $v(Q_i)$ defines the utility attached to health states. As in expected utility generally, the function $v(Q_i)$ is cardinal and unique up to a linear function. Thus the function can be scaled such that $v(Q_i)$ lies within the bounds [0,1], with 0 representing the utility attached to death and 1 representing the utility attached to a full year of health.

Miyamoto et al. (1998) prove that the multiplicative utility form of the QALY multiattribute function is equivalent to the imposition of monotonicity in life duration, standard gamble invariance and the imposition of a zero condition on the utility function. Or, put simply, these three assumptions return a QALY model of multiplicative form. Monotonicity in life duration states that, for a given quality of life a longer life is preferred to a shorter life. Standard gamble invariance states that preferences over prospects relating to duration of life are independent of the given quality of life assumed. This can also be imposed on prospects over quality of life, in which case the preferences are independent of the given duration of life assumed. Standard gamble invariance is commonly referred to as mutual utility independence. The zero condition requires that all (*Q*,0) are equivalent; that is, all quality of life states are the same if you are dead. This is not an arbitrary assumption, but fixes an anchor point. Functional forms of the QALY utility function other than multiplicative are feasible and rely on different assumptions; in particular relating to utility independence. The multiplicative form is by far the most commonly used.

As with expected utility functions generally, risk preference may be incorporated. Following Pliskin et al. (1988) however it is common to assume risk neutrality with respect to duration of life: that is that, for any given fixed quality of life, individuals are indifferent between a certain life, *T*, and the prospect defined over an uncertain duration of life coupled with a life expectancy of *T*. More directly, given a fixed quality of life, individuals are at some level presumed indifferent to the certainty equivalent of any given prospect and the prospect itself. Pliskin et al. (1988) also impose standard gamble invariance and constant proportional trade-off. The latter states that the proportion of remaining life that an individual is willing to trade for an improvement in the quality of life is independent of the remaining duration of life. Once again, these three assumptions (risk neutrality, standard gamble invariance and constant proportional trade-off) return a QALY model of multiplicative form.

Bleichrodt et al. (1997) show that merely imposing risk neutrality over life years and the zero condition can return the multiplicative form of the QALY function, as risk neutrality is consistent with standard gamble invariance with respect to life duration. Moreover, they extend their model to health states worse than death.

Broome (1993) shows that if health state preferences are defined across a range of different lifetimes for each individual, each projecting variable lengths of life and quality of life and the individual applies discounting, then the obvious tractable manner in which to do so is to assume separability. That is, an individual's preferences over their quality of life in any one-year does not affect their preferences over quality of life in any other year. Indeed, Broome argues that his specification of a cardinal utility function describing QALYs, even although it relies on the separability of quality of life across different time periods, is preferred to the other multiplicative versions of a multiattribute QALY function, as the assumption that there is separability over lotteries with regard to quality of life, where length of life has to be held constant (or lotteries over length of life where quality of life has to be held constant). Is an unreasonable assumption. In short, the choice of functional form and the assumptions underlying it matter when defining a QALY.

Indeed, there is a large literature that has developed alternative theories of decision making, either as a competitor to the normative theory of expected utility or based on notions of behaviour deemed rational in some sense, but inconsistent with expected utility. The normative competing theories follow the pattern of developments in decision theory generally where a large number of alternatives to expected utility theory have emerged, prospect theory perhaps being the most robust. Bleichrodt and Pinto (2005) examined a nonlinear QALY that which drops the assumption of linearity imposed on life duration, allowing curvature in the utility attached to life duration – consistent with decreasing marginal utility of life duration. This model and the expected utility QALY model were shown in Bleichrodt and Quiggin (1997) to be consistent with a general class of decision-theoretic models. They found experimental support for their nonlinear QALY model, although as they pointed out this might be an artefact reflecting an individual's valuation of quality of life to be lower in old age. A multiplicative functional form was also supported by their experimental data, although there is other conflicting evidence on this (Pauker 1976; Dolan and Stalmeier 2003). Certainly much remains to be undertaken on the formulation of competing QALY models to the dominant expected utility formulation.

While there is limited acknowledgement that the multiplicative functional form deals with chronic health states alone, Bleichrodt and Quiggin (1997) show that additive independence and symmetry return a QALY model that is capable of describing non-chronic states. Here QALY models revert to different quality of life profiles as acute episodes alter future health profiles, such that:

$$QALY = \sum_{t=1}^{T} H(q_t) \tag{4}$$

where $H(q)$ represents a health profile attributable to a given time interval t. Additive independence of preferences states that all health profiles through time are valued individually. Symmetry implies that the valuation of health quality is not affected by its timing. There is little evidence to support use of the additive QALY functional form (see Spencer 2003).

Even if agreement is reached over the normative approach, the actual calculation of a QALY relies on the measurement of preferences for different health states. In other words, the use of QALYs in health resource allocations moves us from the normative to the positive.

Some have argued that individuals may make systematic biases in attempting to measure preferences associated with their quality and length of life. Dolan and Kahneman (2008) argue that such preferences are liable to be distorted by an individual's own experiences and that, in any case, as health states change individuals will adapt, so who to ask also becomes important. Others argue that the instruments used to measure such preferences are not well understood and may likewise impart biases.

Three instruments are common. The standard gamble is the closest instrument to the expected utility choice situation, and follows from the Von Neumann and Morgenstern definition above. Obviously, individuals must understand, at least intuitively, the concept of probability. They must also be trading as if risk neutral, otherwise risk preferences will contaminate the cardinal values. Both seem unlikely in practice and there is little empirical support (see, for example, Oliver (2005)). The time-trade off instrument measures preference by facing an individual with a choice of a shorter period of full health, then immediate death, versus a longer period in ill health followed by immediate death. At the point of indifference the ratio of the time in full health to ill health provides the cardinal measure of preference of the worse state to the full health state. Again, individuals must recognise what is being chosen. They must also not be exercising a discount rate when making the choice. Constant proportionality must also hold, such that 3 years in full health compared to 4 years in ill health is equivalent to 9 years in full health compared to 12 years in ill health. So length of time in an ill state *per se* does not count. The final common instrument is a visual analogue scale, where individuals simply rate their state of health on a scale; normally calibrated with 0 as death and 1 as full health. This does not represent a choice and therefore is not consistent with preference elicitation. However, Broome (1993) has argued that the QALY may reflect the 'goodness' of or benefit from a state of health, rather than the preference, and this would be captured by this rating value.

There are other contenders in terms of health state valuation. Other instruments such as Years of Healthy Life (HYL) and health-adjusted healthy life are QALYs in all but acronym (Berthelot et al. 1993; Erickson et al. 1995). Mehrez and Gafni (1981) proposed values based on health profiles. Here various health states are considered in different sequences of event (profiles), and individuals trade off the number of years in perfect health against the years in profile that they deem equivalent.

Disability Adjusted Life Years (DALYs), which estimate life expectancy lost and weight this by the number of years lived in disability, are possibly the most commonly proposed alternative. These have not been formally assessed within a decision-analytic framework, but Airoldi and Morton (2007) argue that once age weighting and differences in discounting in the DALY calculation have been made and adjustments made to allow a comparison between loss in quality of life and the disability weighting in the DALY, the two valuation concepts do not differ much. Both Airoldi (2007) and Sassi (2006) found, however, that the actual estimates of health change based on the two approaches do differ systematically. Certainly, if the concepts are similar all of the limitations of the QALY would appear to apply to the DALY also, but note there has been no systematic assessment of the DALY as it relates to expected utility theory. In their assessment, Airoldi and Morton (2009) favour the QALY overall.

Other competing theories have focused on population rather than individual valuations on the basis that this appears more realistic. Such theories are obviously not as well founded on expected utility theory but may have merit. For example, Nord (1992), following Patrick et al. (1973) suggested the use of the person trade-off where the aggregate social value of different health care interventions is sought. Essentially the person trade-off asks individuals to value a health intervention by quantifying directly how many outcomes of one kind they consider equivalent in social value to X outcomes of another kind. Even Nord (1994) acknowledges that this approach is, in practice, demanding if the returned estimates are to have internal validity.

One of the advantages of the person trade-off is that, if the valuation relates to a numerical equivalent (for example, how many people with this health state are equivalent to 10 in perfect health?) the responses reflect views on interpersonal comparison. Explicit information on interpersonal comparison is not required. With the traditional QALY approach aggregation of choices across individuals is required. Here weighting of QALYs as they are gained and lost by different individuals may be required. If the QALY valuation differs even just by age for example, this may require calculation of an array of social weights. Moreover, from a traditional welfare perspective the QALY would only form one part of overall utility unless the health component is separable from other components. The latter is the approach taken by so-called 'extra-welfarism', where QALYs form the basis of preference-based health state measurement.

The QALY is the most prevalent measure of health state valuation. Is it a utility measure also? As we have seen, probably no. For a QALY to be a measure of utility, even where separability is imposed on the utility function between health and other commodities, various restrictive assumptions have to be made to support this approach. Moreover, the standard gamble approach is more closely aligned to the Von Neumann and Morgenstern expected utility model, yet the time trade-off has become the dominant measurement instrument, moving the estimated QALY tariffs further from a utility base. Broome is probably best in summary when he states that the QALY, which (following his definition) relates to an individual who is a discounting QALY minimiser, is best thought of as measuring the benefit or good derived in terms of health from the intervention under scrutiny. This health benefit may be a transformation that translates into utility, but we have little or no information concerning the transformation itself (Broome 1993). On top of that, we have little data on the social weights required to aggregate QALYs for use in matters of resource allocation across individuals, even though this represents the most prevalent use of the QALY measure in Europe at least (for an application of social weighting see Dolan and Tsuchiya 2010). What is fair to state, however, is that QALYs, even if merely considered a measure of health benefit across the dimensions of morbidity and mortality, do appear to provide a baseline, commensurate measure of health outcomes that imparts some information to decision makers in the policy arena.

See Also

- ▶ Health Behaviours, Economics of
- ▶ Health Econometrics
- ▶ Health Economics
- ▶ Health Insurance, Economics of

Bibliography

Airoldi, M. 2007. Gains in QALYs vs. DALYs averted. The troubling implications of using residual life expectancy. LSE Health working paper no. 8. LSE, London.

Airoldi, M., and A. Morton. 2009. Adjusting life for quality or disability: Stylistic difference or substantial dispute? *Health Economics* 18(11): 1237–1247.

Berthelot, J., R. Roberge, and M.C. Wolfson. 1993. The calculation of health adjusted life expectancy for a Canadian province using a multiattribute utility function: A first attempt. In *Calculation of health expectancies: Harmonization, consensus achieved and future perspectives*, ed. J.M. Robine, C.D. Mather, M. Bone, and I. Romieu, 161–172. Montpellier: Eurotext.

Bleichrodt, H., and J.L. Pinto. 2005. The validity of QALYs under non-expected utility. *The Economic Journal* 115: 533–550.

Bleichrodt, H., and J. Quiggin. 1997. Characterizing QALYs under a general rank-dependent utility model. *Journal of Risk and Uncertainty* 15: 151–165.

Broome, J. 1993. QALYs. *Journal of Public Economics* 50: 149–167.

Dolan, P., and D. Kahneman. 2008. Interpretations of utility and their implications for the valuation of health. *Economic Journal* 118(525): 215–234.

Dolan, P., and P. Stalmeier. 2003. The validity of time trade-off values in calculating QALYs: Constant proportional time trade-off versus the proportional heuristic. *Journal of Health Economics* 22(3): 445–458.

Dolan, P., and A. Tsuchiya. 2010. Determining the parameters in social welfare function using stated preference data: An application to health. *Applied Economics* 43(18): 2241–2250.

Erickson, P., R.W. Wilson, and L. Shannon. 1995. *Years of healthy life: Statistical note number 7*. Hyattsville: National Center for Health Statistics.

Keeney, R.L., and H. Raiffa. 1976. *Decisions with multiple objectives: Preferences and value tradeoffs*. New York: Wiley.

Little, I.M.D. 2002. *Ethics, economics and politics*. Oxford: Oxford University Press.

Mehrez, A., and A. Gafni. 1993. Healthy-years equivalents versus quality-adjusted life years: In pursuit of progress. *Medical Decision Making* 13(4): 287–292.

Miyamoto, J.M., P.P. Wakker, H. Bleichrodt, and H.J.M. Peters. 1998. A zero-condition: A simplifying assumption in QALY measurement and multiattribute utility. *Management Science* 44: 839–849.

Nord, E. 1992. Methods for quality adjustment of life years. *Social Science & Medicine* 34: 559–569.

Nord, E. 1994. The QALY – a measure of social value rather than individual utility? *Health Economics* 3(2): 89–93.

Oliver, A. 2005. Testing the internal consistency of the lottery equivalents method using health outcomes. *Health Economics* 14(2): 149–159.

Patrick, D.L., J.W. Bush, and M.M. Chen. 1973. Methods for measuring levels for well-being for a health status index. *Health Services Research* 8: 228–245.

Pauker, S.G. 1976. Coronary artery surgery: The use of decision analysis. *Annals of Internal Medicine* 85: 8–18.

Pliskin, J.S., D.S. Shepard, and M.C. Weinstein. 1980. Utility functions for life years and health status. *Operations Research* 28: 206–224.

Sassi, F. 2006. Calculating QALYs, comparing QALY and DALY calculations. *Health Policy and Planning* 21: 402–408.

Spencer, A. 2003. A test of the QALY model when health varies over time. *Social Science and Medicine* 57: 1697–1706.

Von Neumann, J., and O. Morgenstern. 1944. *The theory of games and economic behaviour*. Princeton: Princeton University Press.

Weinstein, M.C., and W.B. Stason. 1977. Foundations of cost-effectiveness analysis for health and medical practices. *New England Journal of Medicine* 296: 716–721.

Zeckhauser, R., and D.S. Shepard. 1976. Where now for saving lives? *Law and Contemporary Problems* 40: 5–45.

Hearn, William Edward (1826–1888)

M. White

Hearn was born in County Cavan, Ireland and died in Melbourne, Australia. Educated at Trinity College Dublin, he was appointed professor of political economy and other subjects at the University of Melbourne in 1854. Subsequently a member of the Legislative Council of the State of Victoria and contributor to the local press, Hearn is known to economists principally for his *Plutology* (1863).

Plutology explains increasing wealth as a result of the competitive exchange of services. The analysis owes a good deal to Herbert Spencer and Frederic Bastiat. Competition is held to have three general results. It is: beneficent, since prices reflect the minimum cost of procuring a service; just, because recompense is in proportion to merit; and equalizing, since no recompense permanently reflects the effects of chance. As an ‘unfailing rule’, the pursuit of self-interest means services are produced in ‘order of their social importance’. Competition results in a natural social order, ordained by Providence, in which the principles of Darwin’s natural selection are applied to industry (chapter “▶ Employment, Theories Of”).

The price of any service, determined by the extent of demand and supply, oscillates towards the minimum cost of production. The upper price limit is set where the purchaser equates desire for a service with the sacrifice necessary to either directly produce or obtain it from another source. The minimum price must cover any outlays and provide the ‘average’ reward for the vendor’s type of service (chapter “▶ Auspitz, Rudolf (1837–1906)”). The discussion of price formation is not conducted in marginalist terms and owes a good deal, via J.S. Mill, to de Quincey.

The distribution of income is explained according to the general principles of exchange. The manager of an enterprise, for example, contracts with the vendors of labour and ‘capital’ for their services at a fixed price. Discounting all costs and gross returns, the manager then has full title to the output, assuming responsibility for losses and receiving net gains. If capital is supplied in ‘commodity’ form (machinery, buildings), rent is paid; if it is supplied in money form (loans, insurance), interest accrues. Directly following Bastiat’s *Harmonies*, Hearn argues that ground rent cannot be a gratuitous gift of nature as land has a price only if labour is bestowed on it (chapter “▶ Equity”).

The role of a central government in an ‘advanced’ nation is thus basically a nightwatchman, although it may undertake some limited regulation. It is acknowledged, however, that the accumulation path will be impeded to some extent. The most serious problems result from enterprises mistaking market demand and engaging in speculative ventures. Still, fluctuations in output and investment have relatively little importance. ‘Failures, poverty, suffering and privation’ are not part of the ‘ordinary course of events’, any ‘ravages’ are soon repaired and objects destroyed in commercial fluctuations would have mainly been consumed rather than invested. Any ‘disturbances’ are thus ‘incidental’ to the natural laws of economic organization (chapter “▶ Famines”).

Marshall and Edgeworth bestowed high praise on *Plutology*, while Jevons considered its arguments were 'nearly identical' to those in his *Theory*. Subsequent commentary has noted Hearn's dogmatism and plagiarism, especially from John Rae and Bastiat.

Selected Works

1851. *The Cassell prize essay on the condition of Ireland*. London.

1863. *Plutology: Or the theory of the efforts to satisfy human wants*. Melbourne: Robertson; London: Macmillan, 1864.

1867. *The government of England, its structure and its development*. Melbourne: Robertson; London: Longmans Green, Reader & Dyer.

1878. *The Aryan household, its structure and its development*. Melbourne: Robertson; London: Longmans, Green & Co., 1879.

1883. *The theory of legal duties and rights*. Melbourne: Government Printer; London: Trubner & Co.

References

Copland, D.A. 1935. *W.E. Hearn: First Australian economist*. Melbourne: Melbourne University Press.

La Nauze, J.A. 1949. *Political economy in Australia*. Melbourne: Melbourne University Press.

Heavy-Tailed Densities

Rustam Ibragimov

Abstract

This article reviews several frameworks commonly used in modelling heavy-tailed densities and distributions in economics, finance, risk management, econometrics and statistics. The results and conclusions discussed in the article indicate that the presence of heavy tails can either reinforce or reverse the implications of a number of models in these fields, depending on the degree of heavy-tailedness.

Keywords

Heavy-tailed densities; Heavy-tailed distributions; Power laws; Stable distributions; Mean; Variance; Moments; Semi-heavy tails; Diversification; Portfolio choice; Value at risk; Linear estimators; Sample mean; Efficiency; Regression; Robustness; Dependence; α-symmetric distributions; Models with common shocks

JEL Classification

C10; C13; C16; G11

Several notions and classes of heavy-tailed densities and distributions are available in the economic, financial, statistical and probability literature. A unifying property common to such densities and distributions is that their tails are heavier than in the Gaussian case, either in the sense of faster decay to zero or in the sense of comparisons of heavy-tailedness measures, such as kurtosis.

Heavy-Tailed Models

In models involving a heavy-tailed random variable (rv) X it is usually assumed that the distribution of X has power tails, so that

$$P(|X| > x) \asymp \frac{C}{x^{\alpha}}, \alpha > 0, C > 0, \quad as\ x \to +\infty \tag{1}$$

(here and throughout the article, $f(x) \asymp g(x)$ as $x \to +\infty$ means that $\lim_{x\to+\infty} \frac{f(x)}{g(x)} = 1$). The parameter α in (1) is referred to as the tail index, or the tail exponent, of the distribution of X. An important property of rvs X satisfying (1) is that the absolute moments of X are finite if and only if their order is less than the tail index $\alpha : E|X|^p < \infty$ if $p < \alpha$ and $E|X|^p = \infty$ if $p \geq \alpha$.

Examples satisfying (1) include Pareto distributions with densities $f(x) = \alpha x_0^{\alpha}/x^{\alpha+1}$ for $x > x_0 > 0$; $f(x) = 0$ for $x \leq x_0$. In addition, (1) is satisfied for Student's t – distributions with densities $f(x) = \frac{\Gamma((\alpha+1)/2)}{\sqrt{\alpha\pi}\Gamma(\alpha/2)}\left(1 + x^2/\alpha\right)^{-(\alpha+1)/2}$, $x \in R$, where $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt, z > 0$, denotes the Gamma function. Relation (1) also holds for the important class of stable distributions that are closed under portfolio formation.

In addition to distributions that follow power laws (1), several other frameworks for modelling heavy-tailed phenomena have been proposed in the literature, including distributions with finite moments of any order and semi-heavy tails. Such tails are thinner than in the case of any power law (1) but much heavier than those of normal distributions. Semi-heavy tails in this sense are exhibited, for instance, by normal inverse Gaussian and, more generally, generalized hyperbolic distributions (see section 3.2 in McNeil et al. 2005, and references therein), as well as by the important case of log-normal distributions.

Empirical Results

Numerous studies in economics, finance, risk management and insurance have indicated that distributions of many variables of interest in these fields exhibit deviations from Gaussianity, including those in the form of heavy tails (1) (see, among others, the discussion and reviews in Embrechts et al. 1997; Rachev et al. 2005). This stream of literature goes back to Mandelbrot (1963) (see also Fama 1965, and the papers in Mandelbrot 1997), who pioneered the study of heavy-tailed distributions in economics and finance.

The following is a sample of estimates of the tail index α in distributions satisfying (1) for returns on various stocks and stock indices: $3 < \alpha < 5$ (Jansen and de Vries 1991), $2 < \alpha < 4$ (Loretan and Phillips 1994), $1.5 < \alpha < 2$ (McCulloch 1997), $0.9 < \alpha < 2$ (Rachev and Mittnik 2000), $\alpha \approx 3$ (Gabaix et al. 2003). Power laws (1) with $\alpha \approx 1$ (Zipf laws) have been found to hold for firm sizes and city sizes (see Gabaix 1999a, b; Axtell 2001). As discussed by Nešlehova et al. (2006), tail indices less than 1 are observed for empirical loss distributions of a number of operational risks. Silverberg and Verspagen (2007) report the tail indices α to be significantly less than 1 for financial returns from technological innovations. The analysis in Ibragimov et al. (2008) indicates that the tail indices may be considerably less than 1 for economic losses from earthquakes and other natural disasters. Anderson (2006) discusses the heavy-tailedness paradigm in many modern economic and financial markets transformed by the Internet and the development of technology.

Stable Distributions

Canonical examples of power laws (1) are given by stable distributions. For $0 < \alpha \leq 2$ and $\sigma > 0$, the symmetric stable distribution $S_\alpha(\sigma)$ is the distribution of an rv X with the characteristic function (cf) $E(e^{ixX}) = \exp\{-\sigma^\alpha|x|^\alpha\}$, $i^2 = -1, x \in \mathrm{R}$. Throughout the article, we write $X \sim S_\alpha(\sigma)$, if the rv X has the distribution $S_\alpha(\sigma)$.Given two rvs X and Y, the notation $X = {}^dY$ means that the distributions of X and Y are the same.

The parameters α and σ are referred to as the characteristic exponent (index of stability) and the scale parameter of the symmetric stable distribution $S_\alpha(\sigma)$. In general, stable distributions also depend on the skewness parameter β and the location parameter μ. Symmetric stable distributions $S_\alpha(\sigma)$ correspond to the case $\mu = \beta = 0$.

A closed form expression for the density $f(x)$ of a stable distribution is available only in the following cases: normal densities that correspond to the case $\alpha = 2$; Cauchy densities $f(x) = \sigma/\left(\pi\left(\sigma^2 + (x-\mu)^2\right)\right), x \in \mathrm{R}$, with $\alpha = 1$; and the densities $f(x) = (\sigma/(2\pi))^{1/2}$ exp $(-\sigma/2x))x^{-3/2}, x > 0; f(x) = 0, x \leq 0$, of Lévy distributions with $\alpha = 1/2$ and their shifted and reflected versions. While normal and Cauchy distributions are symmetric about the location parameter μ, Lévy distributions are concentrated on the positive semi-axis $(0, \infty)$.

The index of stability α characterizes heaviness of tails of the distribution $S_\alpha(\sigma)$. If $X \sim S_\alpha(\sigma)$, then X satisfies power law (1). Thus, the absolute moments $E|X|^p$ of an rv $X \sim S_\alpha(\sigma)$, $\alpha \in (0, 2)$ are finite if $p < \alpha$ and are infinite otherwise. The same conclusions hold for skewed stable distributions. In particular, second and higher moments are infinite for all non-Gaussian stable distributions with $\alpha < 2$. Cauchy distributions with $\alpha = 1$ have infinite first and higher absolute moments. If the rv X has a Lévy distribution with $\alpha = 1/2$, then $E|X|^p = \infty$ for all $p \geq 1/2$.

The distributions of stable rvs X with $\alpha > 1$ are moderately heavy-tailed in the sense that they have finite first absolute moments: $E|X| < \infty$. In contrast, the distributions of stable rvs X with $\alpha < 1$ are extremely heavy-tailed in the sense that their first absolute moments are infinite: $E|X| = \infty$.

The scale parameter σ is a generalization of the standard deviation; it coincides with the standard deviation for normal distributions with $\alpha = 2$. For $\alpha > 1$, the location parameter μ of a stable distribution coincides with its mean: in particular, $EX = 0$ for symmetric stable rvs $X \sim S_\alpha(\sigma)$ with $\alpha \in (1, 2]$.

Stable distributions are closed under portfolio formation. In particular, if $X_i \sim S_\alpha(\sigma)$, $\alpha \in (0, 2]$, are iid symmetric stable risks, then, for all portfolio weights $w_i \geq 0,\ i = 1, \ldots, n$,

$$\sum_{i=1}^{n} w_i X_i =^d \left(\sum_{i=1}^{n} w_i^\alpha \right)^{1/\alpha} X_1, \qquad (2)$$

or, equivalently, $\sum_{i=1}^{n} w_i X_i \sim S_\alpha(\widetilde{\sigma})$, where $\widetilde{\sigma} = \sigma \left(\sum_{i=1}^{n} w_i^\alpha \right)^{1/\alpha}$ (see Zolotarev 1986; Embrechts et al. 1997; Rachev and Mittnik 2000, for a review of properties of stable distributions).

Multivariate extensions of the stable family such as α – symmetric distributions allow one to model frameworks with a wide range of heavy-tailedness in marginals and dependence among them (see Fang et al. 1990 and the discussion in Ibragimov 2007; Ibragimov and Walden 2007). The class of α – symmetric distributions contains models with common shocks affecting all heavy-tailed risks as well as spherical distributions. Spherical distributions, in turn, include such examples as Kotz type, multinormal, logistic and multivariate α – stable distributions. In addition, they include a subclass of mixtures of normal distributions as well as multivariate t – distributions that were used in the literature to model heavy-tailedness phenomena with dependence and finite moments up to a certain order.

Robustness of Economic Models

Heavy-tailedness has important implications for robustness of many economic models, leading, in a number of settings, to reversals of conclusions of these models to the opposite ones.

This may be illustrated, for instance, by the properties of value at risk (VaR) models and the analysis of diversification and portfolio choice in VaR frameworks under heavy-tailedness (see Embrechts et al. 2002; ch. 12 in Bouchard and Potters 2004; Ibragimov 2005a, b, and references therein).

Given a loss probability $q \in (0, 1)$ and an rv (risk) X we denote by $VaR_q(X)$ the VaR of X at level q, that is, its $(1 - q)$ – quantile: $VaR_q(X) = \inf\{z \in R : P(X > z) \leq q\}$ (in what follows, we interpret the positive values of X as a risk holder's losses).

Throughout the article, $\mathrm{J}_n = \{w = (w_1, \ldots, w_n) : w_i \geq 0, i = 1, \ldots, n, \sum_{i=1}^{n} w_i = 1\}$. For $w = (w_1, \ldots, w_n) \in \mathrm{J}_n$, denote by Z_w the return on the portfolio of risks $X_1, \ldots, X_n$ with weights w.

Denote $\underline{w} = (1/n, 1/n, \ldots, 1/n) \in \mathrm{J}_n$ and $\overline{w} = (1, 0, \ldots, 0) \in \mathrm{J}_n$. The expressions $VaR_q(Z_{\underline{w}})$ and $VaR_q(Z_{\overline{w}})$ are thus the VaRs of the portfolio with equal weights and of the portfolio consisting of only one return (risk). It is natural to think about the portfolio with weights $\underline{w}$ as the most diversified and about the portfolio with weights $\overline{w}$ as the least diversified among all the portfolios with weights $w \in \mathrm{J}_n$.

A simple example where diversification is preferable is provided by the standard case with normal risks. Let $n \geq 2$, $q \in (0, 1/2)$.and let $X_1, \ldots, X_n \sim S_2(\sigma)$ be iid symmetric normal rvs. For the portfolio of $X_i{'}$s with the equal weights $\underline{w} = (1/n, 1/n, \ldots, 1/n)$ we have $Z_{\underline{w}} = (1/n) \sum_{i=1}^{n} X_i =^d (1/\sqrt{n}) X_1$.

Consequently, by positive homogeneity of the VaR, $VaR_q\left(Z_{\underline{w}}\right) = (1/\sqrt{n})VaR_q(X_1) = (1/\sqrt{n})VaR_q$ $(Z_{\overline{w}}) < VaR_q(Z_{\overline{w}})$. Thus, the VaR of the most diversified portfolio with equal weights $\overline{w}$ is less than that of the least diversified portfolio with weights $\overline{w}$ consisting of only one risk Z_1.

Using (2), one can also show (see Ibragimov 2005a, b) that diversification is preferable in the VaR framework for all iid moderately heavy-tailed risks $X_i \sim S_\alpha(\sigma)$ with $\alpha \in (1, 2]$ in the sense that $VaR_q\left(Z_{\underline{w}}\right) \le VaR_q(Z_w) \le VaR_q(Z_w) \le Va$ $R_q(Z_{\overline{w}})$ for all $q \in (0, 1/2)$ and all weights $w \in \mathsf{J}_n$.

The settings where diversification is suboptimal in the VaR framework may be illustrated as follows. Let $q \in (0, 1)$ and let $X_1, \ldots, X_n$ be iid positive risks with a Lévy distribution with the tail index $\alpha = 1/2$ and the density $f(x) = (\sigma/(2\pi))^{1/2}\exp\left(-\sigma/(2x)\right)x^{-3/2}$. Similar to symmetric stable distributions, the portfolios of the risks X_i satisfy (2) with $\alpha = 1/2$. Using (2) for equal weights $w_i = 1/n$, we get $Z_{\overline{w}} = (1/n)\sum_{i=1}^{n} X_i{=}dnX_1$. Consequently, $VaR_q\left(Z_{\underline{w}}\right) = nVaR_q(X_1) = nVaR_q(Z_{\overline{w}}) > VaR_q(Z_{\overline{w}})$. Thus, the VaR of the least diversified portfolios with weights $\overline{w}$ that consists of only one risk is less than the VaR of the most diversified portfolio with equal weights $\underline{w}$.

Relation (2) further implies (see Ibragimov 2005a, b) that the results on diversification suboptimality in the VaR framework continue to hold for all iid extremely heavy-tailed risks $X_i \sim S_\alpha(\sigma)$ with $\alpha \in (0, 1)$ Namely, for such risks, $VaR_q(Z_{\overline{w}})$ $\le VaR_q(Z_w) \le VaR_q\left(Z_{\underline{w}}\right)$ for all $q \in (0, 1/2)$ and all weights $w \in \mathsf{J}_n$.

The results in Ibragimov (2005b) provide portfolio VaR comparisons for convolutions of stable distributions with different tail indices and their extensions to dependence, skewness and heterogeneity, including convolutions of α – symmetric distributions and models with common shocks. Ibragimov and Walden (2007) and Ibragimov et al. (2008) show that the (non-)diversification results in heavy-tailed value at risk models continue to hold for bounded risks. Ibragimov et al. (2008) further use these conclusions to develop a simple model for markets for catastrophic risk in which nondiversification traps may arise.

Implications for Statistical and Econometric Methods

Similar to the portfolio VaR analysis, heavy-tailedness presents a challenge for applications of standard statistical and econometric methods. In particular, as pointed out by Granger and Orr (1972) and in a number of more recent studies (see, among others, ch. 7 in Embrechts et al. 1997, and references therein), many classical approaches to inference based on variances and (auto)correlations such as regression and spectral analysis, least squares methods and autoregressive models may not apply directly in the case of heavy-tailed observations with infinite second or higher moments.

An important simple illustration is provided by the failure of the law of large numbers (LLN) for observations with infinite first moments and variances. When more information about the structure of heavy-tailedness is available, one can obtain more refined results that point out to crucial differences between moderately heavy-tailed and extremely heavy-tailed populations.

Consider the problem of estimating the parameter μ in the simple model

$$X_i = \mu + \eta_i, \tag{3}$$

where η_i are iid errors with an absolutely continuous symmetric distribution. Given a random sample $X_1, \ldots, X_n$ that follows (3), denote by $\widehat{\theta}_n$ (w) the linear estimator $\widehat{\theta}_n(w) = \sum_{i=1}^{n} w_i X_{i.}$

It is well known that, if $E\eta_i^2 < \infty$, then the sample mean is $\overline{X}_n = (1/n)\sum_{i=1}^{n} X_i$ is the best linear unbiased estimator (BLUE) of the population mean $\mu = EX_i$. That is, $\overline{X}_n$ is the most efficient estimator of μ among all unbiased linear estimators $\widehat{\theta}_n(w)$ in the sense of variance comparisons: $Var\left[\overline{X}_n\right] \le Var\left[\widehat{\theta}_n(w)\right]$ for all $w \in \mathsf{J}_n$.

The definition of efficiency based on variance breaks down in the case of heavy-tailed populations with infinite second moments. A natural approach to comparison of performance of estimators under heavy-tailedness is to order them by likelihood of observing their large deviations from the true population parameter. This approach relies on the concept of peakedness of rvs and leads to the following definition.

Let $\widehat{\theta}(v)$ and $\widehat{\theta}(w)$ be two linear estimators of the parameter μ in model (3). The estimator $\widehat{\theta}(v)$ is said to be more efficient than $\widehat{\theta}(w)$ in the sense of peakedness (P-more efficient than $\widehat{\theta}(w)$ for short) if $P\left(\left|\widehat{\theta}(v)-\mu\right|>\varepsilon\right)\leq P\left(\left|\widehat{\theta}(w)-\mu\right|>\varepsilon\right)$ all $\varepsilon\geq 0$, with strict inequality whenever the two probabilities are not both 0 or both 1.

The results in Ibragimov (2007) for general dependent settings such as convolutions of α – symmetric distributions and models with common shocks imply that the sample mean $\overline{X}_n$ is the best linear unbiased estimator of the population mean μ in the sense of P-efficiency for moderately heavy-tailed errors $\eta_i \sim S_\alpha(\sigma)$ with $\alpha > 1$. However, if the errors $\eta_i \sim S_\alpha(\sigma)$ are extremely heavy-tailed with $\alpha < 1$, then P-efficiency of the sample mean is smallest among all linear estimators $\widehat{\theta}(w)$ of the population centre μ with weights $w \in \mathsf{J}_n$.

The conclusions in Ibragimov (2005a) show that, similar to the portfolio VaR analysis and the efficiency properties of linear estimators, many models in economics and related fields are robust to heavy-tailedness assumptions provided the distributions entering these assumptions are moderately heavy-tailed. However, the implications of these models are reversed for distributions with sufficiently heavy-tailed densities.

Conclusion

The results reviewed in this article and those obtained in the recent literature imply that the presence of heavy-tailedness can either reinforce or reverse the implications of many models in economics, finance, econometrics, statistics and risk management, depending on the degree of heavy-tailedness. Typically, the standard implications of the models continue to hold for moderately heavy-tailed distributions. However, these implications may become the opposite ones under sufficient heavy-tailedness. Therefore, the models should be applied with care in heavy-tailed settings, especially in the case of the tail indices close to the value $\alpha = 1$, which in many cases provides the critical robustness boundary.

See Also

- ▶ Lognormal Distribution
- ▶ Pareto Distribution
- ▶ Power Laws

Acknowledgment The author gratefully acknowledges partial support by NSF grant SES-0821024, a Harvard Academy Junior Faculty Department grant and the Warburg Research Funds (Department of Economics, Harvard University).

Bibliography

Anderson, C. 2006. *The long tail*. New York: Hyperion.

Axtell, R.L. 2001. Zipf distribution of U.S. firm sizes. *Science* 293: 1818–1820.

Bouchard, J.-P., and M. Potters. 2004. *Theory of financial risk and derivative pricing: From statistical physics to risk management*, 2nd ed. Cambridge: Cambridge University Press.

Embrechts, P., C. Klüppelberg, and T. Mikosch. 1997. *Modelling extremal events for insurance and finance*. New York: Springer.

Embrechts, P., A. McNeil, and D. Straumann. 2002. Correlation and dependence in risk management: Properties and pitfalls. In *Risk management: Value at risk and beyond*, ed. M.A.H. Dempster, 176–223. Cambridge: Cambridge University Press.

Fama, E. 1965. The behavior of stock market prices. *Journal of Business* 38: 34–105.

Fang, K.T., S. Kotz, and K.W. Ng. 1990. *Symmetric multivariate and related distributions, Vol. 36 of monographs on statistics and applied probability*. London: Chapman & Hall.

Gabaix, X. 1999a. Zipf's law and the growth of cities. *American Economic Review* 89: 129–132.

Gabaix, X. 1999b. Zipf's law for cities: An explanation. *Quarterly Journal of Economics* 114: 739–767.

Gabaix, X., P. Gopikrishnan, V. Plerou, and H.E. Stanley. 2003. A theory of power-law distributions in financial market fluctuations. *Nature* 423: 267–270.

Granger, C.W.J., and D. Orr. 1972. Infinite variance and research strategy in time series analysis. *Journal of the American Statistical Association* 67: 275–285.

Ibragimov, R. 2005a. *New majorization theory in economics and martingale convergence results in econometrics*. Ph.D. dissertation, Yale University, New Haven.

Ibragimov, R. 2005b. *Portfolio diversification and value at risk under thicktailedness*, Harvard University Research Discussion Paper 2086. Available at: http://www.economics.harvard.edu/pub/hier/2005/HIER2086.pdf. Accessed 13 Jan 2009. Forthcoming in *Quantitative Finance*.

Ibragimov, R. 2007. Efficiency of linear estimators under heavy-tailedness: Convolutions of α–symmetric distributions. *Econometric Theory* 23: 501–517.

Ibragimov, R., and J. Walden. 2007. The limits of diversification when losses may be large. *Journal of Banking and Finance* 31: 2551–2569.

Ibragimov, R., D. Jaffee, and J. Walden. 2008. Non-diversification traps in catastrophe insurance markets. *Review of Financial Studies*. Available at: http://rfs.oxfordjournals.org/cgi/content/abstract/hhn021v1. Accessed 13 Jan 2009.

Jansen, D.W., and C.G. de Vries. 1991. On the frequency of large stock returns: Putting booms and busts into perspective. *Review of Economics and Statistics* 73: 18–32.

Loretan, M., and P.C.B. Phillips. 1994. Testing the covariance stationarity of heavy-tailed time series. *Journal of Empirical Finance* 1: 211–248.

Mandelbrot, B. 1963. The variation of certain speculative prices. *Journal of Business* 36: 394–419.

Mandelbrot, B. 1997. *Fractals and scaling in finance: Discontinuity, concentration, risk*. New York: Springer.

McCulloch, J.H. 1997. Measuring tail thickness to estimate the stable index alpha: A critique. *Journal of Business and Economic Statistics* 15: 74–81.

McNeil, A.J., R. Frey, and P. Embrechts. 2005. *Quantitative risk management: Concepts, techniques, and tools*. Princeton: Princeton University Press.

Nešlehova, J., P. Embrechts, and V. Chavez-Demoulin. 2006. Infinite mean models and the LDA for operational risk. *Journal of Operational Risk* 1: 3–25.

Rachev, S.T., and S. Mittnik. 2000. *Stable Paretian models in finance*. New York: Wiley.

Rachev, S.T., C. Menn, and F.J. Fabozzi. 2005. *Fat-tailed and skewed asset return distributions: Implications for risk management, portfolio selection, and option pricing*. Hoboken: Wiley.

Silverberg, G., and B. Verspagen. 2007. The size distribution of innovations revisited: An application of extreme value statistics to citation and value measures of patent significance. *Journal of Econometrics* 139: 318–339.

Zolotarev, V.M. 1986. *One-dimensional stable distributions*. Providence: American Mathematical Society.

Heckman, James (Born 1944)

Richard Blundell, Lars Peter Hansen and Derek Neal

Abstract

James Heckman has made fundamental contributions to the development of methods that allow economists to estimate models of economic behaviour using data on individual decisions. He has also produced numerous important empirical results that advanced understanding of how government policies that regulate labour markets and influence educational opportunity affect economic inequality among individuals and groups.

Keywords

Additive regression models; Black–white labour market inequality in the United States; Childcare; Competing risks model; Consumer theory; Depreciation; Difference-in-difference estimators; Discrete and continuous decision variables; Duration models; Econometric Society; Econometrics; Employment spells; Fixed effects; General Educational Development (GED) test; Heckman, J; Heterogeneity; Human capital; Identification; Instrumental variables; Internal migration; Labour supply; Leisure; Life-cycle labour supply; Marginal treatment effects; Marginal utility of wealth; Matching; Maximum likelihood; Multiple regime models; Non-cognitive skills; Principle assumption; Programme evaluation; Roy model; Selection; Selection bias; State dependence; Training; Treatment effect; Two-step estimator; Uncertainty; Wage differentials

JEL Classifications

B31

Labour Supply and Selection

In Heckman's early work on labour supply we see at least three related contributions. First, he integrated consumer theory and the theory of labour supply. Second, he developed an empirical life-cycle setting for labour supply. Third, he provided an economically coherent framework for the statistical analysis of participation, labour supply and market wages. Heckman's work on labour supply, which originated in the early to mid-1970s, set the scene for the development of his research on selection, on labour market dynamics and on programme evaluation. They are all empirically oriented but with a keen eye on the identification and estimation of structural economic parameters from micro data.

Heckman's initial aim was to estimate the parameters of indifference curves for leisure and consumption. Given these, one could measure the welfare cost of some tax or welfare intervention and also simulate the impact of new policies. There were at least three key unresolved issues in the literature at that time: the econometric problem of non-participation in the presence of childcare costs; the need for a reasonably flexible functional form that could capture variation in hours worked among participants; and the lack of information concerning wage offers among those who do not participate. Heckman successfully addressed all of these issues in two remarkable papers (1974a, b). He recognized that a simple least squares analysis of hours of work, wages and participation would not, by itself, identify preference parameters and that the standard Tobit model alone was also insufficient to deal with the problem. As an alternative, Heckman developed an estimation procedure that allowed the work decision to be based on interrelated choices over hours of work and the use of formal childcare, each with its own separate source of stochastic variation. This approach is the forerunner of many microeconometric developments in this area and continues to set the standard by which models are judged. Indeed, Heckman's development of a likelihood that captures the sampling information on participation and wages can be seen as the beginning of the analysis of endogenously selected samples.

Yet the contributions of this work go beyond the insights concerning selection. His marginal rate of substitution specification for preferences turned out to be a highly innovative way of dealing with non-participation while allowing flexible but heterogeneous preferences, and the endogenous choice of formal and informal childcare jointly with hours of work and participation provided a basis for the analysis of multiple regime models.

In this 'static' labour supply analysis, we find repeated references to the potential importance of a more dynamic setting. In fact, Heckman conducted this work alongside his development of a life-cycle framework for labour supply. The origins of this work are in the first essay of his 1971 Princeton University doctoral thesis. Heckman began by noting that both income and consumption appear to follow a similar hump-shaped path over the life cycle that is out of line with the most basic consumption-smoothing model. Heckman (1974c) provides a beautifully simple, yet complete, integration of intertemporal consumption and labour supply theory and shows that a model with labour supply and uncertainty can easily explain these empirical phenomena.

Heckman extended this life-cycle analysis in two different directions. In Heckman (1976), he incorporated human capital investment and showed how earnings functions that ignored life-cycle labour supply tended to overestimate rates of depreciation. In other work, he developed an empirically implementable form of the intertemporal substitution model for labour supply. Here, he pointed out that given standard neo-classical assumptions the marginal utility of wealth is constant over time for an individual but differs across individuals and is clearly correlated with wages. Since labour supply choices could be written in terms of current wages and the marginal utility of wealth, Heckman's observation pointed to a perfect application of a fixed effects estimator for panel data, and in Heckman and MaCurdy (1980) he applied such an estimator to the panel data analysis of female labour supply. Heckman's model could also be adjusted to account for uncertainty and so became the prototypical intertemporal model of labour supply. It directly recovered the intertemporal substitution elasticity for labour supply and showed immediately the relationship of this intertemporal elasticity with the standard Hicksian and Marshallian elasticities, thereby tying together the 'static' and life-cycle approach to labour supply analysis.

Heckman's empirical investigations of individual labour supply behaviour stimulated further analysis of their statistical implications, and at least two major innovations in econometrics came out of this work. This work yielded new methods for analysing selected samples and also for estimating simultaneous multivariate choice models in which outcomes are a mixture of discrete and continuous decision variables. It is easy to see how important work on labour supply led to progress in these areas. Labour market participation is a choice based, in part, on wage offers that are observed only among those who participate, and household choices concerning participation, hours, and childcare present a mixture of both discrete and continuous decision variables. While the links to labour supply analysis are clear, the applicability of these two developments grew far beyond the study of labour supply.

Heckman's (1979) selection model is one of the most renowned econometric models since the mid-20th century. This work laid the foundation for the subsequent work on returns to training, the study of union wage differentials and to many other microeconometric problems. His approach was innovative but also simple. Starting with an additive regression model, Heckman noted that for normal distributions the conditional mean for a selected sample involves a single additional term that is itself a function of the selection probability. This term or 'control function' may be estimated in a first step from the choice probability model, and thus a computationally convenient two-step estimator is available for the analysis of selected samples.

Future work demonstrated that the selection model and the two-step estimator are more generally applicable in cases where the normality assumption fails. Heckman (1990) and others developed semiparametric extensions for the additively separable model. Heckman and Honoré (1990) derived the general nonparametric identification of the Roy model – a two-regime generalization of the additively separable selection model. Both Heckman and Sedlaceck (1985) and Heckman and Scheinkman (1987) provide empirical analyses of aggregate and sectoral wage distributions when individuals self-select into the labour market and into sectors of the economy.

If actions are a mixture of discrete and continuous decision variables that are simultaneously determined, then there will be a further condition on the econometric model to guarantee it provides a coherent statistical relationship between inputs and response. Heckman labelled this condition the 'principle assumption', and in Heckman (1978) he derives the conditions required for a coherent econometric framework. This work has influenced econometric work in industrial organization. Heckman's condition concerning a jump parameter in the mean of the latent variable underlying a discrete choice is easily mapped into analyses of entry decisions when fixed costs are present.

Panel Data and State Transitions

During the 1970s and 1980s economists gained much greater access to panel data-sets, and this development greatly shaped the research agenda in labour economics and other applied fields. Economists began to focus their attention on the sequence of decisions that firms and individuals make over time and attempted to model and understand the patterns of correlation over time in these decisions.

The current choices of individuals may be correlated with their past choices because of persistent unobserved individual differences (heterogeneity) or because current preferences or opportunities depend on past actions (state dependence). Thus, heterogeneity and state dependence can easily be confounded. Heckman (1981a, b) formally characterizes this identification problem, discusses pitfalls with simple or naive attempts at measuring one or the other effects, and shows how panel data can be used constructively to sort out competing explanations.

Heckman and Singer (1984) confront formally the statistical aspects of unobserved heterogeneity in the context of duration models. Duration models are used in studying job search, mortality, labour supply, marriage and other phenomena. Fully parametric models of duration include precise parameterizations of the unobservable (to an econometrician) attributes of individuals that influence the optimal duration of spells. Heckman and Singer show that these parameterizations can contaminate estimates of the economic parameters that pin down the structural relationship between spell duration and the rate of spell completions. As an alternative they propose, formally justify and implement a maximum likelihood estimator where the unobserved heterogeneity is modelled nonparametrically. Interestingly, in their application to data on unemployment spells, their nonparametric estimator chooses a small number of support points (individual types), and their Monte Carlo results in the same paper show that their estimates of the distribution of unobserved types are never close to the truth. However, their nonparametric treatment of unobserved heterogeneity allows them to choose points of support in a flexible way that avoids contaminating the estimated parameters of interest. The empirical results presented by Heckman and Singer strongly suggest that many of the somewhat puzzling results in the previous literature on the determinants of unemployment spells were the result of researchers trying to simultaneously estimate models with parametric assumptions concerning both the true underlying model of duration times and the distribution of unobserved individual heterogeneity.

The competing risks model has been used in many disciplines. In this model, observed outcomes reflect the minimum realized transition time over a discrete set of possible state transitions. Heckman and Flinn (1982) develop a structural interpretation of the competing risks model in the context of labour markets and apply it to study employment spells. They investigate the identification of the underlying economic parameters of interest as restrictions are removed from the auxiliary statistical specification. In a related paper, Heckman and Honoré (1989) explore identification of the competing risks model of failure times, and they show how introducing regressors into the competing risks model overturns previously established non-identification results. In the previous section, we mentioned Heckman and Honoré's (1990) work on requirements for identification is a generalized Roy model. Here, too, they demonstrated precisely how identification could be achieved through either restrictions on the shape of underlying skill distributions or sufficient variation in prices or exogenous regressors.

Estimation of Treatment Effects

In section "Labour Supply and Selection" we described how Heckman's early work on labour supply led to developments in the analysis of selected samples. Over the years, Heckman began to demonstrate that numerous problems other than labour supply are actually problems where missing data are the key challenge for empirical investigators. The work on identification in general versions of the Roy model is part of this research agenda. In the 1990s Heckman produced a series of related empirical and methodological papers that grew out of sustained research on methods of programme evaluation (see Heckman et al. 1997; see also Ichimura et al. 1998, and Clements et al. 1997). Heckman emphasized that the key impediment to measuring the return to any investment in training or education is the inability to see what those who receive treatment would have experienced in the absence of training.

Heckman's work in this area helped clarify several important points. First, Heckman produced much evidence consistent with the proposition that the treatment effects associated with various training and education programmes vary greatly among participants, even among those who are similar with respect to observed demographic characteristics. Second, outcomes from experiments involving random assignment to treatment and control status do permit straightforward estimates of the average gain from treatment in a given sample, but more structure is required in order to draw inferences concerning the distribution of gains and losses from treatment. Third, the performance of non-experimental methods such as matching or control function models can be greatly improved when researchers take care to balance treated and non-treated samples with respect to the probability of treatment. Careful attention to the support of this probability in each sample as well as its density can greatly improve the performance of non-experimental estimators. For example, Ichimura et al. (1997) demonstrate clearly that the performance of matching estimators improves when samples are re-weighted so that the density of the probability of treatment is the same among the treated and the untreated. Finally, difference-in-difference estimators may be an attractive strategy in situations where researchers have before and after data on treated and non-treated samples, but only in cases where there is evidence that the selection bias associated with treatment takes the form of a subject fixed effect that is constant over time.

Heckman has continued working in this area while devoting special attention to the mapping between various estimators in the literature and the precise set of counterfactual questions that they can address under various assumptions concerning how the data are generated. Heckman and Vytlacil (2005) present results that reflect the culmination of much of this research. In this paper, they explain how numerous estimators employed in the estimation of treatment effects may all be written as weighted averages of the marginal treatment effects (MTE) in the population. The marginal treatment effect is defined as the expected gain from treatment with the observed and unobserved determinants of participation held constant at particular values. Heckman and Vytlacil not only demonstrate how other estimators, such as instrumental variables, can be expressed in terms of the distribution of MTE, but they also demonstrate how MTE can be used as a building block in the construction of estimators that capture the expected treatment effects of specific policies on particular populations. This paper also clarifies an asymmetry in the way that heterogeneity enters models of selection and treatment. Agents may exhibit heterogeneity in terms of what they gain from treatment, but changes in their environment must affect each individual's likelihood of receiving treatment in a similar manner. Heckman goes on to spell out the challenges that researchers face if they wish to estimate models in which selection equations involve random coefficients.

Empirical Work on Inequality

To this point, we have focused almost exclusively on contributions that involve the development and implementation of new methods that seek to overcome some problem involving missing data. However, Heckman has also made noteworthy contributions that involve no methodological innovation but rather the use of simple methods to establish important facts about the sources of economic inequality and their relationship to government policy. His work in the areas of black–white inequality and the economics of education are ripe with examples of this type of work.

During the 1970s economists in the United States devoted considerable attention to the question of whether or not the Civil Rights Act had actually improved the economic well-being of blacks. Butler and Heckman (1978) made an important early contribution to this literature by pointing out that declining labour-force participation rates among less skilled blacks during the post-Civil Rights era could create the impression of economic progress among blacks even if none existed. To measure changes in the distribution of potential real wages facing blacks, researchers needed to address the fact that an increasing fraction of less skilled workers did not report market work and thus did not report wages in most surveys.

Butler and Heckman (1978) greatly influenced future work on black–white economic progress, but Heckman's most important contributions to this literature are summarized in a 1991 *Journal of Economic Literature* piece with John Donohue that cataloged evidence supporting the hypothesis that the Civil Rights Act of 1964 did serve as a catalyst for a discrete episode of black economic progress that was not simply a continuation of existing trends.

Donohue and Heckman (1991) acknowledge that long-term improvements in access to schools and school quality contributed to secular progress for blacks in the labour market in decades before and after the Civil Rights Act. However, they also show that federal government intervention served as an important catalyst for black progress during the civil rights era. They demonstrate 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. Further, they cite previous work by Heckman and his co-authors that demonstrates how the Civil Rights Act broke down de facto and de jure occupational segregation in the South (see Heckman and Payner 1989, and Butler et al. 1989. Donohue et al. 2002, show how private philanthropy and legal activism served as catalysts for improvements in black educational opportunity that pre-date the civil rights era).

Finally, they note a drastic decrease in the rate of net migration among blacks from the South to the North around 1965, a development that strongly suggests the Civil Rights Act did expand opportunity for blacks in the South.

Around the same time, Cameron and Heckman (1993) documented another set of important results concerning outcomes associated with a government programme. Cameron and Heckman demonstrated that persons who receive a high-school diploma by taking the General Educational Development (GED) test do not enjoy employment and earnings outcomes as adults that are in any way equivalent to those observed among high-school graduates. In fact, male GED recipients look quite similar to high-school dropouts with respect to many labour market outcomes.

Since the mid-1990s a growing literature has examined the cost and implied benefits of obtaining a GED for various types of individuals. Heckman and Rubinstein (2001) have taken this literature in a new direction by demonstrating that the difference between GED recipients and persons who finish high school takes the form of differences in non-cognitive skills related to self-esteem, work habits, and other personal traits. This discovery provides a natural explanation for the facts documented by Cameron and Heckman (1993). While the GED does certify that a young person has certain basic skills that are valued by employers, the combination of these skills and the failure to finish high school is a signal to employers that these youth are deficient in other areas.

Conclusion

Taken as a whole, the breadth and depth of Heckman's contributions to economics are stunning. His work on selection models, transition dynamics, and the estimation and identification of treatment effects has changed the way that economists analyse micro-data. At the same time, his empirical work in the economics of education and inequality produced numerous results that shape our understanding of modern labour markets and future research agendas for theorists and empiricists.

He has also shaped the profession as a teacher and advisor of students. During his career, Heckman has served as primary thesis adviser for scores of graduate students, and a significant number of his students have earned tenure in top economics departments, served as editors of journals, and helped produce yet another generation of scholars who take seriously the task of using economics to guide empirical investigations of important questions.

See Also

- ▶ Black–White Labour Market Inequality in the United States
- ▶ Competing Risks Model

Selected Works

1974a. (With O. Ashenfelter.) The estimation of income and substitution effects in a model of family labor supply. *Econometrica* 42: 73–86.

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

1974c. Life cycle consumption and labor supply: An explanation of the relationship between income and consumption over the life cycle. *American Economic Review* 64: 188–194.

1976. A life cycle model of earnings, learning and consumption. *Journal of Political Economy* 84(2, part II): S11–S44.

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

1978b. (With R. Butler.) The impact of the government on the labor market status of black Americans: A critical review. In *Equal rights and industrial relations*, ed. L. Hausman. Madison: Industrial Relations Research Association.

1979. Sample selection bias as a specification error. *Econometrica* 47: 153–161.

1980. (With T. MaCurdy.) A life cycle model of female labour supply. *Review of Economic Studies* 47: 47–74.

1981a. Heterogeneity and state dependence. In *Studies in Labor Markets*, ed. S. Rosen. Chicago: University of Chicago Press.

1981b. Statistical models for discrete panel data. In *Structural Analysis of Discrete Data with Econometric Applications*, eds. C. Manski and D. McFadden. Cambridge, MA: MIT Press.

1982. (With C. Flinn.) New methods for analyzing structural models of labor force dynamics. *Journal of Econometrics* 18: 115–168.

1984. (With B. Singer.) A method for minimizing the impact of distributional assumption in econometric models for duration data. *Econometrica* 52: 271–320.

1985. (With G. Sedlacek.) Heterogeneity, aggregation and market wage functions: An empirical model of self-selection in the labor market. *Journal of Political Economy* 93: 1077–1125.

1987. (With J. Scheinkman.) A method for minimizing the impact of distributional assumption of earnings. *Review of Economic Studies* 54: 243–255.

1989a. (With B. Payner.) Determining the impact of federal anti-discrimination policy on the economic status of blacks: A study of South Carolina. *American Economic Review* 79: 138–177.

1989b. (With R. Butler and B. Payner.) The impact of the economy and the state on the economic status of blacks: A study of South Carolina. In *Markets and Institutions*, ed. D. Galenson. Cambridge: Cambridge University Press.

1989c. (With B. Honoré.) The identifiability of the competing risks model. *Biometrika* 76: 325–330.

1990a. (With B. Honoré.) The empirical content of the Roy model. *Econometrica* 58: 1121–1149.

1990b. Varieties of selection bias. *American Economic Review* 80: 313–318.

1991. (With J. Donohue.) Continuous vs. episodic change: The impact of affirmative action and civil rights policy on the economic status of blacks. *Journal of Economic Literature* 29: 1603–1643.

1993. (With S. Cameron.) The nonequivalence of high school equivalents. *Journal of Labor Economics* 11: 1–47.

1997a. (With H. Ichimura and P. Todd.) Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies* 64: 605–654.

1997b. (With J. Smith, J. and N. Clements.) Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *Review of Economic Studies* 64: 487–535.

1998. (With H. Ichimura, J. Smith and P. Todd.) Characterizing selection bias using experimental data. *Econometrica* 66: 1017–1098.

2001. (With Y. Rubinstein.) The importance of noncognitive skills: Lessons from the GED testing program. *American Economic Review* 91: 145–149.

2002a. (With J. Donohue and P. Todd.) The schooling of southern blacks: The roles of legal activism and private philanthropy, 1910–1960. *Quarterly Journal of Economics* 117: 225–268.

2002b. (With Y. Rubinstein and J. Hsee.) The GED is a ‘mixed signal’: The effect of cognitive and noncognitive skills on human capital and labor market outcomes. Working paper, University of Chicago.

2005. (With E. Vytlacil.) Structural equations, treatment, effects and econometric policy evaluation. *Econometrica* 73: 669–738.

Heckscher, Eli Filip (1879–1952)

Carl G. Uhr

Keywords

Economic history; Heckscher, E. F.; Heckscher-Ohlin trade theory; Imperfect competition; Pure public goods

JEL Classifications

B31

Born into a Jewish family in Stockholm, Heckscher studied history under Hjärne and economics under Davidson at Uppsala University from 1897. In 1907 he became a docent at Stockholm University College of Commerce, and from 1909 to 1929 he was professor of economics and statistics. Then, because of his great research productivity, the college authorities changed his position to research professor, lightened his teaching duties and made him director of the newly established Institute of Economic History. Heckscher continued in this position until he retired in 1945. He succeeded in establishing economic history as a subject of graduate study in Sweden's universities.

In 1950, the Ekonomisk-historiska Institutet, Stockholm, through Bonniers Co., published the *Eli F Heckscher bibliografi 1897–1979* (123 pp.). It contains 1148 entries for his 36 books, 174 articles in professional journals, his chapters in government reports, and the more than 700 short articles he wrote for the weekend issues of Stockholm's leading newspapers. Only a few of his books and articles have been translated and will be referred to by their English titles; other works will be mentioned only by the English translation of their original titles and identified by their numbers as entries in the Heckscher bibliography.

By 1929, when he was able to specialize in economic history, Heckscher had already written a dozen books on such diverse subjects as *Economic Principles* (1910, No. 158), *The Continental System* (1918, No. 443, later republished, Oxford, 1922) and *Economics and History* (1922, No. 478). As a result of his teaching, his contributions to economics are a blend of innovations in economic theory and a new methodology for economic history research, an approach to quantitative research very different from that used by leaders in his field such as Schmoller, Cunningham and Sombart.

Heckscher's most significant contributions to economic theory may be found in two articles. 'Effects of Foreign Trade on Distribution of Income' (1919) is the origin of the modern Heckscher-Ohlin factor proportions theory of international trade, developed further in Ohlin (1933).

'Intermittently Free Goods' (1924) presents a theory of imperfect competition nine years ahead of that by Joan Robinson and Edward Chamberlin, and a discussion of collective goods not priced by the market. Heckscher observed that significant new products are introduced by firms with investment in plants that have a capacity which far exceeds initial demand for their products. The latter are sold at prices which barely cover unit variable costs, and so, for a time, the services of the fixed investment are provided as 'free goods'. Then, as weaker firms are eliminated, demand shifts to the remaining larger firms who use up their production capacity. By and by these firms expand, enjoy economies of scale, differentiate their products and become prosperous oligopolies dividing a mass market into more or less definite shares.

A situation of the opposite kind arises when the smallest feasible production facility has a production capacity which suffices for a growing and indefinitely large demand without affecting the costs and service life of the production unit. This is the case with many so called 'pure' public or collective goods. Heckscher used street illumination as an example of a collective good, which can be used simultaneously by few or many persons, a service that cannot be priced per unit of individual use. The costs of providing this service, then, are usually met by an increase in local government taxes. In that case, and in contrast to that of intermittently free goods, the citizens pay the full-cost price of the service from the outset in their current taxes. Then, as activities in and use of lighted streets increase over time, the citizens derive increased utility per tax dollar spent for street lighting.

At the Institute of Economic History Heckscher's first work, one of his major and most widely known treatises, was *Mercantilism* (1931). His other major work, the fruit of many years of pioneering research devoted to his own country, was *Sweden's Economic History from the Reign of Gustav Vasa* (vols 1 and 2, 1935–6, No. 878; vols 3 and 4, 1950, No. 1146). He also wrote a popular version of this work, *Life and Work in Sweden from Medieval Times to the Present* (1941), No. 1014, republished as An Economic History of Sweden (1954). Among his other books of particular interest are Materialist and Other Interpretations of History (1944, No. 1052), Industrialism, Its Development from

1750 to 1914 (1946, No. 1123) and *Studies of Economic History* (1936, No. 918). It was in this work he presented a new methodology he proposed for economic history research in his essay 'The Aspects of Economic History', pp. 9–69. This was reinforced in his articles, 'A Plea for Theory in Economic History' (1929) and 'Quantitative Measures in Economic History' (1939).

For the analysis of any epoch of economic history – as distinct from factual description of a chronologically arranged body of heterogeneous source materials – Heckscher proposed consideration of a succession of its 'economic aspects', to introduce order and inject economic theory into the interpretation of that epoch. Unlike 'periods', 'aspects' are *not necessarily* time dependent. They are theoretical and imply hypotheses that are, within limits, testable against the given data. A series of aspects, for instance of (*a*) the exchange processes; (*b*) natural resources and technologies; (*c*) labour force and capital; (*d*) forms of enterprise organization; and (*e*) extent and composition of demand, form an economic model of the epoch. This done, the function of the economic historian is to provide a synthetic overview and explanation of the relations between the aspects of the model.

Thus Heckscher bridged the gap between economic history and theory by addressing broad questions or hypotheses to the source materials for intensive and critical study. He always preferred to present his finding supported by statistical data. That done, he was not satisfied until he had explained and illuminated these by economic analysis, that is, by applying cognate principles of economic theory to their interpretation.

See Also

- ▶ Heckscher–Ohlin Trade Theory
- ▶ Stockholm School

Selected Works

1918. *Kontinentalsystemet.* Stockholm: P.A. Norstedt & Söner. Trans. as *The continental system,* ed. H. Westergaard, Oxford: Clarendon Press, 1922.

1919. Effects of foreign trade on distribution of income. *Ekonomisk Tidskrift.* Reprinted in AEA, *Readings in the theory of international trade,* ed. H.S. Ellis and L. Metzler, Philadelphia: Blakiston, 1949.

1924. Intermittently free goods. *Ekonomisk Tidskrift* 553. Reprinted in German in *Ein Beitrag zur Sozialwissenschaft und Sozialpolitik* 59 (1928).

1929. A plea for theory in economic history. *Economic Journal,* Historical Supplement No. 4: 523–54.

1931. *Merkatilismen,* 2 vols. Stockholm: P.A. Norstedt & Söner. Authorized trans. by M.Shapiro as *Mercantilism,* London: G. Allen & Unwin, 1935. Revised edn., New York: Macmillan; London: G. Allen & Unwin, 1955.

1935–6, 1949. *Sveriges Ekonomiska Historia från Gustav Vasa.* Vols. 1 and 2, 1935–6; vols. 3 and 4, 1949, Stockholm: Bonnier.

1939. Quantitative measures in economic history. *Quarterly Journal of Economics* 53: 167–93.

1941. *Svenskt arbete och liv fran medeltiden till nutiden.* Stockholm. Trans. G. Ohlin as *An Economic History of Sweden,* Cambridge, MA: Harvard University Press, 1954.

1950. *Eli F. Heckscher bibliograft 1897–1949.* Stockholm: Bonnier.

Bibliography

Ohlin, B. 1933. *Interregional and international trade.* Cambridge, MA: Harvard University Press.

Heckscher–Ohlin Trade Theory

Ronald W. Jones

Abstract

Heckscher–Ohlin trade theory consists of four principal theorems, viz. the Heckscher–Ohlin trade theorem whereby relatively capital-abundant countries export relatively capital-intensive commodities, the factor-price equalization theorem whereby trade in goods may serve to equalize wage rates between countries, the Stolper–Samuelson theorem whereby an increase in the price of the relatively labour-intensive commodity unambiguously improves the real wage rate, and the Rybczynski theorem stating that an increase in capital endowment by itself must cause some output to fall if prices are held constant. The article discusses the nature and fate of these theorems.

Keywords

Autarky; Comparative advantage; Dimensionality; Distribution of income; Factor- intensity reversal; Factor-price equalization theorem; Free trade; General equilibrium; Heckscher–Ohlin theorem; Hecksher–Ohlin trade theory; International trade theory; Labour–capital ratio; Leontief paradox; Metzler tariff paradox; New trade theory; Reciprocity relationship; Relative factor abundance; Relative factor intensity; Rybczynski theorem; Specialization; Stolper–Samuelson theorem; Tariffs; Terms of trade

JEL Classifications

F1

Eli Heckscher (1919) and Bertil Ohlin (1933) laid the groundwork for substantial developments in the theory of international trade by focusing on the relationships between the composition of countries' factor endowments and commodity trade patterns as well as the consequences of free trade for the functional distribution of income within countries. From the outset general equilibrium forms of analysis were utilized in these developments, which gradually came to be sorted out into four 'core propositions' (Ethier 1974) in the pure theory of international trade.

The Four Theorems

Although all four of the propositions to be discussed are an outgrowth of the seminal work of Heckscher and Ohlin, only one of these propositions bears their name explicitly. The *Heckscher–Ohlin theorem* states that countries export those commodities which require, for their production, relatively intensive use of those productive factors found locally in relative abundance. The twin concepts of relative factor intensity and relative factor abundance are most easily defined in the small dimensional context in which the basic theory is usually developed. Two countries are engaged in free trade with each producing the same pair of commodities in a purely competitive setting, supported by constant returns to scale technology that is shared by both countries. Each commodity is produced separately with inputs of two factors of production that, in each country, are supplied perfectly inelastically. (For a throrough analysis of having endowments respond endogenously, see Findlay 1995). Following the Ricardian distinction, commodities are freely traded but productive factors are internationally immobile.

Although one country may possess a larger endowment of each factor than another, the presumed absence of returns to scale guarantees that only relative factor endowments are important. The home country is said to be relatively labour abundant if the ratio of its endowment of labour to that (say) of capital exceeds the corresponding proportion abroad. This is known as the physical version of relative factor abundance. An alternative involves a comparison of autarky relative factor prices in the two countries: the home country can be defined to be relatively labour- abundant if its wage rate (compared with capital rentals) is lower before trade than is the foreign wage (relative to foreign capital rentals). Since autarky factor prices are determined by demand as well as supply conditions, these two versions need not correspond. In particular, if the home country is, in the physical sense, relatively labour abundant it might nonetheless have its autarky wage rate relatively high if taste patterns at home are strongly biased towards the labour-intensive commodity compared with tastes abroad. In such a case the trade pattern reflects the autarky factor–price comparison: the home country exports the physically capital-intensive commodity. As discussed below, the link between commodity price ratios (the proximate determinant of trade flows) and factor price ratios is more direct than that between commodity price ratios and physical factor endowments. Thus the Heckscher–Ohlin theorem is more likely to hold if relative factor abundance is defined in terms of relative factor prices prevailing before trade. The procedure typically followed in the literature is to assume that both countries share identical and homothetic taste patterns. Such an assumption, in conjunction with the presumed identity of technology at home and abroad (with an even stronger version of homotheticity–linear homogeneity) helps to isolate the separate influence of physical factor supplies and makes the validity of the Heckscher–Ohlin theorem with the physical definition of factor abundance as likely as with the autarky factor price definition.

These assumptions are less than sufficient to guarantee the Heckscher–Ohlin theorem, even in the simple context of two-country, two-factor, two-commodity trade. The potential stumbling block is the fact that even though countries share the same technology, the commodity that is produced by relatively labour-intensive techniques at home may be produced by relatively capital-intensive techniques abroad. This is the phenomenon of factor-intensity reversal. If production processes are independent of each other, there is nothing (other than bald assumption) to rule out its appearance. The bald assumption would assert that regardless of factor endowments one industry always employs a relatively higher ratio of labour to capital than does the other industry, where techniques are chosen with reference to the wage/rental ratio common to both industries. If this is not the case, and if the commodity that is relatively labour-intensive at home is produced by relatively capital-intensive techniques abroad, the phrasing of the Heckscher–Ohlin theorem that explicitly states ‘each country exports the commodity that is produced in that country making relatively intensive use of the factor found in relative abundance in that country’ is fatally flawed. The reason? If the relatively labour-abundant country exports its labour-intensive commodity, it must do so in exchange for the commodity that, in the relatively capital-abundant foreign country, is produced by labour- intensive techniques. Thus if one country satisfies the theorem, the other country cannot (Jones 1956).

In the event of factor-intensity reversal, it must be the case that, whatever the commodity exported by the labour-abundant home country, the ratio of labour to capital employed in its production must exceed the labour/capital intensity adopted in foreign exports. However, this observation is of little value if one wishes to infer from an intensity comparison between exportables and import-competing goods within a given country whether that country is more labour abundant than some foreign country. Such an inference lay behind the celebrated study of Leontief (1953) on United States trade patterns. This research, the conclusions of which came to be known as the Leontief paradox (American exportables are produced by more labour- intensive techniques than are import-competing goods) provided the major stimulus to developing and defining the meaning and conditions supporting the Heckscher–Ohlin theorem.

Earlier work in Heckscher–Ohlin trade models was focused on the pricing relationships embodied in Heckscher–Ohlin theory. Ohlin (1933) stressed the effect which free trade would tend to have on the distribution of income within countries, viz. relative factor prices would move in the direction of equality between trading countries which share the same technology. Ohlin’s mentor, Heckscher, went even further in his pioneering 1919 article. Absolute factor-price equalization was purported to be ‘an inescapable consequence of trade’ (For recent appraisals of each of these economists see Jones 2002, 2006a). Nonetheless, Ohlin’s view of partial equalization seems to have dominated, with the exception of Lerner’s unpublished 1933 manuscript (which surfaced after Samuelson’s articles), until the statement of the factor-price equalization theorem in articles by Samuelson in 1948 and 1949. Rejecting his earlier tacit acceptance of the Ohlin thesis of partial equalization (in the Stolper–Samuelson article, which appeared in 1941), Samuelson proved that within the traditional confines of the 2 × 2 × 2 model (with no factor-intensity reversals and each country incompletely specialized), free trade would drive wage rates to absolute equality in the two countries (and, as well, would equate returns to capital) despite the assumption that labour (and capital) are assumed to be immobile between countries.

The logic of the argument for the simple 2 × 2 case can be stated briefly. In a competitive equilibrium unit cost equals price if the commodity is produced. Thus let A represent the matrix of input–output coefficients, aij, w the vector (pair) of factor prices, and p the vector (pair) of commodity prices. Techniques need not be constant; in general they depend upon prevailing factor prices so that $A = A(w)$. Therefore the competitive profit conditions if both goods are actually produced dictate that:

$$A(w).w = p. \tag{1}$$

If we assume no factor-intensity reversals, $A(w)$ is non-singular. Therefore if countries share the same technology and face the same pair of free-trade commodity prices, they must face exactly the same set of factor prices if each country produces both goods.

This approach may suggest that the crucial issue in the factor-price equalization argument is the unique dependence of factor price vector w on commodity price vector p, and an extensive literature has developed which focuses on this issue. In the 2 × 2 case uniqueness is a simple question – it depends on factor intensities differing between sectors and not reversing. But from the outset Samuelson pointed out that this was not the only issue. The question of uniqueness involves properties of technology alone, whereas under appropriate circumstances two countries in free trade will have factor prices equalized only if factor endowments are reasonably similar. For, if factor endowments are too dissimilar, it will be impossible for both countries to produce both commodities, in which case the equalities in (1) cannot universally hold.

These ideas can be made more precise by considering a concept due to McKenzie (1955), which Chipman (1966) called the 'cone of diversification'. For any factor price vector, w, there is determined a pair of techniques (labour/capital ratios) for the two commodities. Both factors can be fully employed only if the country's endowment vector is contained within the cone spanned by these techniques. Suppose two countries face a common free-trade commodity price vector, p, and that the commonly shared technology associates a unique factor price w corresponding to this p. Then if the endowment vectors of both countries lie within the cone of diversification, their factor prices must be equalized (McKenzie 1955).

Some seven years prior to Samuelson's first factor-price equalization essay there appeared the article by Stolper and Samuelson (1941), which must be ranked a classic not only for its discussion of what became known as the *Stolper–Samuelson theorem*, but because it is one of the first concrete developments of the ideas of Heckscher and Ohlin in the explicit format of a two-factor, two-commodity, general equilibrium model (This theorem became so widely cited that on the golden anniversary of its appearance a conference was held at Stolper's university, the University of Michigan. See Alan Deardorff and Robert Stern 1994). Their argument supposedly concerns the effect of protection on real wages, and in the course of the argument they assume that a tariff does not change the terms of trade so that locally import prices rise. Subsequently, in what has become known as the 'Metzler tariff paradox', Metzler (1949) showed that with sufficiently inelastic demand a tariff might so improve a country's terms of trade that the relative domestic price of imports falls. If so, the Stolper–Samuelson contention that a tariff yields an increase in the real return to a country's relatively scarce factor would be reversed. However, it is now commonly agreed that the Stolper–Samuelson theorem refers to the general phenomenon whereby an increase in the relative domestic price of a commodity (whether brought about by a tariff increase, decrease, or some other reason) must unambiguously raise the real return to the factor of production used relatively intensively in the production of that commodity.

Introducing the production-box diagram technique (for a single country), Stolper and Samuelson illustrate how an increase in the relative price of labour-intensive watches attracts resources from capital-intensive wheat. To clear factor markets, both sectors must then use labour more sparingly. That is, the ratio of capital to labour utilized in each sector rises, which implies an unambiguous increase in labour's marginal productivity measured either in watches or in wheat. Thus regardless of workers' taste pattern, protection has increased the real wage.

The logic of the Stolper–Samuelson argument rests heavily upon the presumed absence of joint production. It takes labour and capital to produce watches, and, in a separate activity, a higher capital/labour ratio is used to produce wheat. In competitive settings any change in a commodity's price must reflect an average of factor price changes so that unit costs change as much as do prices. Therefore one factor price must rise relatively more than either commodity price. Which factor gains depends only upon the factor-intensity ranking. If the price of watches rises, and that of wheat does not, the wage rate must increase by relatively more. And this result follows even if techniques are frozen so that no resources can be transferred between sectors (as they can be in the Stolper–Samuelson discussion) and marginal products are not well defined (Jones 1965).

To round out the quartet of theorems, the Rybczynski theorem (1955) deals with the same model but focuses on the relationship between factor endowments and commodity outputs. Suppose commodity prices are kept fixed in the 2 × 2 setting and an economy is incompletely specialized. Then by the factor-price equalization theorem, factor prices are determined and fixed as well, which implies also that techniques of production remain constant. If the economy's endowment of one factor increases, while its endowment of the other factor remains constant, the economy must in some sense grow (the transformation schedule shifts out). However, this growth is strongly asymmetric: one output actually falls. The factor-intensity ranking selects the loser – the commodity that uses intensively the factor that is fixed in overall supply must decline. The reasoning is simple. As one factor expands, it must be absorbed in producing the commodity using it intensively. But with techniques frozen (since prices are assumed fixed), the expanding sector must be supplied with doses of the non-expanding factor as well. The only source for this factor is the other industry that must, perforce, contract.

Relationships Among the Theorems

All four propositions are based on the same 'mini-Walrasian' general equilibrium model of trade and there are some interesting relationships and distinctions among them. Perhaps most importantly, both the Heckscher–Ohlin theorem and the factor-price equalization theorem refer explicitly to a comparison between (two) countries, whereas the Stolper–Samuelson and Rybczynski propositions are involved with relationships within a single country. This distinction implies that the assumption that countries share an identical technology is not necessary for the latter two propositions. Thus, for example, a country could protect the factor used intensively in its import-competing sector in real terms (according to Stolper and Samuelson) regardless of the level or type of technology adopted by other countries.

The factor-price equalization theorem is a razor's-edge type of result. Should the technology available to two countries differ only slightly, any presumption of exact factor-price equalization in the absence of explicit international factor markets disappears. The Heckscher–Ohlin theorem is a little more robust in this regard. In general, trade patterns depend on all those variables that influence prices: tastes, technology, and factor endowments (not to mention taxes or other distortions). If tastes are identical (and homothetic) but factor endowments are not, the latter difference will tend to dominate the trading pattern even if technologies differ as long as this difference is 'less important'. At issue is a weighing of endowment differences with the Ricardian emphasis on technology differences. A particular variation of the factor-price equalization theorem is more general, and does not need to assume that technologies are identical between countries. It concerns the *dependence* of factor prices only upon commodity prices. It follows as long as the country produces both commodities.

Two versions of the Heckscher–Ohlin theorem have been cited, depending on which definition of relative factor abundance is selected. If the physical factor intensity ranking is chosen as the criterion, the basis for the Heckscher–Ohlin theorem resides in the kind of link between endowment patterns and outputs for a single economy exemplified by the Rybczynski theorem. An extension of this theorem allows a comparison of the transformation schedules for two economies with similar technologies. The relatively labour abundant (physical definition) country will produce relatively more of the labour-intensive commodity at common commodity prices (Jones 1956). Therefore, unless taste differences are sufficiently biased to counter this effect, the labour-intensive good will, in autarky, be cheaper in the labour-abundant country and, with trade, will be exported. The Stolper–Samuelson theorem is closely linked to the alternative form of the Heckscher–Ohlin theorem. Suppose there are no factor-intensity reversals. Then if both goods are produced there is a monotonic relationship between the wage/rent ratio and the relative price of the labour-intensive good such that a rise in the latter is associated with a greater than proportionate increase in the former. Thus the relatively low wage country must, in autarky, have been the relatively cheap producer of the labour-intensive commodity. As mentioned earlier, no caveat must be added about tastes, since these are already incorporated in the autarky factor-price comparison.

Although a comparison of factor endowments between countries is crucial in considering both the Heckscher–Ohlin theorem and the factor-price equalization theorem, such a comparison works in opposite directions for these two propositions. Thus if factor endowment proportions are sufficiently *dissimilar*, trade patterns suggested by the Heckscher–Ohlin theorem *must* hold (aside from the possibility of factor-intensity reversals) whereas free trade *cannot* bring about factor-price equalization. Sufficiently different factor endowments entail one country's transformation schedule being everywhere flatter than the other country's. At least one country must be specialized with trade. By contrast, the factor-price equalization result holds if factor endowments are similar enough so that international differences in the composition of outputs are capable of absorbing these endowment differences at the same set of techniques (and factor prices). If endowments are this close, it would always be possible for demand differences to be so biased that the physically labour-abundant country exports the capital-intensive commodity. Indeed, if such a demand reversal of the Heckscher–Ohlin theorem takes place, free trade must result in factor-price equalization (Minabe 1966).

Samuelson's name occurs so frequently in the literature on Heckscher–Ohlin trade theory that it is often appended to the other two names. One of his results not cited heretofore is the reciprocity relationship (Samuelson 1953). This states that in any general equilibrium model the effect of an increase in a commodity price (say p_j) on a factor return (say w_i) is the same as the effect of an increase in the corresponding factor endowment (V_i) on the output of commodity j. Of course, in each case some other set of variables is being held constant. Thus:

$$\frac{\partial w_i}{\partial p_j} = \frac{\partial x_j}{\partial V_i} \tag{2}$$

with all other commodity prices and all endowments held constant in the left-hand derivative and all other endowments and all commodity prices held constant in the right-hand derivative. This relationship is easy to prove (see, for example, Jones and Jose Scheinkman 1977). It also reveals the *dual* nature of the Stolper–Samuelson and Rybczynski theorems. If an increase in the price of watches lowers capital returns, then an increase in the endowment of capital (at constant prices) would lower the output of watches. In each case it is the presumed labour-intensity of watches that is operative.

In the 2 × 2 setting both the Stolper–Samuelson and Rybczynski theorems reflect the 'magnification effects' (Jones 1965) that stem directly from the assumed lack of joint production. With a '^' over a variable designating relative changes, if watches are labour intensive and wheat capital intensive and if the relative price of watches rises,

$$\widehat{w} > \widehat{p}_{wa} > \widehat{p}_{wh} > \widehat{r}. \tag{3}$$

In addition, should an economy grow, but with labour (L) growing more rapidly than capital (K),

$$\widehat{x}_{wa} > \widehat{L} > \widehat{K} > \widehat{x}_{wh}. \tag{4}$$

Inequality ranking (3) shows commodity price changes trapped between factor-price changes (since two factors are required to make a single good), while inequality (4) shows that in order to absorb endowment changes, the composition of outputs (each of which uses both factors) must change more drastically. Stolper and Samuelson stressed the first inequality in (3), while Rybczynski focused on the last inequality in (4), assuming $\widehat{K}$ equals zero.

Higher Dimensions

International trade theory generally, and Heckscher–Ohlin trade theory in particular, has frequently been criticized for its restriction to the low dimensionality represented by two commodities, two factors, and two countries. In fairness to both Heckscher and Ohlin it should be stressed that their discussions typically were not so confined. But neither were their conclusions as precise as those subsequently developed by Samuelson and others in the 2 × 2 × 2 versions of the four core propositions. And in the years following Samuelson's pioneering work on factor-price equalization, scores of articles have indeed appeared dedicated to the question of robustness of these results in higher-dimensional contexts. A highly detailed discussion of the issue of dimensionality is provided in Ethier (1984), and an earlier critique of the limitations imposed by small numbers of goods and factors is found in Ethier (1974) and Jones and Scheinkman (1977).

Part of the difficulty embedded in the move to higher dimensions lies in the ambiguity involved in what the propositions should state for cases beyond 2 × 2. The one proposition for which this is not the case is the factor-price equalization theorem. Consider the case of equal numbers of factors and produced commodities, with all goods traded and factors immobile internationally. The uniqueness of a factor price vector, w, corresponding to a given commodity price vector, p, is not guaranteed even in the 2 × 2 case; a factor-intensity reversal could lead to two (or more) values of w consistent with a given p. For the $n \times n$ case Gale and Nikaido (1965) provided conditions sufficient to guarantee global univalence of the factor price–commodity price relationship: the $A(w)$ matrix of input–output coefficients should be a 'P- matrix', that is a matrix with all positive principal minors. These conditions have been slightly weakened by Andreu Mas-Colell (1979), and earlier a fundamental interpretation of the conditions was supplied by Yasuo Uekawa (1971). It remains the case, however, that this condition on technology alone is somewhat remote from the issue of factor-price equalization. Just as in the 2 × 2 case, two countries sharing a common technology and each capable of producing the same set of n commodities (at the same traded-goods prices) with n productive factors, will, if techniques of production are the same, have their factor prices driven to equality if their factor endowments are sufficiently close. The concept of the 'cone of diversification' within which both endowment vectors must lie for factor-price equalization is as meaningful and relevant in n dimensions as it is in two.

Although the factor-price equalization theorem has an unambiguous meaning in higher dimensions, it is a theorem that cannot be expected to hold if the number of productive factors exceeds the number of freely traded commodities. The reasoning is basic, and can be linked to Eq. (1). These competitive profit conditions supply n links between factor prices and traded commodity prices, where n is the number of traded commodities. If r, the number of factors, should exceed n, the relationships in Eq. (1) are insufficient in number to provide a solution for the vector w for given p. Other conditions are required, and these are provided by the full employment conditions, one for each productive factor. Thus a nation's endowment bundle, V, becomes a determining variable affecting factor prices that is additional to the commodity price vector, p. For example, in the simple three-factor, two-commodity 'specific-factor' model (Jones 1971; Samuelson 1971), suppose a country faces a given world price vector, p, and experiences a slight increase in its endowment of a factor 'specific' in its use in the first industry. The intensity with which factors are utilized depends upon factor prices, and if these do not change, there is no way in which outputs can adjust to clear all factor markets. The return to the factor specific to the first industry must fall so as to encourage the further use of that factor. Two countries of this type with different endowments will generally have different sets of factor prices with trade, even if they share a common technology. It may be interesting to note that Heckscher's (1919) discussion of the necessity of factor-price equalization is focused on a three-factor, two-commodity numerical illustration (Jones 2006a). As just suggested, such a 3 × 2 setting in general does *not* lead to factor-price equalization.

The Heckscher–Ohlin model with two factors but many commodities available in world markets provides a useful scenario in which to re-examine the Heckscher–Ohlin theorem concerning the pattern of trade. The strong influence of factor endowments on production and trading patterns is revealed by considering two countries sharing the same technology but with different endowment ratios. Suppose commodity prices for traded goods are determined in a world market composed of a number of different countries with potentially a wide variation in technologies. Given world prices, any pair of countries with the same technologies shares a Hicksian composite unit-value isoquant for all traded goods (Jones 1974), made up of strictly bowed-in sections (where only one commodity is produced) alternating with flats (where a pair of commodities is produced). Regardless of the number of commodities, each country engaged in trade need produce only one or two (in the two-factor case), and these commodities will be the ones requiring factors in proportions close to that country's endowment ratio (Not explicitly considered here is that the activity of exporting a commodity may require factor proportions different from those required in production; see Jones et al. 1999). In this setting the spirit of the Heckscher–Ohlin theorem is that each country concentrates its resources on a small range of commodities whose factor requirements mirror closely that country's endowment base; the country exports some or all commodities in this set and imports commodities that are more labour-intensive than these goods as well as those that are more capital-intensive. Two countries whose endowments are fairly similar *may* produce the same pair of goods and thus achieve factor-price equalization with trade. Countries further apart in endowment composition will have disparate sets of factor prices and may produce completely different bundles of commodities (see also Krueger 1977).

With many factors and many commodities a different approach can be taken. The ability of autarky commodity price comparisons to predict trade patterns item by item is severely questioned, so that little hope remains of linking endowment differences to the detailed composition of trade. But statements about aggregates or 'correlations' between trade patterns and autarky prices can be made (Deardorff 1980; Dixit and Norman 1980). A nation's net imports, M, are positively correlated with the comparison of its autarky commodity price vector, p^A, and the vector of free-trade commodity prices, p^T. Thus:

$$\left(p^A - p^T\right)M \geq 0 \tag{5}$$

(see Ethier 1984, p. 139). This idea can be extended to the further relationship between autarky commodity prices and the vector of autarky factor prices (as in (1)) to establish that countries possess a comparative advantage, on average, in commodities using intensively factors that are relatively cheap in autarky (See Deardorff 1982, and Ethier 1984, for more details).

The reciprocity relationship expressed in (2) is quite general in terms of dimensionality and thus serves to link the Rybczynski theorem in a dual relationship to the Stolper–Samuelson theorem. However, when the number of factors exceeds the number of produced commodities, differences between the two types of theorems do appear. This basic asymmetry is linked to the failure of the factor-price equalization theorem when factors exceed commodities in number.

Major efforts have been made to generalize the Stolper-Samuelson and Rybczynski theorems from the 2 × 2 settings to the $n \times n$ setting, and a pair of earlier efforts met with only limited success. Murray Kemp and Leon Wegge (1969) searched for conditions on the original activity matrix, A, or distributive share matrix, θ, that would be sufficient to ensure what is known as the *strong* form of the Stolper–Samuelson theorem: Each factor is associated with a unique commodity such that if that commodity price (alone) increases, the return to the associated factor increases by a relatively greater extent and all other factor returns fall. The conditions they tried are stated by the inequalities in (6):

$$\theta_{ii}/\theta_{ki} > \theta_{ij}/\theta_{kj} \quad \text{for all } i, j \neq i, \text{and } k \neq i. \tag{6}$$

For each factor, i, the ratio of its distributive share in the industry positively associated with that factor (i) to that of any other factor's share in industry i, exceeds the corresponding ratio of these two factors in any other industry. These strong conditions do indeed lead to the desired strong result on factor returns (that is the inverse of the distributive share matrix has positive diagonal terms, greater than unity, and negative off-diagonal elements) for the 3 × 3 case. However, the authors provided a counter-example for the 4 × 4 case and that was that. Even stronger conditions for sufficiency are required, and these were supplied by Jones et al. (1993). These conditions are, in a sense, suggested by the statement of the theorem that for any price change all factor returns except one must fall. That is, they must have a relatively similar fate, suggesting fairly similar intensity use. The inequality that suffices is shown in (7):

$$\theta_{ii}/\theta_{ki} - \theta_{ij}/\theta_{kj} > \Sigma_{s \neq k, i, j} |\theta_{si}/\theta_{ki} - \theta_{sj}/\theta_{kj}| \text{ for all } i; j, k \neq i. \tag{7}$$

That is, condition (6) is not strong enough; the difference between the two terms in (6) must exceed the absolute value of similar differences in all the unintensive factors (whose returns all must fall). As occasionally happens, an article by John Chipman (1969) in the same issue of the same journal provided a condition sufficient for a weaker result, namely that the elements along the diagonal all be positive and exceed unity, regardless of signs off the diagonal. His condition met the same fate – sufficient for the 3 × 3 case but not higher. Mitra and Jones (1999) provided a sufficient condition for the $n \times n$ case.

It is possible to argue that these conditions are so strong as to suggest the Stolper–Samuelson and Rybczynski theorems really do not generalize. However, there is a form of the Stolper–Samuelson theorem that does generalize to higher dimensions with relatively little structure and, arguably, captures the essence of the original 1941 result. Stolper and Samuelson addressed the question of a particular government policy on real wages – the imposition of a tariff. But consider a more general question. Suppose an arbitrary factor of production seeks government aid sufficient to have its *real return* improved in a *non-transparent* fashion, that is, without a direct payment (out of tariff or any other source of government revenue). It is to be done by changes in taxation or government demand that would affect commodity prices. What would be required? Very little, as shown in Jones (1985, 2006b). Suppose there is little or no joint production (to be discussed below), and that there is a sufficient number of commodities (at least equal to the number of factors). These conditions suffice to ensure that there exists a subset of commodities that, if their prices are raised by the same relative amount with no other commodity price changes, the real return to the particular factor is guaranteed to increase. This result should have pride of place in the field of political economy and represents a significant generalization of the original Stolper–Samuelson result. The kind of detailed requirements shown by (6) and (7) is not necessary as long as a single commodity is not by itself required to do the job and as long as (along with Chipman 1969) it is not required that all other factors lose. Indeed, the favoured factor might well appreciate not standing nakedly as the only winner.

There are a few special cases of the $n \times n$ Heckscher–Ohlin setting that deserve mention. The most important might be the contribution of Roy Ruffin (1988). Ruffin redefines the Ricardian setting in which each country has a distinct labour force whose productivities in producing a number of goods differ from those in other countries. Instead of having each type of labour restricted to a single country, Ruffin suggests letting each country be populated by a wide variety of labour types, with the relative supplies of each type differing from country to country. This shifts the focus to relative endowment differences among countries, with the same technologies (a single type of labour is the same no matter where located). The key feature of such a model is that there is not only no joint production of outputs, there is no need for any single factor to have to work jointly with any others to produce commodities. As a consequence, factor prices are always equalized by free trade in commodities. Furthermore, each country's transformation surface looks just like that of a *world* transformation surface in the Ricardian model. In the two-commodity case this is a broken, bowed out, join of the two linear schedules for each country. In higher dimensions there are various dimensional 'facets' down to those of zero-dimension, that is, points at which each type of labour is fully employed in a different commodity. Except for the relative size of these facets, each country's transformation curve looks like that of any other country. At given commodity prices the common 'price plane' is 'tangent' to each surface such that each labour type is assigned to the same commodities in any country. At free-trade prices the relative production pattern in any country exactly mirrors the relative labour supplies and productivity of labour in that country.

Another special version, one that does give the Kemp–Wegge strong results, is the 'produced mobile factor' structure introduced by Jones and Marjit (1985, 1991). Imagine an $\{(n + 1) \times n\}$ specific-factors structure, with n specific factors and a single factor mobile between sectors. This is often taken to be labour, but instead, suppose it is a mobile input that is produced by all the specific factors. (That is, each 'specific' factor produces a particular commodity and, in addition, joins with other factors to produce the mobile factor.) This reduces to an $n \times n$ model with strong Stolper–Samuelson properties.

Fred Gruen and Max Corden (1970) introduced a simple model in discussing the possibility that a country such as Australia might, in levying a tariff, worsen its terms of trade. There are two sectors in the economy, manufacturing and agriculture. The manufacturing sector consists of a single commodity produced by labour and capital. Agriculture has two commodities, wheat and wool, each using labour and land. Thus, this is a special form of 3×3 model. As developed by Jones and Marjit (1992), it is possible to consider the $n \times n$ version of the Gruen–Corden model in which $(n - 2)$ sectors of the economy each use mobile labour and a type of capital specific to that sector to produce a single distinct commodity. In another sector labour and a specific type of capital produce a pair of commodities just as in the original Gruen–Corden case. An application of the Gruen and Corden model is also found in Findlay (1993).

The point of each of these special settings is that Heckscher–Ohlin models in the $n \times n$ case need not be difficult to analyse. However, the most popular two-factor model in the many commodity case may well be more valuable in that it focuses attention on which good or pair of goods a country produces in an international setting. Trade allows a great degree of specialization, and this version of the Heckscher–Ohlin model allows something that the special $\{n \times n\}$models, as well as the $\{(n + 1) \times n\}$ specific-factors model, do not, *viz.* treating the pattern of production as *endogenous* (see also Jones 2007a, b).

Joint Production

Both the Stolper–Samuelson theorem and the Rybczynski theorem are essentially reflections of the asymmetry between factors and commodities. This asymmetry is characterized by the assumption that productive activities are non-joint: in the non- degenerate cases more than one input is required to produce, separately, each output. Thus each commodity price change is a positive weighted average of the changes in rewards to factors used to produce that commodity. This implies that regardless of the ranking of commodity price changes, there is some factor reward that would rise relative to any commodity price rise and at least one factor reward which would rise by relatively less (or fall by more) than any commodity price change. Allowing joint production potentially destroys this asymmetry and thus the basis for the magnification effects.

There is a small literature dealing with this issue (Jones and Scheinkman 1977; Chang et al. 1980; Uekawa 1984). Much depends on the range of output proportions in any productive activity compared with the range of input proportions. For example, in the 2 × 2 case suppose one activity produces primarily the first commodity, but also a small amount of the second, while the other activity reverses these proportions. Furthermore, suppose this 'output' cone of diversification contains the standard 'input' cone of diversification. In this case traditional magnification effects underlying the Stolper–Samuelson and Rybcznyski theorems remain valid. New results emerge if these cones intersect or the input cone contains the output cone (Cones can be made comparable by using distributive shares of inputs and outputs in activities).

Joint production does not, by itself, interfere with the status of the factor-price equalization theorem (Jones 1992; but see Samuelson 1992, for an alternative view). However, joint production does suggest an alteration of the Heckscher–Ohlin theorem. Instead of concentrating on the link between factor endowments and the location of commodity outputs (and therefore trading patterns), the focus is on the location of productive activities. Each activity requires, as before, an array of inputs, and the allocation of endowment bundles among countries helps to determine where these activities are located. The pattern of commodity trade must then reflect, as well, the output composition of these activities.

Concluding Remarks

The theory of international trade that has developed from the seminal writings of Heckscher and Ohlin is fundamentally based on the twin observations that countries differ from each other in the composition of their factor endowments and that productive activities are distinguished by the different relative intensities with which factors are required. As this theory has been developed four core propositions have served to summarize its content. The strict validity of each of these propositions has been seen to depend upon further specification of the technology (for example, ruling out factor intensity reversals, joint production, and non-constant returns to scale, and imposing, for some results, that countries share the same technology), demand (for example, requiring all individuals to possess identical homothetic taste patterns), or dimensionality (for example, requiring a small number of factors and commodities, or a matching number of both). To conclude this discussion of the core propositions it is possible to point out the less precise, broad message of each.

The Heckscher–Ohlin Theorem

Production patterns reflect different compositions of endowments and, unless demand differences are significant, so will patterns of trade. International trade encourages specialization in production in those activities requiring factors in proportions similar to the endowment bundle and allows a country to import commodities whose factor requirements are far from proportions found at home. In some of the writings on 'new trade theory', assumptions are made that all varieties of a certain type of product are produced using the same factor proportions. By assumption this rules out the Heckscher–Ohlin theorem as an explanation of trade patterns. However, if varieties differ in quality, each variety could differ in factor requirements as well, serving to re-establish the relevance of the Heckscher–Ohlin theorem.

The Factor-Price Equalization Theorem

Even if the international mobility of factors of production is ruled out by national frontiers, free trade in commodities helps to even out disparities in demand relative to supply of factors and to diminish the discrepancy between factor returns among countries. Two or more countries sharing the same technology will find that free trade brings factor returns to absolute equality if their endowments are sufficiently similar and they produce in common a sufficient number of commodities (at least equal to the number of distinct productive factors).

The Stolper–Samuelson Theorem

Changes in relative commodity prices, such as those brought about by trade or interferences in trade, have strong asymmetric effects on factor rewards. If no joint production prevails, some factors find their real rewards unambiguously raised and other rewards are unambiguously lowered by relative price changes. If, further, the number of factors equals the number of produced commodities, as in the original 2 × 2 setting, and production is non-joint, relative commodity price changes can be constructed which, without the aid of any direct subsidies, will raise the real reward of any particular factor regardless of its taste pattern.

The Rybczynski Theorem

Unbalanced growth in factor supplies tends, at given commodity prices, to lead to stronger asymmetric changes in outputs. If the numbers of factors and commodities are evenly matched and production is non-joint, this asymmetry entails that growth in some, but not all, factors (when commodity prices are given) serves to force an actual reduction in one or more outputs. By similar reasoning, differences in the composition of endowments among countries with similar technologies results in stronger asymmetries in production patterns when all face free trade commodity prices. If tastes are somewhat similar, these endowments differences are apt to support the trading patterns described by the Heckscher–Ohlin theorem.

See Also

- ▶ Factor Price Equalization (Historical Trends)
- ▶ General Equilibrium
- ▶ International Trade Theory
- ▶ Leontief Paradox

Bibliography

Chang, W., W. Ethier, and M. Kemp. 1980. The theorems of international trade with joint production. *Journal of International Economics* 10: 377–394.

Chipman, J. 1966. A survey of the theory of international trade, Part 3. *Econometrica* 34: 18–76.

Chipman, J. 1969. Factor price equalization and the Stolper–Samuelson theorem. *International Economic Review* 10: 399–406.

Deardorff, A. 1980. The general validity of the law of comparative advantage. *Journal of Political Economy* 88: 941–957.

Deardorff, A. 1982. The general validity of the Heckscher–Ohlin theorem. *American Economic Review* 72: 683–694.

Deardorff, A., and R. Stern, eds. 1994. *The Stolper–Samuelson theorem: A golden jubilee*. Ann Arbor: University of Michigan Press.

Dixit, A., and V. Norman. 1980. *Theory of international trade*. Cambridge: Cambridge University Press.

Ethier, W. 1974. Some of the theorems of international trade with many goods and factors. *Journal of International Economics* 4: 199–206.

Ethier, W. 1984. Higher dimensional issues in trade theory. In *Handbook of international economics*, ed. R. Jones and P. Kenen, vol. 1. Amsterdam: North-Holland.

Findlay, R. 1993. Wage dispersion, international trade and the services sector. In *Trade, growth and development: The role of politics and institutions*, ed. G. Hansson. London: Routledge.

Findlay, R. 1995. *Factor proportions and growth*. Cambridge, MA: MIT Press.

Flam, H., and M.J. Flanders, eds. 1991. *Heckscher–Ohlin trade theory.* Cambridge, MA: MIT Press.

Gale, D., and H. Nikaido. 1965. The Jacobian matrix and the global univalence of mappings. *Mathematische Annalen* 159: 81–93.

Gruen, F., and W. Corden. 1970. A tariff that worsens the terms of trade. In *Studies in international economics*, ed. I. McDougall and R. Snape. Amsterdam: North-Holland.

Heckscher, E. 1919. The effect of foreign trade on the distribution of income. *Ekonomisk Tidskriff*: 497–512. Translated as chapter 13 in American Economic Association, *Readings in the theory of international trade*. Philadelphia: Blakiston, 1949, and a new translation is provided in Flam and Flanders (1991).

Jones, R. 1956. Factor proportions and the Heckscher–Ohlin theorem. *Review of Economic Studies* 24: 1–10.

Jones, R. 1965. The structure of simple general equilibrium models. *Journal of Political Economy* 73: 557–572.

Jones, R. 1971. A three-factor model in theory, trade, and history. In *Trade, balance of payments, and growth*, ed. J.N. Bhagwati et al. Amsterdam: North-Holland.

Jones, R. 1974. The small country in a many-commodity world. *Australian Economic Papers* 13: 225–236.

Jones, R. 1985. Relative prices and real factor rewards: A reinterpretation. *Economic Letters* 19: 47–49.

Jones, R. 1992. Jointness in production and factor-price equalization. *Review of International Economics* 1: 10–18.

Jones, R. 2002. Heckscher–Ohlin trade models for the new century. In *Bertil Ohlin: A centennial celebration (1899–1999)*, ed. R. Findlay, L. Jonung, and M. Lindahl. Cambridge, MA: MIT Press.

Jones, R. 2003. Joint output and real wage rates. *International Review of Economics and Finance* 12: 513–516.

Jones, R. 2006a. Eli Heckscher and the holy trinity. In *Eli Heckscher, international trade, and economic history*, ed. R. Findlay et al. Cambridge, MA: MIT Press.

Jones, R. 2006b. 'Protection and real wages': The history of an idea. *Japanese Economic Review* 57: 457–466.

Jones, R. 2007a. Key international trade theorems and large shocks. *International Review of Economics and Finance*.

Jones, R. 2007b. Specific factors and Heckscher–Ohlin: An intertemporal blend. *Singapore Economic Review* 52: 1–6.

Jones, R., and S. Marjit. 1985. A simple production model with Stolper–Samuelson properties. *International Economic Review* 19: 565–567.

Jones, R., and S. Marjit. 1991. The Stolper–Samuelson theorem, the Leamer triangle, and the produced mobile factor structure. In *Trade, policy, and international adjustments*, ed. A. Takayama, M. Ohyama, and H. Otah. New York: Academic Press.

Jones, R., and S. Marjit. 1992. International trade and endogenous production structures. In *Economic theory and international trade: Essays in memoriam J. Trout Rader*, ed. W. Neuefeind and R. Riezman. Berlin/New York: Springer-Verlag.

Jones, R., and J. Scheinkman. 1977. The relevance of the two-sector production model in trade theory. *Journal of Political Economy* 85: 909–935.

Jones, R., S. Marjit, and T. Mitra. 1993. The Stolper–Samuelson theorem: Links to dominant diagonals. In *General equilibrium, growth and trade II: Essays in honor of Lionel W. McKenzie*, ed. R. Becker, R. Jones, and W. Thomson. New York: Academic Press.

Jones, R., H. Beladi, and S. Marjit. 1999. The three faces of factor intensities. *Journal of International Economics* 48: 413–420.

Kemp, M., and L. Wegge. 1969. On the relation between commodity prices and factor rewards. *International Economic Review* 10: 407–413.

Krueger, A. 1977. *Growth, distortion, and patterns of trade among many countries*, Princeton studies in international finance, no. 40. Princeton: Princeton University Press.

Leontief, W. 1953. Domestic production and foreign trade: The American capital position re-examined. *Proceedings of the American Philosophical Society* 97: 332–349.

Lerner, A. 1933. Factor prices and international trade. Mimeo. Published in *Economica* 19 (1952): 1–15.

Mas-Colell, A. 1979. Two propositions on the global univalence of systems of cost functions. In *General equilibrium, growth, and trade*, ed. J. Green and J. Scheinkman. New York: Academic Press.

McKenzie, L. 1955. Equality of factor prices in world trade. *Econometrica* 23: 239–257.

Metzler, L. 1949. Tariffs, the terms of trade, and the distribution of national income. *Journal of Political Economy* 57: 1–29.

Minabe, N. 1966. The Heckscher–Ohlin theorem, the Leontief paradox, and patterns of economic growth. *American Economic Review* 56: 1193–1211.

Mitra, T., and R. Jones. 1999. Factor shares and the Chipman condition. In *Trade, welfare and econometrics: Essays in honor of John S. Chipman*, ed. J. Melvin, J. Moore, and R. Riezman. New York: Routledge.

Ohlin, B. 1933. *Interregional and international trade*. Cambridge, MA: Harvard University Press.

Ruffin, R. 1988. The missing link: The Ricardian approach to the factor endowment theory of trade. *American Economic Review* 78: 759–772.

Rybczynski, T.M. 1955. Factor endowments and relative commodity prices. *Economica* 22: 336–341.

Samuelson, P. 1948. International trade and the equalization of factor prices. *Economic Journal* 58: 163–184.

Samuelson, P. 1949. International factor-price equalization once again. *Economic Journal* 59: 181–197.

Samuelson, P. 1953. Prices of factors and goods in general equilibrium. *Review of Economic Studies* 21: 1–20.

Samuelson, P. 1971. Ohlin was right. *Swedish Journal of Economics* 73: 365–384.

Samuelson, P. 1992. Factor-price equalization by trade in joint and non-joint production. *Review of International Economics* 1: 1–9.

Stolper, W., and P. Samuelson. 1941. Protection and real wages. *Review of Economic Studies* 9: 58–73; also in Deardorff and Stern (1994).

Uekawa, Y. 1971. Generalization of the Stolper–Samuelson theorem. *Econometrica* 39: 197–213.

Uekawa, Y. 1984. Some theorems of trade with joint production. *Journal of International Economics* 16: 319–333.

Hedging

Gregory Connor

Abstract

Hedging is defined with a state-space model of risky outcomes. Full and partial hedging are compared, and the feasible set of hedging positions related to the available collection of traded assets. Three types of counter-parties for hedging trades are distinguished. The risk premium for a hedging asset is defined, and its relationship to economy-wide risk factors explained. The case of mean-variance preferences provides a useful formula for the optimal hedge position. Corporations undertake many hedging transactions, even though the shareholders of the corporation do not typically benefit from any risk reduction. Some explanations of corporate hedging are set out.

Keywords

Adverse selection; Asset pricing; Bankruptcy; Bid ask spread; Brownian motion; Corporate hedging; Differential information; Futures markets; Hedge portfolio; Hedging; Keynes, J. M.; Mean-variance preference model; Moral hazard; Normal backwardation; Portfolio insurance; Progressive taxation; Risk aversion; Risk premium; Speculation; State space models; Taxation of corporate profits

JEL Classifications

G1

Hedging is the purchasing of an asset or portfolio of assets in order to insure against wealth fluctuations from other sources. A hedge portfolio is any asset or collection of assets purchased by one or more agents for hedging. A grain dealer may hedge against losses on an inventory of grain by selling grain futures; a Middle Eastern businessman may hedge against political turmoil (and the resulting losses) by buying gold; a pension fund may hedge against capital losses on its equity portfolio by buying stock index put options.

A Competitive Equilibrium Model of Hedging

The fundamental concepts of hedging can best be described in the state space model. Consider a one-period economy with M agents and one end-of-period consumption good. For simplicity, assume that there is no consumption at the beginning of the period. Each agent possesses a real asset which produces a random amount of the consumption good at the end of the period. Agents have homogeneous beliefs. There are N possible states of nature, with probabilities $Pr(1),\ldots,$ $Pr(N)$. The agents haveconcave, possibly state-dependent utility functions and wish to maximize the expected utility of end-of-period consumption. Let $U_j(C_j,\ \theta_i)$ denote the end-of-period utility of agent j given that his consumption is C_j and the state of nature is θ_i.

A *financial asset* is a claim to a random amount of end-of-period output which is traded between agents at the beginning of the period. A *hedge portfolio* is a particular type of financial asset or collection of financial assets which protects an agent against some particular risky outcome(s).

The analysis is simplest if we assume that the hedge portfolio consists of a mixed asset-liability with positive payoffs in some states and negative payoffs in other states, balanced so as to give a competitive equilibrium price of zero. Under this formulation a hedge portfolio is a portfolio which pays off positively in states where the agent would otherwise have a high marginal utility of consumption (that is, ‘bad’ states) and negatively in states where he would otherwise have a low marginal utility of consumption. If the agent’s marginal utility is equalized across the relevant states after purchasing the hedge portfolio, then he is fully hedged; if the hedge position lowers but does not eliminate the disparity, then he is partially hedged.

Who takes the other side of the hedging transaction? There are three possibilities. First, if there exist two agents who have real asset cash flows which vary inversely, then they can trade in a way which allows both to hedge simultaneously. For example, the grain dealer who holds an inventory of grain may be able to sell a futures contract to a bread producer who has committed himself to using grain at a later stage of his production process. Both parties consider themselves as hedging. Second, one agent may be less risk-averse towards certain states of nature than another. The less risk-averse may be willing to sell the hedge asset to the more risk-averse at a price which produces mutual gains in expected utility. Third, the hedging agent may be able to trade small quantities of the hedge asset with many agents, who can then eliminate all or most of the risk of the trade by combining the asset with many others (that is, by diversifying away the risk). For example, insurance companies can sell fire insurance policies to many individuals and leave very little risk to be absorbed by the company's shareholders.

Let the number of distinct types of assets be *K* and let *Y* denote the *NxK* matrix of their payoffs in the *N* possible states of nature. The set of available trades is *span* (*Y*) where *span*(·) denotes the subspace spanned by the matrix. In an economy without frictions, agents will create new financial assets until all mutually beneficial trade opportunities are in *span*(*Y*). All mutually beneficial trades have been consummated if there exist positive scalars $\lambda_1 \ldots \lambda_M$ such that

$$\lambda_j U_j'\left(C_j, \theta_i\right) = \lambda_h U_h'\left(C_j, \theta_i\right) \quad i = 1, \ldots, N; \quad j, h = 1, \ldots, M \tag{1}$$

where U'denotes the first derivative with respect to consumption. The invisible hand drives agents towards creating all the types of financial assets which can lead to mutually beneficial trades. However, there are many external factors which can offset this tendency. If agents have some control over outcomes, then moral hazard problems may limit hedging opportunities. For example, agents may not be able to hedge against changes in labour income if work requires imperfectly observable effort. If agents have special knowledge, then adverse selection can similarly limit trade. If a car owner knows more about its quality than a prospective buyer, then the owner cannot sell his car at a reasonable price when he experiences financialdistress. The administrative costs of trade can also limit hedging before the full efficiency condition (1) is fulfilled.

The model described above is static. In an intertemporal model, dynamic strategies increase the set of hedging opportunities beyond the linear span of the matrix of asset payoffs. Agents can create a rich set of payoff claims by dynamically varying the proportions invested in the individual assets. With continuous trading, this process reaches its natural limit: if an asset price follows Brownian motion, then the continuously adjusted portfolio consisting of only the risky asset and riskless asset can be constructed which replicates the payoff to any put or call option on the risky asset.

The proliferation of complex financial assets, such as options on futures and interest rate and currency options, and the increased sophistication of traders has led to a bewildering array of dynamic hedging strategies, especially by large institutional investors. *Portfolio insurance* provides a good example of the kind of sophisticated new hedging instrument which can be created with a dynamic trading strategy. Consider a pension fund with a large equity portfolio and an aversion to large capital losses on this portfolio. The portfolio insurance strategy can put a floor on the random rate of return to the pension fund's portfolio. The return floor can be any rate lower than the available riskless rate (it can be a negative net return, so that the fund bounds its losses rather than assuring itself a small gain). The strategy works as follows. At the starting date of the insurance strategy, the fund has most of its money invested in equities and a small proportion in a riskless asset (that is, government notes). If the equities fall in price, the fund sells some of the equities and places the cash in the riskless asset. If equity prices continue to fall, the fund increases the proportion of investment in the riskless asset. If there is a sustained fall in equity prices, the fund will end the insurance programme invested entirely in the riskless asset. It will have earned a rate equal to the pre-chosen minimally acceptable return. The fund makes a 'soft landing' at this minimal value: the proportion of money invested in the riskless asset approaches 1 as the value of the portfolio approaches the minimally acceptable level.

Portfolio insurance is not a free lunch. In exchange for the return floor, the pension fund sacrifices some of the upside potential of pure equity investment. For example, if the equity market declines sharply and then rises, the fund will miss the upturn, since it will have defensively decreased its position in equities before the upturn.

There are numerous other dynamic hedging strategies, not only in equity markets but in fixed income, currency and options markets. In terms of the volume of trade, hedging in financial markets now greatly outpaces activity in commodities futures markets, the original and classic example of markets often used for hedging.

Risk Premia and Hedging

An economically interesting question is whether agents 'pay a premium' to hedge. Assume again that the current price of the hedge portfolio is set to zero by appropriate balancing of the asset and liability sides of the hedge (a futures contract is a natural example). If the expected cash flow is negative (positive) next period, then the hedge portfolio carries a positive (negative) implicit asset pricing risk premium. If the expected cash flow is zero, then the implicit asset pricing risk premium is zero. If we used returns rather than prices, then the expected return premium would have the opposite sign from the asset pricing premium.

Much of the early literature on hedging was centred on hedging in commodities futures contracts. One of the key questions was whether agents who sold futures pay a positive risk premium. Keynes (1930) considers this problem for the case of commodities futures contracts. He argues that the natural supply of short sellers (sellers of futures contracts) outnumbers long hedgers (buyers) in this market. Therefore, the implicit risk premium for holding a futures contract should be negative, in order to induce other agents (henceforth called speculators) to absorb the excess hedging demand of short hedgers. This will be true if the futures price increases on average over the life of the contract, so that the expected cash flow from holding the contract is positive.

The empirical evidence for this positive trend (sometimes called *normal backwardation*) in commodities futures market prices is weak at best. Keynes's analysis implicitly assumes that the commodity futures market is isolated from other assets so that hedgers must pay other agents (the speculators) a premium to induce them to take a position in the market. In an integrated set of asset markets, hedgers need not pay any premium to induce other agents to trade. Rather, the existence of a risk premium depends upon the covariance between the payoffs to the hedge asset and theeconomy-wide risks faced by all the agents. If the hedge asset is uncorrelated with market-wide risks, then it will carry no risk premium, even though it may have a high value to a particular hedger due to his specific income stream. The hedge asset which protects against market risk will carry a risk premium.

There is another source of return to speculators, which is not captured in the competitive pricing model. Speculators may charge an explicit or implicit bid ask spread when trading with hedgers. If hedgers buy and hold for a long period, then this is equivalent to the asset pricing risk premium described above. However, if hedgers trade frequently, then the bid ask spread can lower their realized returns, and raise the realized returns of speculators, without affecting the observed long-run return premium of the hedge asset as reflected in transactions prices. This may explain the lack of empirical evidence for normal backwardation in commodities futures market. This effect of the bid ask spread was not recognized in most of the early literature on commodity futures markets.

The bid ask spread need not be explicit. Even open-floor markets will contain a set of implicit bid ask spreads, to the extent that traders' strategies reflect a greater willingness to sell at higher prices and to buy at lower ones. One can view the feverish activity which is common in floor trading as speculators searching for transactions at the outer edges of an implicit bid ask spread. Hedgers, who are off the floor and are more anxious to complete a particular trade, take the losing side of the implicit spread.

The Role of Hedgers in a Market with Heterogeneous Information

The model in sections "A Competitive Equilibrium Model of Hedging" and "Risk Premia and Hedging" assumes homogeneous information across agents. If agents have differential information about the payoffs to assets, then the trading strategies of rational agents cannot take this simple competitive form. Agents must treat trade opportunities as signals of the information of other agents about the value of the trade.

The presence of differential information can lead to fewer hedging opportunities and/or raise the expected cost of hedging. Milgrom and Stokey (1982) show that rational agents will not trade solely because they have different beliefs about the value of an asset. If agents are distinguished merely by their differential information, then they will refuse all trades, since the willingness of the other agents to trade signals that the terms are unfavourable. This means that a financial asset market will fail to open in the absence of other motives for trade. This is a market failure due to adverse selection. The need of some agents for hedging can provide an additional reason for trade which overcomes the adverse selection problem and eliminates the market failure. Hedgers will be willing to trade even if they suspect that the other party to the transaction has superior information. Informed agents will be gaining at the expense of hedgers, but they will also be providing an insurance or liquidity service to hedgers, and so hedgers may be willing to trade with them despite their informational disadvantage. This in turn has the side effect of permitting superior information to be reflected in market prices.

I follow Glosten and Milgrom (1985) and assume that there exists a costless, competitive, risk-neutral market maker who intervenes in all trades. This is for analytical convenience and is not necessary to the basic model. Suppose that certain agents (hedgers) have a strong preference for a given asset, that is, their preference is such that they will buy (and sell) some non-zero amount at a price higher (lower) than the market-clearing equilibrium price. This implies that they are willing to trade even if they must pay a bid ask spread around the equilibrium price. Informed agents, henceforth speculators, will also trade despite a bid ask spread as long as the expected profit from their superior information is larger than the bid ask spread. The market maker's resulting equilibrium bid ask spread will allow the hedgers to trade at an expected loss and the speculators at an expected gain, leaving the market maker with an expected profit of zero (the equilibrium condition). The market maker will respond to the net demands of all traders (which partially reveals the net demand of speculators) to adjust the bid and ask prices, and so (partially) capture in the market price the superior information of speculators.

One interesting feature of this model is the symbiotic roles of speculators and hedgers. Without speculators (informed traders), the hedgers would lose liquidity; without hedgers, speculators would lose the opportunity to profit from their superior information. Without both speculators and hedgers, the price in the market would no longer provide a useful signal for agents making production and consumption decisions. Kyle (1984) develops a model in which the symbiotic relationship is made clear and describes the effects of more or fewer speculators or hedgers on the informational efficiency and liquidity of the market. Some of the results are counterintuitive: for instance, increasing the number of hedgers, who are uninformed, can increase the informational efficiency of prices.

Hedging in a Mean Variance Model

The mean variance preference model provides a useful framework for empirical and applied analysis of hedging. Suppose that an investor's utility is given by the expected value of his random end-of-period endowment, X, minus some multiple of the variance of this endowment:

$$U(X) = E[X] - a\,\mathrm{var}[X]$$

Let the hedging instrument have a current price of zero and random end-of-period payoff Y. It is easy to show that the investor's optimal position in the hedging instrument is

$$w = \frac{E[Y]}{(2avar[Y])} - \frac{\mathrm{cov}[X, Y]}{\mathrm{var}[Y]}.$$

(See Rolfo 1980, for the derivation). The first additive part of this optimal hedging position, $E[Y]/(2avar[Y])$, is called the *speculative hedge.* The second part,$-cov[X, Y]/var[Y]$, is called the *pure hedge.* Some analysts (such as Duffie 1989) argue that in practice uninformed hedgers should ignore the speculative hedge and set the hedge position equal to the pure hedge. The justification is that the speculative hedge requires predictions about the expected payoff on the hedging instrument, whereas the pure hedge uses only the covariance of the hedging instrument and the random endowment. In many cases, covariances are more stable over time and more precisely estimated than expected payoffs. Setting the hedge position equal to the pure hedge is equivalent to minimizing variance instead of optimizing over a mean variance criterion.

The mean variance preference model provides a useful link to empirical analysis of hedging. Suppose that we observe a sample of realized random endowments and hedging instrument payoffs. Consider an ordinary least squares regression of the random endowment on the payoffs to the hedging instrument:

$$X = \alpha + \beta Y + \varepsilon.$$

The coefficient β estimates (minus) the pure hedge. The R-squared from this regression estimates the proportion of endowment variance which is eliminated by setting the hedge position equal to the pure hedge.

Risk Premia on Hedge Portfolios and General Equilibrium Pricing

In section "Risk Premia and Hedging", I described two types of hedge portfolios – those with and those without risk premia– and how the distinction between them depends on the covariance between the hedge portfolio returns and the market-wide risks in the economy. In this section, I describe the relationship between hedge portfolios which protect against market-wide risks and the general equilibrium pricing of assets.

Let Q_t denote the discounted expected utility of lifetime consumption for some agent at time t:

$$Q_t = E_t\left[\sum_{\tau=t+1}^{\infty} \rho^{\tau} U(C_\tau, \theta_\tau)\right]$$

where ρ is the agent's discount factor and $U(.,.)$ is his utility function. Let f_t. denote the change in discounted expected utility given the change in the agent's time t wealth:

$$f_t = \frac{\partial Q_t}{\partial W_t}$$

where W_t is his wealth at time t. Note that, at time $t - 1$, f_t is a random variable. Let r_{it} denote the return from time $t - 1$ to t of the i^{th} financial asset. If the agent holds an equilibrium amount of this asset then the following first-order condition is satisfied:

$$E_{t-1}[r_{it} f_t] = \frac{\partial U(C_{t-1}, \theta_{t-1})}{\partial C_{t-1}}$$

which can be rewritten (using $E[ab] = E[a]E[b] + cov[a, b]$) as:

$$E_{t-1}[r_{it}] = r_{0t} + \left(\frac{1}{\gamma}\right) cov_{t-1}[r_{it}, f_t] \qquad (2)$$

where $\gamma = \frac{\partial\, U(C_{t-1}, \theta_{t-1})}{\partial\, C_{t-1}}$ and r_{0t} is the expected return on an asset with a riskless payoff at time t. Suppose that, at time $t - 1$, f_t equals asum of a set of K uncorrelated random variables $Z_{1t}, \ldots, Z_{Kt}$:

$$f_t = Z_{1t} + \ldots + Z_{Kt} \qquad (3)$$

The variables $Z_{1t}, \ldots, Z_{Kt}$ describe the K random shocks which affect the agent's marginal utility. They could be interest rate movements, output shocks, inflation shocks, and so on. Assume that there exists a set of K portfolios with returns $r^*_{1t}, \ldots, r^*_{Kt}$ such that the j^{th} portfolio has perfect negative correlation with Z_{jt}:

$$cov_{t-1}\left[r_{jt}^*, Z_{jt}\right] = \left(var_{t-1}\left[r_{jt}^*\right]var_{t-1}\left[Z_{jt}\right]\right)^{\frac{1}{2}}. \tag{4}$$

These portfolios are potential hedges against the K types of risks which affect the investor (The agent will short sell the portfolio to hedge since the portfolio return varies inversely with marginal utility). I call $r_{1t}^*, \ldots, r_{Kt}^*$ an indexing set of hedge portfolios since the portfolios index the random shocks to the agent's marginal utility. Using (3) and (4) we can rewrite (2) as:

$$E_{t-1}[r_{it}] = r_{0t} + \beta_{i1t}\pi_{1t} + \ldots + \beta_{iKt}\pi_{Kt} \tag{5}$$

Where

$$\beta_{ijt} = \frac{cov_{t-1}\left[r_{it}, r_{jt}^*\right]}{var_{t-1}\left[r_{jt}^*\right]}$$

and

$$\pi_{jt} = E_{t-1}\left[r_{jt}^* - r_{0t}\right].$$

Equation (5) is an asset pricing relationship: it says that the expected return on any asset equals the riskless return plus a linear combination of the covariances of the asset's return with an indexing set of hedge portfolios.

Explaining Corporate Hedging

The analysis in the previous sections considers hedging from the perspective of an individual investor and consumer. In practice, most financial hedging activity is undertaken by corporations rather than by individuals. If corporations issue widelyheld common stock, then hedging activity between them appears unnecessary. The shareholders of the corporation are its ultimate risk bearers, and they will not benefit from hedging activity at the corporate level. So, for example, an investor with shares in both an oil-producing industry and an oil-consuming industry will not want firms in the two industries to hedge with oil futures, taking opposite hedge positions (one long oil futures and the other short oil futures). From the shareholder's perspective, the cash flows from these offsetting hedges will be diversified away at the portfolio level. The ultimate shareholder pays the transactions costs of hedging by the corporations, but does not experience any aggregate risk reduction in his portfolio.

There are several explanations for the prevalence of corporate hedging. One explanation relies on the costs of financial distress. By hedging, the corporation lowers the probability of bankruptcy or near-bankruptcy, and so increases its average market value. Another explanation relates corporate hedging to the agency costs of hiring managers to run the firm. The firm's managers have a non-diversified exposure to the profitability of the firm; hedging by the corporation more closely aligns the interests of the shareholders and managers, and also allows the shareholders to pay the managers less on average since the managers' risk exposure is reduced. A third explanation relies on the progressiveness of the corporate tax system, which encourages corporations to hedge so as to smooth taxable earnings and thereby lower their average tax bill.

See Also

- ▶ Capital Asset Pricing Model
- ▶ Futures Markets, Hedging and Speculation
- ▶ Mean-Variance Analysis
- ▶ Risk Aversion
- ▶ State Space Models

Bibliography

Arrow, K. 1964. The role of securities in the optimal allocation of risk bearing. *Review of Economic Studies* 31: 91–96.

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

Breeden, D. 1984. Futures markets and commodity options: Hedging and optimality in incomplete markets. *Journal of Economic Theory* 32: 275–300.

Brennan, M. 1958. The supply of storage. *American Economic Review* 48: 50–72.
Cootner, P. 1967. Speculation and hedging. *Food Research Institute Studies, Supplement* 7: 65–105.
Duffie, D. 1989. *Futures markets*. Englewood Cliffs: Prentice-Hall.
Duffie, D., and C. Huang. 1985. Implementing Arrow–Debreu equilibrium by continuous trading of a few long-lived securities. *Econometrica* 53: 1337–1356.
Glosten, L., and P. Milgrom. 1985. Bid, ask, and transaction prices in the specialist market with heterogeneously informed traders. *Journal of Financial Economics* 14: 1–00.
Grauer, F., and R. Litzenberger. 1979. The pricing of commodity futures contracts, nominal bonds and other risky assets under commodity price uncertainty. *Journal of Finance* 34: 69–84.
Gray, R. 1961. The search for a risk premium. *Journal of Political Economy* 69: 250–260.
Grossman, S. 1976. On the efficiency of competitive stock markets when traders have diverse information. *Journal of Finance* 31: 573–585.
Hoffman, G. 1954. Past and present theory regarding futures trading. *Journal of Farm Economics* 19: 1–11.
Houthakker, H. 1957. Can speculators forecast prices? *Review of Economics and Statistics* 39: 143–152.
Keynes, J.M. 1930. A treatise on money. volume 2: The applied theory of money. Reprinted in *The collected writings of John Maynard Keynes*, vol. 7. London: Macmillan, 1971.
Kyle, A. 1984. Market structure, information, futures markets, and price formation. In *International agricultural trade: Advanced readings on price formation, market structure, and price instability*, ed. G. Storey, A. Schmitz, and A. Sarris. London: Westview.
Leland, H. 1980. Who should buy portfolio insurance? *Journal of Finance* 35: 581–594.
Malinvaud, E. 1972. The allocation of individual risk in large markets. *Journal of Economic Theory* 4: 312–328.
Merton, R. 1973. An intertemporal capital asset pricing model. *Econometrica* 41: 867–887.
Milgrom, P., and N. Stokey. 1982. Information, trade and common knowledge. *Journal of Economic Theory* 26: 17–27.
Rolfo, J. 1980. Optimal hedging under price and quantity uncertainty: The case of a cocoa producer. *Journal of Political Economy* 88: 100–116.
Smith, C., and R. Stulz. 1985. The determinants of firms' hedging policies. *Journal of Financial and Quantitative Analysis* 20: 391–405.
Working, H. 1953a. Futures trading and hedging. *American Economic Review* 43: 314–343.
Working, H. 1953b. Hedging reconsidered. *Journal of Farm Economics* 35: 544–561.
Working, H. 1967. Tests of the theory concerning for trading on commodity exchanges. *Food Research Institute Studies, Supplement* 7: 5–48.

Hedonic Functions and Hedonic Indexes

Jack E. Triplett

Hedonic Functions

A hedonic function is a relation between prices of varieties or models of heterogeneous goods – or services – and the quantities of characteristics contained in them:

$$P = h(c) \tag{1}$$

where P is an n-element vector of prices of varieties, and (c) is a $k \times n$ matrix of characteristics. The theory providing its economic interpretation rests on the *hedonic hypothesis* – heterogeneous goods are aggregations of characteristics, and economic behaviour relates to the characteristics.

The hedonic hypothesis implies that a transaction is a tied sale of a bundle of characteristics, so the price of a variety is interpreted as itself an aggregation of lower-order prices and quantities. Characteristics are assumed the true arguments of utility functions; they are the inputs to the production process, in the case of heterogeneous materials, capital goods, or labour services. Hence:

$$Q = Q(c, M) \tag{2}$$

where Q is utility (output), M is a vector of other, homogeneous consumption goods (productive inputs), and for expositional simplicity we specify only one heterogeneous good in the system, with characteristics (c). For a heterogeneous labour type, productive characteristics are typically assumed to have been acquired through investment in human capital, so that (1) is a hedonic wage equation, and human capital characteristics appear in (2). It is common to assume, for durable goods and for labour, that services of characteristics are proportional to their stocks, through characteristics may decay at varying rates. Analysis of consumer behaviour toward characteristics of goods is frequently linked to the literature on household production, but the two subjects are conceptually distinct, and the latter is ignored here, in the interest of brevity.

Production of the heterogeneous good is the joint production of a bundle of characteristics:

$$t(c, K, L) = 0 \tag{3}$$

Characteristics may be attached to goods through externalities (air quality as a housing characteristic) or by an act of nature (risk as an attribute of jobs) as well as by explicit production decisions of producers.

Equations (2)–(3) exhibit the extreme form of the hedonic hypothesis: *only* the characteristics of heterogeneous goods enter behavioural relations. Plausible cases exist where both quantities of goods and of their characteristics matter, particularly where there are complementarities in (2) between characteristics and other inputs or outputs (two small cars are not necessarily equivalent to a large one with the same total quantities of characteristics because consumption also requires input of driving time), or when conventional scale economies are present in (3). For present purposes, such additional structure is dispensed with because it complicates the exposition, and because it is more relevant to investigating the demand and supply of characteristics than for explaining hedonic functions. For the same reasons, we cannot explore interesting cases of production where both inputs and outputs are heterogeneous.

It is well-established – but still not widely understood – that the form of $h(\cdot)$ cannot be derived from the form of $Q(\cdot)$ or of $t(\cdot)$, nor does $h(\cdot)$ represent a 'reduced form' of supply and demand functions derived from $Q(\cdot)$ and $t(\cdot)$. Establishing this result requires consideration of buyer and seller behaviour toward characteristics.

The Buyer, or User, Side

It is expositionally convenient to represent the user's choice of characteristics as a two-stage budgeting process. Suppose that (2) can be written

$$Q = Q(q(c), M) \tag{4}$$

where $q(\cdot)$ is an aggregator over the characteristics (c). Then conditional on M and a utility (output) level Q^*, the allocation of characteristics (choice of variety) can be determined by minimizing the cost of attaining the sub-aggregate $q(c)$. Thus, if q^* is a value of $q(\cdot)$ such that $Q^* = Q(q^*, M)$, the optimal choice of (c) is the solution to:

$$\min_c h(c), \qquad \text{s.t.} \;\; q(c) = q^* \tag{5}$$

Marginal conditions for an optimum are (where the subscript shows partial derivative with respect to c_i *or* c_j):

$$q_i / q_j = h_i / h_j \tag{6}$$

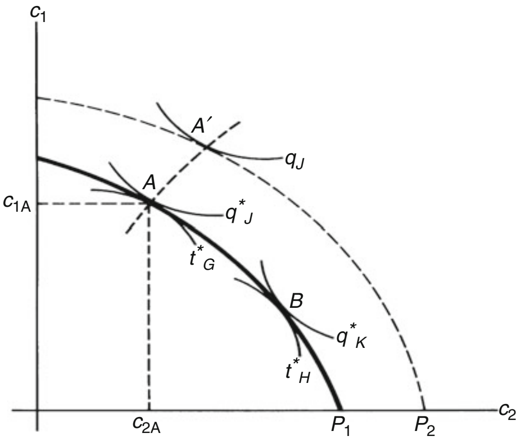
The ratio of marginal 'sub-utilities' of c_i and c_j must equal the ratio of acquisition costs for incremental units of c_i and c_j. Note that the ratio h_i/h_j is the slope of $h(\cdot)$ in the c_i/c_j plane, variety price held constant.

Suppose for illustration a non-linear, two-characteristic, continuous hedonic function such that, for any fixed price P^*, the graph of

$$P^* = h(c_1, c_2) \tag{7}$$

has the form of the contours P_1 and P_2 in Fig. 1. The locus P_1 connects all varieties selling for the price P_1 – point A designates a variety described by the vector [P_1, $c_{1\mathrm{A}}$, $c_{2\mathrm{A}}$], point B by [P_1, $c_{1\mathrm{B}}$, $c_{2\mathrm{B}}$], and so forth. The slope of P_1 at any point gives relative marginal acquisition costs for characteristics c_1 and c_2. The solution to the choice of variety problem is shown in Fig. 1 by the tangency of q_J^* – a partial or conditional indifference curve (isoquant) for user J – and P_1.

A quantal choice problem is contained in this optimization: The buyer selects the variety whose embodied characteristics are closest to the optimal ones. When the spectrum of varieties is continuous in c_1, c_2, the quantal choice is trivial so long as only one unit of the good is bought; Lancaster (1971), following Gorman (1980, but written in 1956), models the non-continuous case by



Hedonic Functions and Hedonic Indexes, Fig. 1

specifying the P_1, P_2,... contours as piece-wise linear, and permitting the buyer to obtain an optimal set of characteristics by combining two varieties.

The remainder of the user optimization problem proceeds as in other two-stage allocations. Total expenditure on characteristics (the level of 'quality' when only one unit is bought) is determined by:

$$\begin{array}{c}\max Q(q(c), M)\\ \text{subject to}: q(c)\cdot v(c) + P_M M = y\end{array} \qquad (8)$$

where $v(c)$, the price of the composite commodity q – or alternatively, the 'price of quality' – is the slope of the hedonic surface above an expansion path such as AA′ in Fig. 1. With respect to any good i in M, the solution entails:

$$Q_q/Q_i = v(c)/p_i \qquad (9)$$

and the set of such conditions determines total expenditure on characteristics (equals the price of the model chosen).

The characteristics–space problem has many similarities to normal 'goods space' problems. The hedonic frontiers P_1, P_2, ... provide analogues to conventional budget constraints (isocost lines) and serve to constrain the agent's optimization problem in characteristics space. These are the constraints themselves, *not* the cost functions of conventional duality theory. The constraints may be non-linear; if so, characteristics prices are not fixed, but are uniquely determined for each buyer by the buyer's location on the hedonic surface (compare the slope of P_1 at A and B). It is observed that varieties having differing characteristics are available at the same price and are chosen by different buyers (in Fig. 1, model B is chosen by buyer K). This suggests that divergence of tastes and technologies is an essential part of the theory for hedonic functions, and that 'representative consumer' (firm) models do not describe market outcomes.

If there are a large number of buyers, Rosen (1974) shows that each frontier P_1, P_2, ..., will trace out an envelope of tangencies with relations such as q^*_J, q^*_K, ... As with any envelope, the form of $h(\cdot)$ is independent of the form of $Q(\cdot)$ – except for special cases – and is determined on the demand side by the distribution of buyers across characteristics space. This is an important result for the interpretation of hedonic functions

Forming Measures of 'Quality'

It is well known that (4) implies weak separability of Q on (c), which permits consistent aggregation over the characteristics in (c) – see Blackorby et al. (1978). It is natural to take such an aggregate as a measure of 'quality'.

One can thus use weak separability on characteristics to rationalize the common practice of writing scalar 'quality' in the utility or production function, as for example Houthakker's (1952) model of quantity and quality consumed – a model that has many empirical progeny, and much appeal for its simplicity. Weak separability on characteristics also provides the analytic bridge between characteristics–space models and Fisher and Shell's (1972) notion of 'repackaging', in which quality change enters the utility function by scalar multiplication of the good whose quality changed. Because hedonic functions have mostly been *used* for purposes (like constructing a 'quality-adjusted' price index for automobiles: Griliches 1971) for which separability was assumed (usually implicitly), separability assumptions on characteristics are thus a common thread through most analysis of 'quality'.

Obviously, when Q is not separable on (c), no consistent scalar measure of ‘quality’ can be formed. It is not hard to think of cases where characteristics separability is not realistic (are refrigerator characteristics separable from what is stored in them, or transportation equipment characteristics separable from energy consumption?). One should note that characteristics–space approaches could be adapted to certain non-separable cases (computing the cost per mile of constant-quality transportation services), where scalar approaches may be more problematic. Moreover, since weights for the aggregator are the marginal subutilities q_1 and q_2, the quality measure will depend on relative characteristics prices – properly, on the position of the P-contours in Fig. 1 – whenever substitution among characteristics quantities is possible; a scalar quality measure is therefore not in general unique, even when consistent. These points suggest that a major advantage of hedonic, or characteristics-space, methods is their potential for dealing with non-separable cases and with changing relative characteristics costs regimes, though there is little demonstration of this potential in existing empirical work.

The Production Side

A comparable theory shows how a price-taking producer selects the optimal variety or varieties to sell, given (1) and (3). For a particular level of input usage or production cost, a two-characteristic form of $t(\cdot)$ yields transformation surfaces, for producers or production processes G and H, like t^*_{G} and t^*_{H} in Fig. 1. Revenue from increments of characteristics added to the design can be computed from partial derivatives of the hedonic function. Optimal product design is determined by:

$$t_1/t_2 = h_1/h_2 \tag{10}$$

The quantity produced of the optimal design is determined in the usual way by setting the marginal cost of quantities of the optimal design (equation omitted here for brevity) equal to the variety price (given by the hedonic function, $h(\cdot)$).

The production-side theory is problematic, compared with the user case, because in the absence of scale economies in the production of varieties, producers would build ‘custom products’, offering all product designs on the hedonic surface where variety price exceeds cost, rather than specializing in the most profitable variety. The competitive, large numbers case is thus not an appealing one, unless product design is to an extent fixed by sellers’ endowments at least in the short run (the normal assumption for labour markets, and for land).

If there are a large number of sellers, Rosen (1974) shows that, except for special cases, each hedonic frontier P_1, P_2,. . . will trace out an envelope of tangencies with relations such as t^*_{G} and t^*_{H}. As in the user case, the form of $h(\cdot)$ is therefore influenced on the supply side by the distribution of sellers across characteristics space and by their output scales, but the form of $h(\cdot)$ cannot in general be derived from the form of $t(\cdot)$.

Special Cases

If $q(\cdot)$ is identical for all users, then only a single set of q^* contours appears in Fig. 1, and each hedonic frontier P_1, P_2, . . . traces out the associated q^* contour. In this case, the form of $h(\cdot)$ is determined by the form of $q(\cdot)$, up to a monotonic transformation, and should conform to the principles of classical utility theory (which means that each hedonic frontier, P, bows inward, toward the origin, rather than as drawn in Fig. 1).

If $t(\cdot)$ is identical fo all sellers, then only a single set of t^* contours appears in Fig. 1, and each hedonic frontier P_1, P_2,. . . traces out the associated t^* contour. In this case, the form of h $(\cdot)$ is determined by the form of $t(\cdot)$, and the usual reasons for assuming convexity of production sets apply, so that the P-frontiers should bow outward from the origin, in the manner of a normal production transformation curve.

If there is no diversity on either side of the market, only one design will be available at each model price. The hedonic frontiers degenerate into a series of points, one for each model price.

Of these possibilities, uniformity of $t(\cdot)$ across sellers (except for labour services) is the most likely, especially in the long run when access to technology is freely available. Uniformity of $q(\cdot)$ is improbable, and appears inconsistent with available evidence.

Functional Forms for Hedonic Functions

Neither classical utility nor production theory can specify the functional form of $h(\cdot)$. The P-frontiers can bow in, bow out, or take the form of straight lines (or even irregular shapes). In particular, and contrary to assertions that have appeared in the literature, nothing in the theory rules out the semi-logarithmic form (which has often emerged as best in goodness-of-fit tests in both labour and product market hedonic studies). Though non-linear in P and (c), the semi-log is nevertheless linear in the $[c_i, c_j]$ plane and thus even has some 'nice' properties (because all buyers and sellers face the same characteristics prices, for equal expenditure on, or revenue from, characteristics).

Hedonic 'Demand' Studies

Hedonic functions have sometimes been used to generate demand or 'willingness to pay' estimates (particularly, of the value of air quality or neighbourhood amenities in land and housing prices, and of risk in labour markets). However, as Fig. 1 shows, buyers J and K, though located on the same hedonic price surface, may face different characteristics prices as a consequence of their preference functions; the slopes of the P-function at A and B do not represent exogenous price variance that determines characteristics allocations.

Unless one is willing to assume that all buyers have identical tastes, cross-section characteristics demand studies founder for the same reason as cross-section 'goods' demand studies: Variations in quantities reflect taste differences and not shifts in the slope of the budget constraint. Moreover, the situation depicted in Fig. 1 cannot be reduced to a demand estimation problem by treating it as some variant of an econometric demand–supply identification problem, despite some attempts to do so in the literature.

Hedonic Indexes

A hedonic price index is one that makes use of information from the hedonic function. Adding time dummy variables to a multi-period regression on (1) is a favourite empirical procedure, but is by no means the only way to compute a hedonic price index. A characteristics price index is any index that is defined on the characteristics of goods, or on behavioural functions in which characteristics are arguments. A hedonic price index is thus a particular implementation of a characteristics price index. Almost any empirical application of hedonic functions (e.g. use of hedonic wage regressions to estimate race or sex discrimination in labour markets) can be interpreted as an index number, so the theory of characteristics–space indexes has wide applicability. To conserve space, the following is couched in terms of a cost-of-living index but application to other contexts can be made by suitable extensions (Triplett 1983).

The Exact Characteristics–Space Index

A cost-of-living (COL) index shows the minimum change in cost between two periods that leaves utility unchanged. Using (1) and (2), the minimum cost of attaining utility level Q^* in any period is:

$$\begin{aligned} C^* &= C(P_M, h(\cdot), Q^*) \\ &= \min_{C,M} [P_M M + h(c) : Q(c, M) = Q^*] \end{aligned} \quad (11)$$

The form of the cost functional, C, depends on the form of $Q(\cdot)$ and the budget constraint; the hedonic function makes up that portion of the budget constraint that pertains to the acquisition of characteristics. The cost-of-living index between periods r and s is then:

$$\underset{r,s}{\mathrm{COL}} = C\left(P_{Mr}, h(\cdot)_r, Q^*\right) / C\left(P_{Ms}, h(\cdot)_s, Q^*\right) \quad (12)$$

Generally, the full index is intractable, and there is need to consider a less comprehensive measure that is more nearly congruent with the problem at hand. For the separable utility function (4) an exact 'subindex' (Pollak 1975) can be computed that involves only the heterogeneous good. Define the cost functional d by:

$$d = d(h(\cdot), q^*) = \min_c \left[h(c) : q(c) = q^*\right] \quad (13)$$

Then the *characteristics price index* is:

$$I_{r,s} = \left[d(h(\cdot)_r, q^*\right] / \left[d(h(\cdot)_s, q^*\right] \quad (14)$$

where the subscripts designate characteristics costs in period r and s, respectively. Expression (14) is the ratio of the costs, under two characteristics price regimes, of a constant-utility collection of characteristics.

Note that (14) does *not* hold characteristics constant – it is not the price of the same, or 'matched', variety in two periods. Rather, (14) permits substitution among characteristics as relative characteristics costs change, in a manner analogous to the normal COL defined on goods – (14) would be implemented by finding a variety (bundle of characteristics) in period s that was *equivalent* in utility to the one chosen in period r, but which minimized consumption costs in the relative price regime of period s.

Information Requirements

The normal 'goods' COL index requires knowledge of the utility function. The form of the characteristics price index (14) depends on the form of the utility function (or the 'branch' utility function, $q(\cdot)$) *and* the form of the hedonic function, h $(\cdot)$. Both are unobservable or must be estimated. The reason (14) requires more information than the analogous 'goods-space' COL index is that in general the hedonic function is non-linear and therefore its form enters into $d(\cdot)$. In contrast, 'goods' COL indexes assume a bounding hyperplane, whose linearity implies a mirror-image duality between the utility function and the consumption cost function. Use in characteristics space of the demand-systems approaches that have been used to estimate goods–space COL indexes (Braithwait 1980) is complicated by the non-linearity of the hedonic function and by the necessity to estimate both the demand equations and the budget constraint.

Note that, contrary to assertions that have appeared in the literature, imposing 'nice' functional forms (that is, those with properties of classical utility theory) on the hedonic function does nothing to identify index (14) – unless the special case of uniform preferences, where the hedonic function sketches out the characteristics–space preference map, obtains.

Bounds and Approximations: Empirical Hedonic Price Indexes

It is evident that (14) is not an index number that can be computed from the hedonic function alone, so it is not an empirical hedonic price index. It is important to specify the relation of empirical hedonic indexes to (14).

In the usual goods case the budget constraint is assumed a hyperplane. Accordingly, bounds on goods–space COL indexes are fixed-weight (Laspeyres or Paasche) indexes – the denominator of the Laspeyres index, for example, is the equation for the reference period budget constraint, and the numerator is the equation for another budget constraint. Fixed-weight indexes are also convenient approximations to COL indexes, since they require only knowledge of one actual budget constraint and two price regimes, and one knows that the fixed-weight index differs from the true index only by the expenditure saving from substitution.

For characteristics–space price indexes, it is natural to follow an analogous procedure and construct approximations to (14) from the characteristics–space budget constraint, when it is known. The characteristics–space budget constraint is precisely the information provided by the hedonic function. Accordingly, hedonic price index numbers – those computed from hedonic functions – can be interpreted as approximations to the true characteristics–space indexes (14) in the same sense that fixed weight Laspeyres and Paasche price indexes approximate goods COL indexes: The approximations are, in each case, based solely on the budget surface, where the true indexes, in each case, require knowledge of the utility function.

Hedonic indexes differ from goods–space approximating indexes in two major respects. In the characteristics–space case, the form of the budget surface must be estimated empirically. When the hedonic function is linear or is semi-log, the *P*-contours of Fig. 1 are linear – each budget constraint is a hyperplane. Otherwise, the constraints are non-linear. Secondly, and as a corollary, the form of the approximating hedonic index depends on the form of the hedonic function. A third, subsidiary, point is that with usual procedures, the hedonic index records the shift in the whole hedonic surface, rather than, as goods–space fixed-weight indexes are usually calculated, a shift in a single selected budget hyperplane.

Hedonic indexes may also be bounds on the true index, though this interpretation requires more careful empirical specification of the hedonic function than has often been the case, and it is not clear whether they are the *best* bounds. The theory of bounds for characteristics–space indexes is not well worked out.

See Also

- ▶ Characteristics
- ▶ Index Numbers
- ▶ Separability

Bibliography

Blackorby, C., D. Primont, and R.R. Russell. 1978. *Duality, separability and functional structure: Theory and economic applications*. New York: North Holland.

Braithwait, S.D. 1980. The substitution bias of the Laspeyres price index: An analysis using estimated cost-of-living indexes. *American Economic Review* 70: 64–77.

Fisher, F., and K. Shell. 1972. *The economic theory of price indices: Two essays on the effects of taste, quality, and technological change*. New York: Academic.

Gorman, W.M. 1980. A possible procedure for analysing quality differentials in the egg market. *Review of Economic Studies* 47: 843–856.

Griliches, Z. (ed.). 1971. *Price indexes and quality change: Studies in new methods of measurement*. Cambridge: Harvard University Press.

Houthakker, H.S. 1952. Compensated changes in quantities and qualities consumed. *Review of Economic Studies* 19: 155–164.

Lancaster, K. 1971. *Consumer demand: A new approach*. New York: Columbia University Press.

Pollak, R.A. 1975. Subindexes of the cost of living. *International Economic Review* 16: 135–150.

Rosen, S. 1974. Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy* 92: 34–55.

Triplett, J.E. 1983. Concepts of quality in input and output price measures: A resolution of the user value–resource cost debate. In *The U.S. National Income and Product Accounts: Selected topics*, ed. Murray F. Foss, Conference on Research in Income and Wealth, Vol. 47. University of Chicago Press for the National Bureau of Economic Research.

Hedonic Prices

Lars Nesheim

Abstract

Hedonic price functions describe the equilibrium relationships between characteristics of products and their prices. They are used to predict prices of new goods, to adjust for quality change in price indexes, and to measure consumer and producer valuations of differentiated products. They emerge as market outcomes from both competitive and non-competitive markets. The functional form is determined by the distribution of buyers and their preferences, the distribution of sellers and their costs, and the structure of competition in the market.

Keywords

Arbitrage; Assortative matching; Bundling; Cost-of-living indexes; Hedonic preferences; Hedonic price functions; Household production; Imperfect competition; Laspeyres index; Misspecification; Nonparametric methods; Paasch index; Pure strategy Nash equilibrium; Willingness to pay

JEL Classifications

D0; D4

A hedonic price function describes the equilibrium relationship between the economically relevant characteristics of a product or service (or bundle of products) and its price. For example, in a simple labour economics model the hedonic wage function might describe how the wages of a worker depend on education, experience and skill. In a simple housing economics model, the hedonic house price might describe how the price of a house depends on geographic location, size, and quality. In each case, the hedonic price function describes equilibrium (not necessarily competitive) valuations of the economically relevant characteristics of the product.

In empirical applications, statistical estimates of hedonic price functions have primarily been used to calculate quality adjusted price indexes for goods and to measure consumer valuations or producer costs of product characteristics. They have been used to study markets for agricultural products, automobiles, labour, houses, computers, and myriad other differentiated commodities. They have been used to measure quality change in private goods markets and to measure consumer valuations of changes in public goods such as clean air, schools or transport infrastructure. In all these applications, hedonic methods are crucial because the goods in question are not homogenous and their value to buyers and sellers varies systematically with characteristics.

Key questions to be answered when developing a hedonic model to analyse a product market are what are the economically relevant characteristics of the product and what is the market environment that generates the hedonic equilibrium price. Given answers to these questions, a key theoretical goal of hedonic analysis is to determine the theoretical relationship between these market equilibrium prices and underlying structural features of the economy such as producer costs and consumer preferences. Two key empirical goals of hedonic analysis are to understand when statistical estimates of hedonic relationships provide good out-of-sample predictions of prices and to understand what structural information these statistical relationships provide about costs and preferences.

General Hedonic Demand

Hedonic models make various assumptions about whether the space of feasible characteristics is discrete or is a continuum, and whether the characteristics embodied in different products can be bundled or unbundled. This section discusses a general model of hedonic demand that encompasses these special cases. The supply side of the market and various notions of equilibrium are discussed in section "Market Equilibrium".

Each consumer who participates in the hedonic market derives utility from a vector of characteristics $z \in Z_m \subseteq \mathbf{R}^{n_z}$. The bundle z is obtained either by buying a single product that embodies z or by buying a set of products that together produce z. In either case the hedonic cost or price is $p(z)$. The set Z_m is the feasible set given current market conditions. The set Z_m could be a finite set or it could be a continuum. Each consumer also has the option not to participate in the hedonic market, in which case they obtain reservation utility u_0. Assume that characteristics are defined so that utility is increasing in each element of z. Also, assume that utility is decreasing in $p(z)$.

Every consumer is represented by a type $x \in X \subseteq \mathbf{R}^{n_x}$. The space X is the space of all consumer types. The vector x is a vector of consumer characteristics (such as income, education or preference parameters) that affects utility. Consumer heterogeneity is an important feature of hedonic models.

Given hedonic price $p(z)$, consumer x chooses $z \in Z_m$ to maximize utility $u(x, z, p(z))$.) That is, they solve

$$\max_{\{z \in Z_m\}} \{u(x, z, p(z))\}. \tag{1}$$

The solution $z = d(x)$ is the hedonic demand function (or correspondence) for consumer x.

Several features of the model are important. First, z is a complete list of the product characteristics that both affect consumer utility and are known to the consumer at time of purchase. In the housing market example, z could measure geographic location, age of the dwelling, lot size, number of rooms, size of the yard, and so on. Second, there may be additional characteristics of the good that affect *ex post* utility but that are not known to the consumer at time of purchase. In such cases, the utility function should be interpreted as the expected utility from purchasing a good with known characteristics z. Third, buyer utility depends on x and on z. Two consumers, x_1 and x_2, with $x_1 \neq x_2$, will generally choose different bundles $(z_1, p(z_1))$ and $(z_2, p(z_2))$ and will obtain different levels of utility.

Continuous Choice Version

To specialize to the case where Z_m is a compact convex subset of $\mathbf{R}^n$, both u and p are differentiable and the consumer maximization problem has an interior solution, the first-order condition describing the consumer's hedonic demand is

$$\frac{\partial u(x,z,p(z))}{\partial z}+\frac{\partial u(x,z,p(z))}{\partial p}+\frac{\partial p(z)}{\partial z}=0 \quad (2)$$

which can be rewritten as

$$\frac{\partial p(z)}{\partial z}=-\left(\frac{\partial u(x,z,p(z))}{\partial z}\Big/\frac{\partial u(x,z,p(z))}{\partial p}\right). \quad (3)$$

The marginal price at z equals the marginal rate of substitution of the consumer x who chooses z. In the quasi-linear utility case $u(x,z, p(z)) = u(x, z) - p(z)$ and Eq. (3) becomes

$$\frac{\partial p(z)}{\partial z}=\frac{\partial u(x,z)}{\partial z}. \quad (4)$$

These results are the basis for the intuition that the slope of the hedonic price function measures consumers' marginal willingness to pay. Figure 1 illustrates. Consumers x_1 and x_2 optimally choose bundles z_1 and z_2 respectively. At z_1, the marginal price equals the marginal willingness to pay of consumer x_1. However, it is less than the marginal willingness to pay of consumer x_2. At z_2, the marginal price equals the marginal willingness to pay of x_2 but is greater than the marginal willingness to pay of x_1.

The hedonic price function reveals precise information about consumers x_1 and x_2 at points z_1 and z_2 respectively. At all other person-location pairs, it reveals only bounds on willingness to pay. It also reveals very little about how consumers x_1 and x_2 will react to large changes in the shape of the price function. More precise information requires the estimation of consumer preferences.

Discrete Choice Version

If the marginal conditions in (3) and (4) are replaced by inequalities, the qualitative interpretations above apply equally to economies in which Z_m is finite. Suppose there are J elements in Z_m. Let z_j be the j'th element in Z_m and let $p_j = p(z_j)$ for $j = 1, \ldots, J$. In the quasi-linear case, if consumer x chooses z_j, then

$$u(x,z_j)-p_j \geq u(x,z_k)p_k$$

for all $k \in \{1, \ldots, J\}$.

Consider the set of consumers who choose z_j and for whom

$$u(x,z_j)-p_j = u(x,z_k)-p_k \quad (5)$$

for some $k \neq j$. These consumers are indifferent between bundle z_j at price p_j and bundle z_k at price p_k. The difference in prices between z_j and z_k exactly compensates for the difference in utilities. For these indifferent consumers, willingness to pay for z_j over z_k is

$$p_j - p_k = u(x,z_j)-u(x,z_k).$$

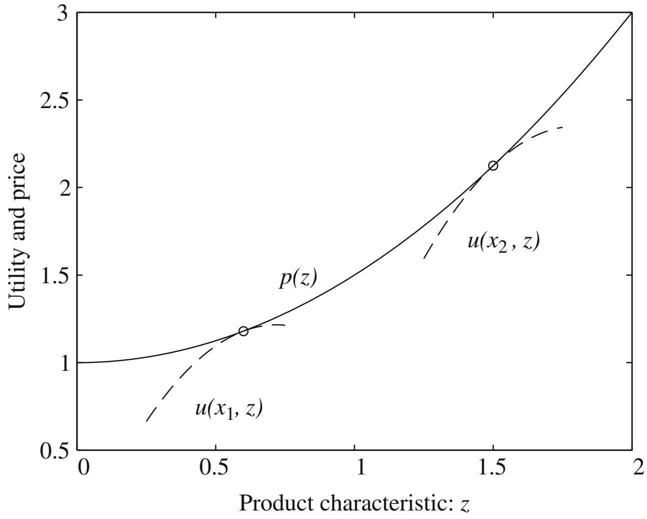
This is the discrete analog of the marginal willingness to pay.

Equation (5) only holds for those who are indifferent between j and k. For those who are not indifferent, the willingness to pay for z_j over z_k is strictly larger than the price. That is

$$u(x,z_j)-u(x,z_k) > p_j - p_k.$$

When the set of available alternatives Z_m is finite, the hedonic price function provides a precise measure of willingness to pay for consumers who are indifferent between options and provides bounds on willingness to pay for consumers who strictly prefer one option to others.

Hedonic Prices, Fig. 1



Single Product Demand Version

In single product demand models, the vector z measures the characteristics of the unique product type that is chosen. These models assume that households cannot buy two separate products with characteristics z_1 and z_2 and combine their characteristics to obtain some other bundle z_3 (Rosen 1974). These models do allow consumers to choose both a product type z and a quantity. To see this, rewrite the utility function in (1) as

$$u(x, z, p(z)) = \max_{\{q\}} \{\tilde{u}(x - qp(z), z, q)\}$$

where q is the quantity of product type z and x is income. This is the primary model used to study location choices and demand for land in urban economic models. See Fujita (1991).

Home Production Version

Home production models assume that consumers purchase a vector of goods in quantities $q \in \mathbf{R}^n_+$ at market prices $\pi \in \mathbf{R}^n_+$ and produce the bundle z from the goods purchased. See Gorman (1980), Lancaster (1966), and Muellbauer (1974). In home production hedonic models, consumers have a technology $f : Z \times \mathbf{R}^n \to \mathbf{R}^m$ describing the production possibility frontier. Given purchases of q units of market goods, any bundle z that satisfies the restriction $f(z, q) = 0$ is feasible.

Given market prices π and technology f, the cost of obtaining the bundle z is

$$p(z) = \min_{\{q\}} \{\pi \cdot \mathrm{q} \ \text{ subject to } \ f(z, q) = 0\} \quad (6)$$

Thus, the hedonic price $p(z)$ is the minimum cost of obtaining bundle z given market prices π and technology f. Given $p(z)$, consumers maximize the utility given in (1). The single-product demand model is a special case of the home production model.

In the Gorman–Lancaster version of the model, the technology is linear and $f(z, q) = z - Aq$ where A is a $n_z \times n_q$ matrix. Each market good contains a fixed quantity of characteristics. The total amount available for consumption is the sum of characteristics across all goods purchased.

Hedonic Cost of Living Index

In each of these models, one can calculate various hedonic cost of living indexes. See Pollak (1989) for details of many alternatives. This section discusses one alternative.

Consider a consumer who purchases a vector of quantities of homogenous goods q with linear prices π and a single differentiated product with characteristics z and hedonic price $p(z)$. When prices are (π, p), the cost of obtaining utility level u_0 is

$$c(\pi, p, u_0) = \min_{\{q,z\}} \{\pi \cdot q \quad \text{subject to} \quad u(q,z) \geq u_0\} \tag{7}$$

If prices change from (π_0, p_0) to (π_1, p_1), then the constant utility hedonic cost of living index is

$$\frac{c(\pi_1, p_1 \ , u_0)}{c(\pi_0, p_0, u_0)}.$$

This cost index hold utility constant and allows consumers to alter consumption of q and z in response to changing prices. When consumer preferences are unknown, this theoretical index cannot be calculated. With data on prices and quantities, empirical alternatives include the Laspeyres index and the Paasch index.

Let (q_0, z_0) solve (7) when prices are (π_0, p_0) in period zero. Let prices in period one be (π_1, p_1). Then a hedonic Laspeyres index is

$$L(q_1, p_1, q_0, p_0, x_0, z_0) = \frac{\pi_1 \cdot q_0 \ + p_1(z_0 \ )}{\pi_0 \cdot q_0 \ + p_0(z_0)} \geq \frac{c(\pi_1, p_1 \ , u_0)}{c(\pi_0, p_0, u_0)}.$$

This index holds the consumption bundle (q_0, z_0) constant at initial levels. Like the standard Laspeyres index, it is an overestimate of the cost of living index because it ignores a consumer's ability to alter consumption in response to changing prices. If some components of z are exogenous (for example, public goods like air quality or public safety), alternative indexes can be defined by including the time varying exogenous elements of z as arguments in the cost function.

One major problem with the index is that the set of available products often changes rapidly over time. If product z_0 is not traded in period 1, then p_1 (z_0) will not be observed. Pakes (2003) shows that an estimate of p_1 (z_0) based on observed prices is an upper bound under certain circumstances. A better option is to calculate the virtual price $p_1^V(z_0)$ that makes the household indifferent between purchasing z_0 at price $p_1^V(z_0)$ and purchasing z_1 (the product actually chosen in period 1) at price $p_1(z_1)$. The virtual price satisfies

$$p_1^V(z_0) = p_1(z_1) - (u(x, z_1) - u(x, z_0)).$$

Data on prices and quantities can be used to bound the virtual price. Precise results require estimation of consumer preferences.

Another major problem is that statistical authorities, as discussed in section "Estimating Hedonic Prices" , do not observe the elements of z that enter consumer preferences. A third major problem is that time constraints and cost constraints place severe limitations on data collection and analysis for use in practical price index calculations. Triplett (2004) provides a comprehensive overview of these issues.

Market Equilibrium

Hedonic prices emerge as equilibrium outcomes from a market environment. They might emerge from a purely competitive environment in which neither buyers nor sellers have power to influence prices or they might emerge from an imperfectly competitive environment in which either buyers or sellers have market power. They may be observed in arms-length transactions or unobserved as in black-market wage contracts or implicit marriage contracts.

In general, the hedonic price function in a market is a nonlinear function of the characteristics z. Its functional form is determined by the distribution of buyers and their preferences, by the distribution of sellers and their costs, and by the type of equilibrium in the market. Special cases exist where more can be said. If bundles of characteristics can be unbundled, arbitrage leads to a linear hedonic price (Rosen 1974). In the Gorman–Lancaster model, the hedonic price function is piece-wise linear (see Pollak 1983, or Heckman and Scheinkman 1987). In the Tinbergen (1956) model, the hedonic price is quadratic. When both buyer utility and seller costs depend on z only through an index $q(z)$, the hedonic price function satisfies $p(z) = \tilde{p}(q(z))$.

Competitive Hedonic Equilibrium

Consider a one-dimensional Tinbergen–Rosen model in which consumers of type $x \in \mathbf{R}_+$ choose $z \in \mathbf{R}_+$. Assume that consumer utility is $u(x, z) =$

$x\tilde{u}(z)$ where $x\frac{\partial \tilde{u}z}{\partial z} > 0$. Note that $\frac{\partial^2 u(x,z)}{\partial z \, \partial x} = \frac{\partial \tilde{u}z}{\partial z} > 0$. Assume that the distribution of consumer types is described by the distribution function $F_x(x)$ with density function $f_x(x)$ and support $\mathbf{R}_+$.

Treat the supply side symmetrically. Assume that firms of type $y \in \mathbf{R}_+$ have costs of producing one unit of product z of $c(y,z) = \frac{\tilde{c}(z)}{y}$ where $\left(\frac{1}{y}\right)\frac{\partial \tilde{u}(z)}{\partial z} > 0$. Note that $\frac{\partial^2 c(y,z)}{\partial z \, \partial y} = \left(\frac{-1}{y^2}\right)\frac{\partial \tilde{c}(z)}{\partial z} < 0$. The distribution function describing the distribution of firms is $F_y(y)$ with density $f_y(y)$ and support $\mathbf{R}_+$.

Given a differentiable price, consumers solve

$$\max_{\{z\}} \{x\tilde{u}(z) - p(z)\}.$$

Assume there is a unique interior optimizer. The consumer first order condition is

$$x\frac{\partial \tilde{u}(z)}{\partial z} - \frac{\partial p(z)}{\partial z} = 0.$$

This equation implicitly defines the buyer demand function $z = d(x)$ and the inverse demand function $x = \tilde{d}(z) = \left(\frac{\partial p(z)}{\partial z} \middle/ \frac{\partial \tilde{u}(z)}{\partial z}\right)$. Note that the consumer second order condition implies that $\partial \tilde{d}(z) > 0$. As a result, the distribution function describing the distribution of demand is $Fx\left(\tilde{d}(z)\right) = F_x\left(\frac{\partial p(z)}{\partial z} \middle/ \frac{\partial \tilde{u}(z)}{\partial z}\right)$.

By the same reasoning, the firms' first-order conditions define the inverse supply function $y = \tilde{s}(z) = \left(\frac{\partial \tilde{c}(z)}{\partial z} \middle/ \frac{\partial p(z)}{\partial z}\right)$ which also is monotonic. As a result the distribution of supply can be written $F_y\left(\frac{\partial \tilde{c}(z)}{\partial z} \middle/ \frac{\partial p(z)}{\partial z}\right)$.

An equilibrium hedonic price function is one that equates the distributions of supply and demand. Formally, a function $p(z)$ is an equilibrium price function if it satisfies the differential equation

$$F_x\left(\frac{\partial p(z)}{\partial z} \middle/ \frac{\partial \tilde{u}(z)}{\partial z}\right) = F_y\left(\frac{\partial \tilde{c}(z)}{\partial z} \middle/ \frac{\partial p(z)}{\partial z}\right) \tag{8}$$

for almost all $z \in Z_m$ and if $p(z_{\min})$ ensures that all buyers and sellers obtain at least their reservation utilities.

Some simple conclusions stem from this analysis. First, since $\frac{\partial^2 u(x,z)}{\partial z \, \partial x} > 0$ and $\frac{\partial^2 c(y,z)}{\partial y \, \partial z} < 0$, the equilibrium involves positive assortative matching between buyers and sellers. Second, the equilibrium price depends on u, the preferences of buyers, c, the costs of sellers, and on F_x and F_y, the distributions of both types of agents. Third, the price function is the envelope of seller cost and buyer utility.

In more general cases and in cases of higher dimension, the differential Eq. (8) often does not have nice numerical properties. However, one can solve the equilibrium problem by solving the associated social welfare maximization problem which is an optimal transportation problem (an infinite dimensional linear programming problem with special structure). Recent results in this area include Gretsky et al. (1999) and Chiappori et al. (2006).

Oligopoly Hedonic Equilibrium

When there is imperfect competition in hedonic markets, firms set prices to maximize profits. Assume individual demand is derived from the discrete choice model in section "Discrete Choice Version". Let $p = (p_1, \ldots, p_J)$ and $z = (z_1, \ldots, z_J)$. Given p and z, let $D_j(p, z, x) \in [0, 1]$ be the demand of consumer x for product j. Let $f_x(x)$ be the density of consumer types with support X.

Aggregate demand for good j is

$$q_j(p,z) = \int_X D_j(p,z,x) f_x(x)dx.$$

Given the strategies of all firms $k \neq j$, firm j solves

$$\max_{\{z_j p_j\}} \{p_j q_j(p,z) - c(j, q_j, z_j)\}.$$

The first order conditions are

$$q_j + p_j\frac{\partial q_j}{\partial p_j} - \frac{\partial c_j}{\partial q}\frac{\partial q_j}{\partial p_j} = 0 \tag{9}$$

$$p_j \frac{\partial q_j}{\partial z_j} - \frac{\partial c}{\partial q_j}\frac{\partial q_j}{\partial z_j} - \frac{\partial c}{\partial z_j} = 0. \tag{10}$$

A pure strategy Nash equilibrium is a set of strategies (z_j, p_j) for each firm $j = 1, \ldots, J$ such that each firm maximize profits given the strategies of its competitors. In a Nash equilibrium, the equilibrium hedonic prices p and characteristics z are determined by the distribution of buyers and their preferences, the costs of the competitors and by the competitive structure of the market. Buyers preferences u and the distribution f_x determine the structure of demand. This demand structure combined with the costs of competitors and the number of competitors determine the fierceness of competition. See, for example, Berry et al. (1995).

Estimating Hedonic Prices

Ideal Case: z Is Perfectly Observed

The theory of hedonic prices places no restrictions on the hedonic price functional form. The lack of theoretical predictions has led to controversy about functional form in empirical hedonic price work. Different researchers have used linear models, log linear models, Box–Cox models, and fixed-effect models. To estimate hedonic quality adjustments for use in price indexes, many statistical authorities adopt the even more restrictive ‘time-dummy’ model in which the hedonic price function takes the form

$$p_t = \beta_0 + \beta_1 z_{1t} + \beta_2 z_{2t} + \beta_3 \cdot D_t + \varepsilon_t \tag{11}$$

where D_t is a vector of time dummies. See Triplett (2004) for a detailed discussion. This version restricts the hedonic price function to be linear in characteristics and to have coefficients that are constant over time. The time-dummy model is rarely theoretically justified and the constant coefficient restriction is usually rejected in empirical tests. Nevertheless, Triplett (2004) argues that in many cases of interest to statistical authorities the restriction works as an approximation and does not make much empirical difference for estimates of hedonic price indexes.

There is no theoretical justification for restrictive parametric empirical models of hedonic prices unless prior knowledge of the market and the products traded exists to support the restrictions. When data-sets are large and the dimension of z is small, there is little empirical justification for parametric models either. In such cases, hedonic price functions should be estimated nonparametrically unless prior knowledge sufficient to restrict the model exists. Such nonparametric regressions can be easily estimated on desktop computers.

When sample size is small or the dimension of z is large, however, then unrestricted nonparametric methods are often impractical. In these cases, prior information should first be used to impose structure on the hedonic relationship. In some cases, it is then feasible to use semiparametric methods to estimate the hedonic relationship without imposing further structure. In many (if not most) cases, however, there is no choice but to impose further structure that is supported neither by data nor by theory. If the primary use of the method is to predict prices out-of-sample, then goodness of fit and stability with respect to changing market conditions can be useful criteria to choose functional form. If the primary use is to estimate marginal willingness to pay in some dimension, then semiparametric methods that allow for flexibility in the dimension of interest might be of most use. Tests for robustness should be implemented and interpretations of results should consider potential misspecification biases.

Practical Case: z Is Imperfectly Observed

Empirical estimates of hedonic price functions may be biased due to omitted variables or mismeasured variables. Assume the goal is to estimate the hedonic price $p(z)$ and that the methods used will rely on estimation of conditional expectations. Discussion of estimation of $\ln(p(z))$ or methods based on other statistics such as the median would proceed along similar lines.

Let $z = (z_1, z_2)$ be the set of all hedonic characteristics and let $\tilde{z} = (\tilde{z}_1, \tilde{z}_2)$ be the set of variables that the econometrician observes. Assume that z_1 is observed without error so that $\tilde{z}_1 = z_1$. Assume that $\tilde{z}_2$ is a vector of proxy variables (or instrumental variables) and that $z_2 = g(\tilde{z}_2; \varepsilon_2)$ where ε_2 is a vector of unobservables. Let $p(z_1, z_2)$ be the theoretical hedonic price function. Observed prices $\tilde{p}$ satisfy

$$\tilde{p} = p(\tilde{z}_1, g(\tilde{z}_2, \varepsilon_2)) + \eta \qquad (12)$$

where η is measurement error, $E(\eta) = 0$, and η is assumed independent of $(\tilde{z}_1, \tilde{z}_2, \varepsilon_2)$. The unobserved characteristic case, is the case where $g(\tilde{z}_2, \varepsilon_2) = \varepsilon_2$ and $f_{\varepsilon_2}(\varepsilon_2 | \tilde{z}_1, \tilde{z}_2) = f_{\varepsilon_2}(\varepsilon_2 | \tilde{z}_1)$. Then ε_2 is the unobserved characteristic of the product.

Under these assumptions, the expectation of $\tilde{p}$ conditional on $(\tilde{z}_1, \tilde{z}_2)$ is

$$\begin{aligned} & E(\tilde{p} | \tilde{z}_1, \tilde{z}_2) \\ & = \int p(\tilde{z}_1, g(\tilde{z}_2, \varepsilon_2)) f_{\varepsilon_2}(\varepsilon_2 | \tilde{z}_1, \tilde{z}_2) d\varepsilon_2 \\ & = (\tilde{z}_1, \tilde{z}_2) h \end{aligned} \qquad (13)$$

where $f_{\varepsilon_2}(\varepsilon_2 | \tilde{z}_1, \tilde{z}_2)$ is the density of ε_2 conditional on $(\tilde{z}_1, \tilde{z}_2)$. This is the best predictor (in the integrated squared error sense) of $\tilde{p}$ given data on $(\tilde{z}_1, \tilde{z}_2)$. However, in general $h(\tilde{z}_1, \tilde{z}_2) \neq p$ $(\tilde{z}_1, \tilde{z}_2)$ and little can be said about the relationship between the two without more information.

Researchers have employed instrumental variables techniques or prior information that places structure on g, on p, or on $f_{\varepsilon 2}$ to cope with this problem. See Chay and Greenstone (2005) and Bajari and Benkhard (2005) for examples.

Estimating Hedonic Preferences

In most cases, the full set of consumer characteristics that affect choices is not observed. The econometrician observes only a subset of consumer characteristics such as education, income, age, and household structure. For example, suppose the consumer has two characteristics (x, ε) and x is observed while ε is not. Recall the consumer first order condition

$$\frac{\partial p(z)}{\partial z} = \frac{\partial u(x, \varepsilon, z)}{\partial z}. \qquad (14)$$

This equation defines the hedonic demand function $z = d(x, \varepsilon)$.

When data on (x, z, p) are available, u cannot be estimated directly using (14) because z is an endogenous variable. As in Fig. 1 where households with different values of x choose different value of z, households with different values of ε will choose different values of z.

Additional restrictions can help identify u. Ekeland et al. (2004) show that the utility function can be identified nonparametrically if $\frac{\partial u}{\partial z}$ is additively separable. That is if,

$$\frac{\partial u(x, \varepsilon, z)}{\partial z} = u_0(x) + u_1(z) + \varepsilon$$

where u_0 and u_1 are arbitrary nonparametric functions.

More generally, Heckman et al. (2005) prove that the demand function d (x, ε) can be estimated using data on (z, x) alone if ε is statistically independent of x. They further show that the function u is not identified with data from a single market unless prior information is used to restrict u. For example, if marginal utility is weakly separable so that $\frac{\partial\, u(x, \varepsilon, z)}{\partial\, z} = v(q(z, x), \varepsilon)$ where q is a known function, then the function v can be estimated.

Heckman et al. (2005) also show how to use multi-market data to estimate the unrestricted Eq. (14). Because cross-market variation in prices is tied to cross-market variation in the distributions of buyers and sellers, it is functionally independent of within market variation in z and x. As a result, this cross-market variation in prices can then be used to identify and estimate the function u.

See Also

- ▶ Compensating Differentials
- ▶ Household Production and Public Goods
- ▶ Inflation Measurement
- ▶ Location Theory
- ▶ Non-parametric Structural Models

Bibliography

Bajari, P., and C. Benkhard. 2005. Demand estimation with heterogeneous consumers and unobserved product characteristics: A hedonic Approach. *Journal of Political Economy* 113: 1239–1276.

Berry, S., J. Levinsohn, and A. Pakes. 1995. Automobile prices in market equilibrium. *Econometrica* 63: 841–890.

Chay, K., and M. Greenstone. 2005. Does air quality matter? Evidence from the housing market. *Journal of Political Economy* 113: 376–424.

Chiappori, P.-A., R. McCann, and L. Nesheim. 2006. Nonlinear hedonic pricing and matching models: Finding equilibria through linear programming. Manuscript.

Ekeland, I., J. Heckman, and L. Nesheim. 2004. Identification and estimation of hedonic models. *Journal of Political Economy* 112: S60–S109.

Fujita, M. 1991. *Urban economic theory: Land use and city size*. Cambridge: Cambridge University Press.

Gorman, T. 1980. A possible procedure for analysing quality differentials in the egg market. *Review of Economic Studies* 47: 843–856.

Gretsky, N., J. Ostroy, and W. Zame. 1999. Perfect competition in the continuous assignment model. *Journal of Economic Theory* 88: 60–118.

Heckman, J., R. Matzkin, and L. Nesheim. 2005. *Nonparametric estimation of nonadditive hedonic models*, Working Paper No. CWP03/0 5. London: CeMMAP, Institute for Fiscal Studies.

Heckman, J., and J. Scheinkman. 1987. The importance of bundling in a Gorman–Lancaster model of earnings. *Review of Economic Studies* 54: 243–255.

Lancaster, K. 1966. A new approach to consumer theory. *Journal of Political Economy* 74: 132–156.

Muellbauer, J. 1974. Household production theory, quality, and the hedonic technique. *American Economic Review* 64: 977–994.

Pakes, A. 2003. A reconsideration of hedonic price indexes with an application to PC's. *American Economic Review* 93: 1578–1596.

Pollak, R. 1983. The treatment of 'quality' in the cost-of-living index. *Journal of Public Economics* 20: 25–53.

Pollak, R. 1989. *The theory of the cost-of-living index*. New York: Oxford University Press.

Rosen, S. 1974. Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of Political Economy* 82: 34–55.

Tinbergen, J. 1956. On the theory of income distribution. *Weltwirtschaftliches Archiv* 77: 155–173.

Triplett, J. 2004. Handbook on hedonic indexes and quality adjustments in price indexes: Special application to information technology products. OECD Science Technology and Industry Working Papers, 2004/0 9. OECD Publishing. doi:10.1787/643587187107.

Hedonism

James Griffin and Derek Parfit

From the Greek *hedone*, 'pleasure', this term is used of two different theses, one a psychological thesis about motivation (psychological hedonism), the other a thesis about what is intrinsically valuable in a person's life (ethical hedonism).

Psychological hedonism refers to the claim that a person acts solely to promote his own pleasure. (It is usual to limit the scope of the claim to acts that meet some minimal standards of rationality, that is, excluding confused, involuntary or habitual acts).

Ethical hedonism refers to the claim that only pleasure is intrinsically valuable, that all other things that are valuable are so only instrumentally as means to pleasure. This root form of ethical hedonism is not strictly an ethical view at all (so the name is something of a misnomer); it is rather a view about what constitutes the quality of an individual life. This root form is often combined with a view about action, namely that all action should aim at maximizing pleasure. This combination can easily turn into a view about rationality (e.g. that a rational agent acts to maximize his own pleasure – what could be called egoistic hedonism), or into a view about morality (e.g. that each person should act to maximize pleasure for persons generally – universalistic hedonism).

It is hard to supply a satisfactory analysis of the concept of 'pleasure' or to understand its relation to 'happiness'. Pleasure is not a physical sensation (think of the pleasure of country walks), nor a psychological one (think of the pleasure from one's work). What we enjoy is so heterogeneous that no unified account of 'pleasure' may be possible. Sidgwick (1907), aware of how different our states of minds can be when we enjoy ourselves, thought that the unifying feature was desire: 'pleasure', he proposed, is 'desirable consciousness'. J.S. Mill thought that the relation between ‘pleasure’ and ‘happiness’ was relatively simple: he defined ‘happiness’ as ‘pleasure and the absence of pain’. But the terms mark different features of life: a martyr might go to the stake happily, but is unlikely to do so with pleasure. The lack of breadth of physical and psychological explanations of ‘pleasure’ has led to more behavioural ones: for example, what we find pleasant is what we do, or would, give ourselves to eagerly. Similarly, ‘pain’ applies to more than just physical pains, and the painfulness even of physical pains is a matter not only of our sensations but also of how we react to them. So the need to get a sufficiently broad analysis of ‘pain’ has also led to behavioural explanations; for example what we find painful is what we wish to avoid, have alleviated, etc. The terms ‘pleasure’ and ‘pain’ sometimes become so broad in the course of the statement or defence of either psychological or ethical hedonism that they become, in effect, technical terms. When that happens, we have to ask what their technical sense is. For example, near the end of his life Sigmund Freud refused strong pain-killing drugs, preferring, he said, to think in torment than to be confused in comfort. We might wish to use ‘pleasure’ in such a way that we should say that Freud found clear thought in pain more ‘pleasant’ than confused pleasure. But we might also think that the technical sense of the word would now be so far removed from the ordinary sense that it would be better to find another way of speaking altogether.

Psychological hedonism is an empirical thesis, and widely thought to be false. As Butler, Hume and others point out, one’s actions are often explained by desires for things other than one’s own pleasure or avoidance of pain (e.g. the desire to eat is often more effective than the desire for pleasure from eating). Perhaps the clearest counter-examples are the desires that many people have about what will happen after their deaths, when (they assume) they will not exist. A psychological hedonist might reply that what really explains such actions are desires for the pleasure of knowing how things will go after one’s death. However, a simple thought experiment would often show this to be false. Suppose that a father working to provide for his family after his death is told that he can choose between (1) his family’s being provided for though he will never know it, and (2) his family’s not being provided for though he will think that they will be, and that, after he makes his choice, he will be made to forget it. Many persons would choose (1) rather than (2), thereby showing that their action is not prompted by any concern for their own future mental states. (The falsity of psychological hedonism leads to the *paradox of hedonism*, namely, that typically one cannot promote one’s own pleasure by aiming to promote it, that pleasure is usually an unintended accompaniment of action aimed at another goal.)

The root form of ethical hedonism is a thesis about what affects the quality of a single life. It too is widely disputed. Even on Sidgwick’s broad account of ‘pleasure’ as ‘desirable consciousness’, anything that affects the quality of life must enter consciousness. But we desire, when fully informed, things other than states of consciousness and, moreover, we seem to regard them as making our lives better. For instance, we may desire a good reputation among people we do not know, or posthumous fame, or to do something important with our lives. An ethical hedonist may object that these desires are irrational, but this cannot be plausibly claimed about all of these desires. And we have many desires of this kind. This point is made forcefully by Robert Nozick with a piece of science fiction, his *Experience Machine* (a variant on earlier Pleasure Machines). Suppose we could plug into a machine that would give us any state of consciousness we desired. Would we plug in? What could matter except how life feels from the inside? Most people would respond that they want not just the experience of helping their children or doing something important, but also actually to do these things; most people also want to be in touch with reality, even at the cost of some desirable consciousness. This suggests that the notions of ‘quality of life’ or ‘well-being’ cannot be understood entirely in hedonistic terms, even when ‘hedonism’ is generously defined.

In the last two centuries ethical hedonism has figured most prominently as the value theory of classical utilitarianism. However, it is not essential to utilitarianism; other value theories can be, and are, substituted for it. Modern economists, having borrowed the conceptual framework of utility maximization from classical utilitarians, largely ignored hedonism in favour of a more neutral value theory: what is valuable is the fulfilment of desire, where it is left open what the objects of desire may be (pleasure, no doubt, being one but not necessarily the only one). Modern philosophers in this tradition are divided; some wish to stay in the hedonist tradition, at least broadly interpreted, and insist that only states of consciousness affect the quality of a life, while others drop that requirement and prefer to develop the notion of the fulfilment of informed desire.

Bibliography

A classic discussion of these issues is Henry Sidgwick, *The Methods of Ethics*, 7th edn, Bk I, chs 4 and 9; Bk II, chs 1–4; Bk III, ch. 14; Bk IV, ch. 1. For a good discussion of Sidgwick's views, see Schneewind (1977), ch. 11. For modern discussions, see Brandt (1979), ch. 13; Edwards (1979); Parfit (1984), Appendix I; Griffin (1986), Pt I.

Brandt, R.B. 1979. *A theory of the good and the right*. Oxford: Clarendon Press.

Edwards, R.B. 1979. *Pleasures and pains*. Ithaca: Cornell University Press.

Griffin, J. 1986. *Well-being*. Oxford: Clarendon Press.

Parfit, D. 1984. *Reasons and persons*. Oxford: Clarendon Press.

Schneewind, J.B. 1977. *Sidgwick's ethics and victorian moral philosophy*. Oxford: Clarendon Press.

Sidgwick, H. 1874. *The methods of ethics*, 7th edn. London: Macmillan, 1907.

Hegelianism

R. P. Bellamy

The origins and concerns of the political ideas of the German philosopher G.W.F. Hegel (1770–1831) are traditionally thought to be religious rather than economic. However, a preoccupation with issues of political economy is present in his earliest theological writings and lies at the centre of his wider philosophical project (Hegel 1793–1800). Broadly speaking, Hegel wished to construct an ethical theory appropriate for the specific problems of the modern world. He believed ancient and medieval societies had been bound together by a communal code of behaviour, with social roles mirroring a putative natural or divine order. The harmony of the natural macrocosm and the social microcosm had been sundered in modern societies by a growing awareness of individuality on the part of their members. Hegel traced this development to two sources: the primacy accorded to the individual conscience within Christianity, especially the Lutheranism he personally espoused, and the individualism encouraged by the capitalist mode of production. Contrary to recent influential critics (e.g. Popper 1945), Hegel did not wish to stifle individual liberty by returning to the organic community theorized by Plato. Instead, he sought to describe the conditions necessary for the freedom of each person to be compatible with the freedom of all.

Hegel traces the development of this consciousness of subjective freedom in the *Phenomenology* (1807) and the *Lectures on World History* (1822–30). He regards the symbol of Christ, of the divine present within humankind, as emblematic of this sense of personal freedom and simultaneously the death of any notion of a transcendent God standing outside of human existence. The individual becomes the fount and locus of all value, confronting a material world with judgements he or she has chosen and endowing it with meaning. This process is given substance through human labour and the physical transformation of nature to suit human purposes, an idea Hegel borrows from Locke. Drawing on the stadial model of economic development advocated by the political economists of the Scottish Enlightenment, particularly Sir James Steuart and Adam Smith, he went on to elaborate how this new ethic had spawned a completely new civilization.

Commercial society broke the old ties of dependence of agrarianism and feudalism by freeing humanity from a subordination to nature. Humans no longer live in a created world, but create their own environment. However, the exchange economy generates new social bonds by involving individual producers within mutual service relationships. ‘Civil Society’ (*burgerliche Gesellschaft*), according to Hegel, is united by a ‘system of needs’ (Hegel 1821, para. 189). The division of labour reduces our self-sufficiency and makes us dependent on others for the provision of our wants. Production too becomes a cooperative venture, both in the interests of efficiency and because more specialized skills are required. As our technical ability to create new commodities increases, so does the complexity of our needs. The labour process becomes ever more subdivided and the interrelationships deriving from mutual services more intricate. Hegel regards these developments as double-edged. On the one hand, he fully embraces the classical liberals’ praise of market society as increasing individual liberty. To a certain extent he endorses their claim that the interrelatedness of the system of needs makes it self-regulating. Such duties as the obligation to obey promises, notions of fair exchange, bans on stealing etc. . . . emerge within civil society itself, and he agrees with Hume that certain criteria of justice derive from the mutual self-interest of property owners in conditions of scarcity. He appropriately locates police functions within civil society. On the other hand, he does not believe that the needs of the market alone can lead to a well-ordered community. Hegel points to two potential sources of instability. First, he expands the insights of Smith, Adam Ferguson and Schiller’s *Letters on the Aesthetic Education of Man* into the ennervating and alienating effects of modern industrial labour. With Smith’s famous pin-factory example in mind, he notes the stupefying and mechanical nature of factory work, predicting that it will ultimately be taken over by machines. Second, he foresaw capitalism’s propensity for periodic crises of overproduction. The business classes’ uncontrolled pursuit of conspicuous consumption leads them to produce more goods than there are consumers. The bottom falls out of the market and workers, who, because of the extent of the division of labour, rely entirely upon this single commodity for their employment, will lose their livelihood. This group becomes ‘a rabble of paupers’ (Hegel 1821, para. 244) outside of society and unprovided for by Humean economic justice.

Hegel gave the problem of poverty considerable thought and he dwells on it in a number of writings. He suggests two solutions, state charity funded by taxation and the direct creation of employment by state interference in the economy. He rejected the latter as merely exacerbating the problem, since overproduction was its root cause. The former, whilst more appealing, is equally inadequate. He notes that poverty is relative as well as absolute, and that charity can therefore create a stigma which increases the inferiority of the recipients and undermines their self-respect. Hegel’s solution was to introduce a political dimension into social decision making. He parts company with classical political economists here, maintaining that our understanding of the true nature of society is incomplete as long as we remain within the restricted perspective of the market mentality. Like the more mercantilist Steuart, he contends that an awareness of our mutual obligations can grow through membership of occupational associations (*Korporation*) and social groups (*Stände*). He advocated a system of indirect democracy, whereby representatives from these bodies are sent to a national assembly which can enact social legislation. Hegel maintained that participation within these institutions would moderate the individualist self-seeking which led to economic crises. People would appreciate their mutual debts, implicit in capitalist production, and alter their behaviour accordingly to further the common good of the whole community.

Some commentators have regarded this solution as a sleight of hand (Avinieri 1972; Plant 1977). Following Marx, they regard Hegel as having correctly expounded the contradictions of capitalist society, but assert he has merely carried them into the political sphere by enfranchising functional groups rather than individuals. Ending poverty requires the radical restructuring of productive relations demanded by communism (Marx 1843). Hegel failed to make this step because he limited the philosopher’s task to understanding society rather than changing it. However, Hegel's purpose was to preserve modern individuality. For him Marxism would have represented an unacceptably anachronistic return to the organic communities of the past.

Liberals also dispute Hegel's political response. They accuse him of subverting liberty by imposing a corporate mentality upon the free transactions of individuals within society. This misunderstanding of Hegel's intentions stems from their view of the relation of society to the state. Whereas liberals regard the state as merely providing the minimal means necessary for our pursuit of our private projects, without fear of undue hindrance from others, Hegel defines it in terms of certain shared ethical norms presupposed by all our activities. The public sphere is not the outcome of individual choices but what is presumed by them, the medium within which they are formed. This is the ethical life or *Sittlichkeit* of a community, which the state represents and upholds.

Whilst the corporatist policies of fascist authors, such as Giovanni Gentile (1875–1944), seem to justify the fears of both liberal and Marxist writers, others have understood him better. The British idealists in particular, such as T.H. Green (1836–82) and Bernard Bosanquet (1840–1923), shared his concern with poverty and suggested schemes for the state regulation of industry, education and poor relief which provided the intellectual origins for later proposals for the welfare state. Like Hegel, they regarded social and political institutions as instrumental in fostering an awareness of the complex of mutual rights and duties necessary for the adoption of such policies. They were similarly ambivalent about the degree to which poverty arose from a weakness of will on the part of the poor or social conditions. Nevertheless, an unresolved paradox persists in Hegel's theory. He claims community is an unconscious presupposition of maximizing individuals in commercial society, but it is not at all obvious how the market would operate once people become conscious of this fact and adopt the community-minded behaviour Hegel believed they would. Clearly civil society would then be thoroughly politicized; whether or not with the dire consequences liberals fear, or in a self-contradictory manner as Marx opined, is beyond the competence of this article to judge.

See Also

▶ Dialectical Materialism

Bibliography

Avinieri, S. 1972. *Hegel's theory of the modern state*. Cambridge: Cambridge University Press.

Bosanquet, B. 1899. *The philosophical theory of the state*. London: Macmillan.

Chamley, P. 1963. *Economie politique et philosophie chez Steuart et Hegel*. Paris: Presses Universitaires de France.

Gentile, G. 1946. *Genesis and structure of society*. Trans H.S. Harris. Urbana: University of Illinois Press, 1960.

Green, T.H. (1878–80). *Lectures on the principles of political obligation,* ed. Paul Harris and John Morrow. Cambridge: Cambridge University Press, 1986.

Hegel, G.W.F. 1793–1800. *Early theological writings*. Trans. T.M. Knox. Chicago: Chicago University Press, 1948.

Hegel, G.W.F. 1807. *The phenomenology of spirit*. Trans. A.V. Miller. Oxford: Oxford University Press, 1977.

Hegel, G.W.F. 1817–19. *Die Philosophie des Rechts: Die Mitschriften Wannenman (Heidelberge 1817/18) und Homeyer (Berlin 1818/19),* ed. K.-H. Itling. Stuttgart: Kelt-Cotta, 1983.

Hegel, G.W.F. 1818–31. *Vorlesungen über Rechtsphilosophie, 1818–31,* 4 vols, ed. K.-H. Itling. Stuttgart–Bad Cannstatt: Frommann-Holzboog, 1973–4.

Hegel, G.W.F. 1821. *Philosophy of right*. Trans. T.M. Knox. Oxford: Oxford University Press, 1958.

Hegel, G.W.F. 1822–30. *Lectures on the philosophy of world history: Introduction*. Trans. H.B. Nisbet. Cambridge: Cambridge University Press, 1975.

Marx, K. 1843. *Critique of Hegel's philosophy of right*. Trans. J. O'Malley. Cambridge: Cambridge University Press, 1970.

Pelczynski, Z.A. (ed.). 1984. *The state and civil society: Studies in Hegel's political philosophy.* Cambridge: Cambridge University Press.

Plant, R. 1977. Hegel and political economy. *New Left Review* 103: 79–92; 104: 103–113.

Popper, K. 1945. *The open society and its enemies*, vol. 2. London: Routledge & Kegan Paul.

Vincent, A., and R. Plant. 1984. *Philosophy, politics and citizenship: The life and thought of the British idealists*. Oxford: Blackwell.

Heilbroner, Robert L. (1919–2005)

William Milberg

Abstract

Robert Heilbroner was among the most popular historians of economic thought in the 20th century and a prominent critic of neoclassical economics and free-market capitalism. His *The Worldly Philosophers* explained how the great economists struggled to understand Western capitalism's rapid economic growth and accompanying inequities and social tensions. Heilbroner's probing 'scenarios' of capitalism's future drew mainly from the works of Smith, Marx and Schumpeter. His insistence that economic issues are integrally tied to moral and psychological concerns gave his work a rare depth and spoke to the political nature of all social thought.

Keywords

Budget deficits; Capital accumulation; Capital budgeting; Capitalism; Economic determinism; Globalization; Heilbroner, R.; History of economic thought; Labour power; Lowe, A.; Marx's analysis of capitalist production; Neoclassical economics; Private property; Profit; Schumpeter, J.; Smith, A.; Socialism; Wage relation

JEL Classifications

B31

One of the most prominent critics of the economics profession and of free-market capitalism, Robert Heilbroner was also responsible for motivating generations of college students to become economists. *The Worldly Philosophers: The Life, Times and Ideas of the Great Economists*, Heilbroner's classic treatment of the history of economic thought, captivated generations of readers with its elegantly written, witty, and probing discussions of how these thinkers struggled to understand Western capitalism's rapid economic growth, and industrialization and its accompanying inequities and social tensions. First published in 1953, *The Worldly Philosophers* is in its seventh edition, has been translated into 22 languages, and remains one of the best-selling books on economics of all time. Heilbroner went on to publish 25 books and over 100 articles on the history of economics and the future of capitalism, focusing at various times on the role of the state, big business, technology, morality, psychology, private property and power. His influence went well beyond the academy, as his books were written in an accessible style and elucidated issues of concern to a broad public. He was a regular contributor to *The New Yorker* and *The New York Review of Books*, and for years served on the editorial board of the interdisciplinary journals *Dissent* and *Social Research*.

Robert Heilbroner was born in New York City in 1919, attended Horace Mann School for Boys and graduated summa cum laude from Harvard University in 1940. He worked briefly for the Office of Price Administration in Washington, D. C., before serving in the army as an interpreter in Japan in the Second World War. After the war Heilbroner came back to New York and worked as a freelance writer while he studied at the New School for Social Research. Invited to join the economics faculty at the New School, Heilbroner was granted a doctorate from the New School for his already published book *The Making of Economic Society*. Heilbroner spent his entire career at the New School for Social Research, where he helped build the programme in political economy that remains to this day one of the few Ph.-D. programmes in the United States which emphasizes heterodox economics and the history of economic thought.

Heilbroner was a democratic socialist, as critical of authoritarian Soviet socialism as of dogmatic, free-market capitalism. In lucid prose, Heilbroner conveyed the consequences for everyday life of the deep and seemingly abstract economic forces which create and distribute income and wealth. His identification of these forces as embedded in politics and culture reinforced their everyday relevance.

Economics in Context

The purpose of economics, Heilbroner wrote, was ‘to give meaning to economic life’. Such meaning, he argued, is necessarily forward looking: ‘There is a deep human need to be situated with respect to the future ... to rescue us from a conception of social existence as all contingency and chance’ (1990a, p. 1112). Heilbroner believed that any effort to understand contemporary society required a serious consideration of the history of ideas and societies. Like his mentor Adolph Lowe, Heilbroner relied heavily on the insights of Smith, Marx and Schumpeter in his own efforts to analyse such large questions as the prospects for socialism, the viability of capitalism, and such problems as the trend of dangerous environmental degradation or the inequalities raised by the globalization of production and finance. He described these three economists as ‘great scenarists’, not because any of their long-run predictions proved right – mostly, he admitted, they were wrong – but because they provided ‘a plausible framework within which to face that most fearsome of psychological necessities – looking into the future’. Heilbroner considered these scenarios to be ‘the most significant accomplishment of economics’ (1995, pp. 5–6).

Heilbroner insisted on understanding capitalism as a particular stage in the long history of human efforts to solve the ‘economic problem’ of material provisioning and social reproduction. Knowledge of how different societies have confronted these problems gives crucial perspective to our own efforts to do so today. Thus the starting point for understanding contemporary economic life is to identify the distinguishing features of the current economic system: capitalism. Modern economics, Heilbroner argued, has largely avoided this first crucial step, ignoring rather than illuminating the rich array of social, psychological and moral forces that propel capitalist societies. ‘[B]ehind the veil of conventional economic rhetoric’, he wrote in a short autobiographical essay (2000, p. 287), ‘we can easily discern an understructure of traditional behavior – trust, faith, honesty, and so on – as a necessary moral foundation for a market system to operate, as well as a concealed superstructure of power.’ Heilbroner noted with outrage that even the word ‘capitalism’ had disappeared from the economics textbooks. He saw economics as an ‘explanation system’ of capitalism, and insisted on the relevance of economics to large questions of political economy – the role of the state, the sustainability of environmental health, the problem of world poverty and the danger of nuclear war – rather than small questions of optimal allocation under conditions of scarcity.

Capitalism’s Nature and Logic

Rejecting neoclassicism, Heilbroner turned to Smith and Marx for the central building blocks of social analysis, since both identify a logic of capitalist development which explains capitalism’s endurance and its inherent limitations. Marx is especially important because of his focus on the particularity of the capitalist drive for the expansion of wealth and the exigencies of power, politics and psychology brought on by this accumulation drive. As Heilbroner elaborated in a series of books written in the 1980s – *The Nature and Logic of Capitalism*, *Marxism For and Against*, and *Behind the Veil of Economics* – all efforts to solve the economic problem of material provisioning, be they organized by tradition, command or markets, are aimed at the production of a material surplus above the needs of subsistence. Only in capitalism, however, does this take a general form – self-expanding value. The commodity is but a way station towards the accumulation of value in ‘a never-ending metamorphosis of M-C-M′’ (1988, p. 37). This circuit of capital hinges on the institution of private property in the means of production, which ‘organizes and disciplines’ society and serves as an instrument of power ‘because its owners can establish claims on output as their quid pro quo for permitting access to their property' (1988, p. 39). Profit goes to owners of capital and not only validates the activities of particular owners but perpetuates the M-C-M′ circuit. As Heilbroner writes, 'Profit is for capitalism what victory is for a regime organized on military principles. ..' (1988, p. 41).

The wage relation is crucial to understanding capitalism's uniqueness. Workers are free to offer their labour power for wages, unlike forced labour in many traditional and especially feudal and slave societies. But private property relations keep workers from retaining the full value of their efforts. A second unique feature of capitalism is the distinctiveness of its private and public realms, each relying on the other for its sustenance. Capitalism thus has a unique political agenda in that the precise role and scope of the state vis-à-vis the private sector is constantly contested and debated. Despite the freedom embodied in the wage relation and the reliance of the private sector on the state, capitalism has functioned under both democratic and anti-democratic political systems. Heilbroner himself was an outspoken advocate for an active role for government in creating a decent society and productive economy. In *The Debt and the Deficit: False Alarms, Real Possibilities* (1989, co-authored with Peter Bernstein) Heilbroner argued for Keynesian deficit spending and capital budgeting by the US government.

For all its identifiable deep structures and logic, capitalism for Heilbroner is constantly changing, buffeted by other social forces. This is partly the result of history's dialectical nature: as problems are resolved through social change, the new conditions present a new set of problems. In *The Future as History* and *An Inquiry into the Human Prospect* Heilbroner focused on various implications of this unsettling aspect of social reality, in particular long-run environmental consequences of economic development. Capitalism is all the time and everywhere contingent on independent ideas, political struggles and ethical dilemmas, and these have resulted in a variety of capitalisms around the world. For Heilbroner, theories of economic determinism – be they Marxist or neoclassical – that reduce capitalism to a system of markets cannot adequately explain social change, since economics, politics and morality are linked: '[T]he engines of history do not draw all their energies from economic drives and institutions. If socialism failed, it was for political, more than economic reasons; and if capitalism is to succeed it will be because it finds the political will and means to tame its economic forces' (1996, p. 195).

The Paradox of Progress

If it was Marx who best articulated the 'nature and logic' of capitalism, it was Smith who provided the most important insights into the psychology of individuals in capitalist society. Heilbroner considered Smith's *Theory of Moral Sentiments* to be as important as *The Wealth of Nations* and insisted that the socialized individual of *Theory of Moral Sentiments* was not only consistent with but necessary to the successful working of the nascent capitalism described in *The Wealth.* Smith's writings on empathy and, most importantly, subservience ('the principle of authority') and the drive for self-betterment are the psychological foundations of the 'society of perfect liberty'; that is, they are the psychological dynamics that make capitalism function. Like the other classical economists, Smith also emphasized capitalism's dark side. Capitalism's advance brings unprecedented wealth creation and the possibility of 'perfect liberty'. It also brings stagnation, poverty, inefficiency, systemic corruption and moral decay. The result was what Heilbroner termed 'the paradox of progress'. For Heilbroner these insights were important not only for students of intellectual history but also for those seeking to understand the prospects for capitalism today. 'Capitalism's uniqueness in history', he wrote in *Twenty-First Century Capitalism*, 'lies in its continuously self-generated change, but it is this very dynamism that is the system's chief enemy' (1993, p. 130). Both of these insights – the embeddedness of the economy in a broader social, political and psychological fabric, and the inherent problems in capitalist development – Heilbroner attributes to Adam Smith, although they are not part of the canonical reading of Smith as a proponent of laissez-faire.

Smith's influence on Heilbroner went beyond the issue of the psychology of individual agents in capitalism and into the existential question of the purpose of theory itself. In his *Essays on Astronomy* (published in 1758 and excerpted in Heilbroner, 1986), Smith wrote that '[T]he repose and tranquility of the imagination is the ultimate end of philosophy. . .Philosophy, by representing the invisible chains which bind together all these disjointed objects, endeavors to introduce order in this chaos of jarring and discordant appearances, to ally this tumult of the imagination.' Discussing this passage, Heilbroner wrote that 'We theorize ... to restore our peace of mind' (1986, p. 16).

Analysis and Vision in Economics

Heilbroner's embrace of the classicals and rejection of the neoclassicals hinged on the Schumpeterian distinction between 'analysis' and 'vision'. Schumpeter (1954, p. 41) defined vision as the 'preanalytic cognitive act' that is inevitable and ideological. Analysis is the largely deductive process that follows from the theory's foundations. As intellectual historian, Schumpeter separated economic analysis from its vision, leading to the *History of Economic Analysis*, published posthumously. Heilbroner embraced the Schumpeterian categories, and especially the notion that vision is an inevitable part of the process of theorizing, since 'All systems of thought that describe or examine societies must contain their political character, knowingly and explicitly, or unknowingly and in disguise.' (1990b, p. 109) But Heilbroner resisted Schumpeter's separation of vision and analysis, since connecting the two allowed a greater appreciation of how economic scenarios are formed. *The Worldly Philosophers* was enormously popular not only because it included juicy biographical details about the early economists but because Heilbroner revealed the lively imagination and political engagement of the 'great scenarists'. Vision, for Heilbroner, embodied much of the creativity that informs economic problem-solving and modelling. And it is through the vision that ethical and epistemological principles are brought into theory. Vision is the expression 'of the inescapable need to infuse 'meaning' – to discover a comprehensive framework – in the world' (1990a, p. 1112). For Heilbroner, it was precisely the persistent denial of the role of vision that leaves modern economics so limited as a tool for understanding social life. In *The Crisis of Vision in Modern Economic Thought*, Heilbroner (and co-author William Milberg) developed this theme in the context of contemporary debates in macroeconomics.

In the final chapter of the seventh edition of *The Worldly Philosophers*, Heilbroner wrote of 'the end of economics', playing on the dual meaning of end as both purpose and termination. For all its technical sophistication, modern economics has largely failed to accomplish the purpose of a worldly philosophy: to give meaning to economic life. Heilbroner saw the narrowness of modern economics as an abandonment of the grand aspiration for social thought that Smith, Ricardo, Malthus, Marx, Mill, Keynes and Schumpeter each held in their day. Heilbroner lamented, 'The new vision is Science, the disappearing one capitalism' (1953, p. 314). Heilbroner was an important public intellectual of the second half of the 20th century. While he remained to his last days a severe critic of modern economics, his personal warmth, his kindness, his humaneness, his commitment to equality, opportunity and democracy, and his love of deep and serious debate on pressing social issues endeared him to a broad group of professional economists, social scientists, students and a socially-concerned public.

See Also

- ▶ Capitalism
- ▶ Marx's Analysis of Capitalist Production
- ▶ Schumpeter, Joseph Alois (1883–1950)
- ▶ Smith, Adam (1723–1790)
- ▶ Wealth

Selected Works

1953. *The worldly philosophers: The lives and times of the great economists*, 7th edn. New York: Simon and Schuster, 1999.

1959. *The future as history*. New York: Harper and Brothers.

1963. (With W. Milberg.) *The making of economic society*, 12th edn. Englewood: Prentice-Hall, 2006.

1974. *An inquiry into the human prospect*, rev. edn. New York: W.W. Norton, 1980.

1980. *Marxism: For and against*. New York: W.W. Norton.

1982. The socialization of the individual in Adam Smith. *History of Political Economy* 143, 427–439.

1985. *The nature and logic of capitalism*. New York: W.W. Norton.

1986. *The essential Adam Smith*. New York: W. W. Norton.

1988. *Behind the veil of economics*. New York: W.W. Norton.

1989. (With P. Bernstein.) *The debt and the deficit: False alarms/real possibilities*. New York: W.W. Norton.

1990a. Analysis and vision in the history of modern economic thought. *Journal of Economic Literature* 28, 1097–1114.

1990b. Economics as ideology. In *Economics as discourse*, ed. W. Samuels. Boston: Kluwer Academic.

1993. *Twenty-first century capitalism*. New York: W.W. Norton.

1995. (With W. Milberg.) *The crisis of vision in modern economic thought*. New York: Cambridge University Press.

1996. *Teachings from the worldly philosophy.* New York: W.W. Norton.

2000. Robert Heilbroner. In *A biographical dictionary of dissenting economists*, 2nd edn, ed. P. Arestis and M. Sawyer. Aldershot: Edward Elgar.

Bibliography

Schumpeter, J. 1954. *History of Economic Analysis*. New York: Oxford University Press.

Helfferich, Karl (1872–1924)

K. Schmidt

Helfferich was an economist with a particular expertise in currency problems; at times, he was also a civil servant, a banker and a politician. He was born in Neustadt/Palatinate and died in a railway accident in Bellinzona, Switzerland. Helfferich studied in Munich, Berlin and Strasbourg, where he took his PhD (1894). In the heated discussion during the years between 1895 and 1901 over whether or not Germany should stay with the gold standard or move to bimetallism, he fought vigorously for the former position. In 1899 he became a lecturer at the University of Berlin. From 1901 to 1906 he was in the Colonial Department of the Foreign Office in charge of currency and transport matters in the German colonies of that time. He then joined the Deutsche Bank, first in a high position in Istanbul and later as director in Berlin. Early in 1915 Helfferich became the secretary of state in the German Treasury Office. In financing the war, he made recourse far less to additional taxes than to borrowing, including borrowing from the Reichsbank – a method which was strongly criticized later on because of its inflationary consequences. In the following year Helfferich took the same post in the Office of the Interior, from which he resigned one year later. From 1920 until his death, Helfferich was a member of the German Reichstag and strongly influenced the policy of the Deutschnationale Volkspartei.

As a scholar he taught and wrote primarily on monetary and currency matters. But he also did some substantial work in other economic fields, such as trade policy, national income and wealth, and in politics. His most important scientific publication is the book *Das Geld,* which between 1903 and 1923 went into six editions (English edn, 1927). It was one of the best textbooks of its time covering in a very systematic way historical, theoretical, organizational and political issues.

Most important was Helfferich's role in the German currency reform of 1923. It was he who invented the idea by introducing an auxiliary currency (originally the *Roggenmark,* finally the *Rentenmark)* to provide for a stable-value legal tender as well as the stabilization of the Mark. This combination of aims and the restoration of confidence in the new currency by making the *Rentenmark* redeemable in *Rentenbriefe* (which were issued on the basis of the agricultural and industrial property) were the decisive conditions for the success of the currency reform of 1923. In 1923 Helfferich was recommended for the presidency of the Reichsbank, but for political reasons the opportunistic Dr Schacht was given the position.

Selected Works

1898. *Die Reform des deutschen Geldwesens nach der Gründung des Reiches*, 2 vols. Leipzig: Duncker & Humblot.
1901. *Handelspolitik.* Leipzig: Duncker & Humblot.
1903. *Des Geld*, 6th ed. Leipzig: Hirschfeld, 1923. English edition as *Money;* trans. L. Infield and ed. with Introduction by T.E. Gregory. London: Benn, 1927; New York: Adelphi, 1927.
1913. *Deutschlands Volkswohlstand 1888/1913*, 5th ed. Berlin: Stilke, 1923.

Bibliography

von Lumm, K. 1926. *Karl Helfferich als Währungspolitiker und Gelehrter*. Leipzig: Hirschfeld.
Reichert, J.W. 1929. Helfferich, Karl. In *Handwörterbuch der Staatswissenschaften,* Supplement to 4th ed. Jena: Fischer.

Heller, Walter Perrin (1942–2001)

Ross M. Starr

Keywords
American Economic Association; Arrow, K.; Baumol–Tobin money demand model; Consumption function; Coordination failure; Council of Economic Advisers; Demand for money; General equilibrium; growth, models of; Heller, W.P.; Money and general equilibrium; Non-clearing markets in general equilibrium; Stanford Institute for Theoretical Economics; Walter, W.W.

JEL Classifications
B31

Walter Perrin Heller was a leading 20th-century economic theorist, and an early member of the University of California, San Diego, faculty (from 1974 to his death in 2001). He annually taught the UCSD graduate core microeconomic theory course on welfare economics.

Heller came from an academic family distinguished in the economics discipline: his father, Walter W. Heller, was Professor of Economics at the University of Minnesota and served as chairman of the President's Council of Economic Advisers in the Kennedy and Johnson US presidential administrations. Walter P. Heller's undergraduate education took place at Oberlin College and at the University of Minnesota, particularly under the guidance at Minnesota of Professor Leonid Hurwicz (1990 recipient of the US National Medal of Science). Heller's intellectual home was Stanford University. He received his Ph.D. there in 1970 with the dissertation advice of Nobel Prize winner Kenneth J. Arrow. For three decades he participated in the Stanford summer economic theory workshop at the Institute for Mathematical Studies in the Social Sciences (IMSSS) and its successor, the Stanford Institute for Theoretical Economics (SITE). Prior to joining the UCSD faculty, he was on the economics faculty of the University of Pennsylvania.

Heller served as an associate editor of the *Journal of Economic Theory* and on the executive committee of the American Economic Association. His research treated the stability of economic growth, microeconomic foundations of macroeconomics and of the demand for money, and resource allocation under conditions of market failure due to incompleteness or monopoly. In the late 1980s and the 1990s, the research focused on a fundamental issue in the theory of unemployment, namely, *coordination failure*, or the inability – even of complete markets in price equilibrium – successfully to match supply and demand, workers and employers. Kenneth Arrow remarked at Heller's memorial service at Stanford on 16 July 2001:

> Economic theory backed by serious mathematical reasoning was just beginning to be recognized when Walt started his graduate work...Walt was one of the leaders in using new ways – not merely for clarification – but for changing the way the economy was considered. He contributed to many aspects of [economic] theory . . . His long-standing project of studying the coordination failures of the economic system brought out, in an essentially novel way, the previously unclarified meaning of Keynesian insights. This work . . . is a vital continuing part of modern economic thought . . .

Stability of economic growth: A growth model over time in general competitive equilibrium (at each instant) may nevertheless be on an intertemporally inefficient path (Hahn 1966, 1968; Malinvaud 1953). Further, an efficient path may be unstable (Samuelson and Solow 1956). Heller (1971, 1975) demonstrated that inefficiency and instability depend on myopia; in the presence of complete intertemporal capital markets (futures markets for capital), stability and efficiency of the growth path are established.

Demand for money: Heller was among the first to apply the full formal structure of an Arrow–Debreu model to the analysis of a monetary economy (1972; 1974; 1976 with R. Starr). The Baumol–Tobin money demand model with transaction costs (Tobin 1956) is shown to be consistent with full general competitive equilibrium.

Foundations of macroeconomics: The Keynesian consumption function was long recognized anecdotally to be a result of capital market imperfection, but Heller and Starr (1979b) represents the first mathematical formalization of this notion. Unemployment equilibrium was long thought inconsistent with Walrasian general equilibrium pricing; Heller and Starr (1979a) demonstrate that expectations of uncleared markets may be self-fulfilling in equilibrium even at competitive equilibrium prices.

Coordination failure: When the formation of markets is itself a resource using activity, then some markets may not form or announce prices in equilibrium (1986; 1992; 1999) with resulting inefficiency and unemployed resources. In a model with a non-competitive (oligopoly or monopoly) sector, even with a full set of markets, there may be multiple Pareto ranked equilibria (1998).

Heller's work is elegantly written so that the underlying intuition is clear and is supported by mathematical structure. The Walter P. Heller Prize for excellence in research – instituted by Heller's colleagues – is awarded annually to a UCSD graduate student.

See Also

- ▶ Arrow, Kenneth Joseph (Born 1921)
- ▶ General Equilibrium
- ▶ Heller, Walter Wolfgang (1915–1987)
- ▶ Money and General Equilibrium
- ▶ Non-clearing Markets in General Equilibrium

Selected Works

1971. Disequilibrium dynamics of competitive growth paths. *Review of Economic Studies* 38, 385–400.

1972. Transactions with set-up costs. *Journal of Economic* Theory 4, 465–78.

1974. The holding of money balances in general equilibrium. *Journal of Economic Theory* 7, 93–108.

1975. Tatonnement stability of infinite horizon models with saddle-point instability. *Econometrica* 43, 65–80.

1976. (With R. Starr.) Equilibrium with non-convex transactions costs: Monetary and non-monetary economies. *Review of Economic Studies* 43, 195–215.

1979a. (With R. Starr.) Unemployment equilibrium with myopic complete information. *Review of Economic Studies* 46, 339–59.

1979b. (With R. Starr.) Capital market imperfection, the consumption function, and the effectiveness of fiscal policy. *Quarterly Journal of Economics* 93, 455–63.

1986. Coordination failure under complete markets with applications to effective demand. In

Essays in honor of Kenneth J. Arrow. volume 2, equilibrium analysis, ed. W. Heller, R. Starr and D. Starrett. New York: Cambridge University Press.

1992. Underemployment as a coordination problem with savings and increasing returns. In *Economic analysis of markets and games: Essays in honor of Frank Hahn*, ed. P. Dasgupta. et al. Cambridge MA: MIT Press.

1998. (With A. Edlin and M. Epelbaum.) Is perfect price discrimination really efficient?: Welfare and existence in general equilibrium. *Econometrica* 66, 897–922.

1999. Equilibrium market formation causes missing markets. In *Markets, information, and uncertainty: Essays in economic theory in honor of Kenneth J. Arrow*, ed. G. Chichilnisky. New York: Cambridge University Press.

Bibliography

Hahn, F. 1966. Equilibrium dynamics with heterogeneous capital goods. *Quarterly Journal of Economics* 80: 633–640.

Hahn, F. 1968. On warranted growth paths. *Review of Economic Studies* 35: 175–184.

Malinvaud, E. 1953. Capital accumulation and efficient allocation of resources. *Econometrica* 21: 233–268.

Samuelson, P., and R. Solow. 1956. A complete capital model involving heterogeneous capital goods. *Quarterly Journal of Economics* 70: 537–562.

Tobin, J. 1956. The interest elasticity of the transactions demand for cash. *Review of Economics and Statistics* 38: 241–247.

Heller, Walter Wolfgang (1915–1987)

Joseph A. Pechman

Keywords

American Economic Association; Council of Economic Advisers; Full-employment surplus; Germany, economics in; Heller, W. W.; Potential gross national product; Progressive taxation; Public finance; Revenue sharing; Tax expenditures

JEL Classifications

B31

Heller was born in Buffalo on 27 August 1915. He grew up in Seattle and Milwaukee, and graduated from Oberlin College. He received a doctorate in economics from the University of Wisconsin, where he studied with Harold M. Groves, who greatly influenced a generation of public finance scholars. He spent his entire academic career as professor of economics at the University of Minnesota.

Heller made important scholarly contributions to the study of public finance, but his major claim to fame was his highly successful term as chairman of the Council of Economic Advisers under Presidents John F. Kennedy and Lyndon B. Johnson from 1961 to 1964. After leaving the government, he was influential as a consultant and adviser to presidents, Congress and business. He wrote widely on current economic developments, tax policy, and state-local finance, and was also known as a stimulating lecturer and commentator on economic policy issues. In 1974, he served as president of the American Economic Association.

Heller began his professional career as an expert on state and local taxation. He wrote his doctoral dissertation on the administration of state income taxes, and later originated the idea of federal revenue sharing with the states and local governments. The details of revenue sharing were developed by a task force appointed by President Johnson, but it was enacted by Congress only after it was recommended by President Richard M. Nixon in 1972. The revenue sharing legislation was extended until the end of September 1986.

During the Second World War, Heller moved to the Treasury Department, where he contributed to the development of tax policy to finance the war. In 1947–8, he was tax adviser to the US Military Government in Germany, where he played an important role in designing the currency and fiscal reforms that helped launch the post-war German economic revival. He also served as a consultant to the Treasury Department during the late 1940s and early 1950s. He has been a strong advocate of progressive taxation and was one of the first to recognize that unnecessary deductions and tax preferences narrow the income tax base, require higher marginal tax rates to raise the necessary revenues, and distort economic decisions.

As chairman of the Council of Economic Advisers, Heller supported innovative macroeconomic policies to promote economic growth and stability. He persuaded President Kennedy to propose a major tax cut to stimulate demand, advocated the enactment of an investment tax credit and liberalized depreciation allowances to increase investment incentives. His Council developed the first, and most successful, voluntary wage–price guidelines to help contain inflationary pressures as the economy moved to full employment.

Heller's Council pioneered fiscal analysis based on the concepts of potential gross national product – the output the economy would produce at full employment – and the full-employment surplus. It is also noted for its advocacy of the neoclassical Keynesian synthesis of fiscal and monetary policies required to achieve full employment and increase economic growth. To reach full employment, it proposed the use of stimulating budget and monetary policies. To increase growth at full employment, it stressed the need for a full-employment surplus and monetary ease to support private investment in plant and equipment, combined with public investments in education, research, and development. It also urged the dismantling of barriers to free trade among nations to achieve the benefits of international specialization and exchange.

As a result of the policies pursued by the Kennedy and Johnson administrations, the nation enjoyed a long period of economic growth and prosperity without inflation. From the fourth quarter of 1960 to the fourth quarter of 1964 (when Heller left his CEA post), US real GNP grew at an average annual rate of 4.9 per cent, consumer prices rose 1.2 per cent a year, and long-term federal bond yields never exceeded 4.2 per cent.

Heller combined his advocacy of sound economic policies with an understanding of the need to help the disadvantaged and underprivileged. He helped to persuade President Johnson to design and implement an anti-poverty programme to provide economic opportunities for low-skilled workers and a decent income for those who cannot earn their own livelihood. 'We cannot relax our efforts to increase the technical efficiency of economic policy', he wrote in 1966. 'But it is also clear that its promise will not be fulfilled unless we couple with improved techniques of economic management a determination to convert good economics and a great prosperity into a good life and a great society.'

Selected Works

1952. Limitations of the federal individual income tax. *Journal of Finance* 7(2): 185–202.

1959. (With C. Penniman.) *State income tax administration*. Madison: University of Wisconsin Press.

1966. *New dimensions of political economy*. Cambridge, MA: Harvard University Press.

1968. A sympathetic reappraisal of revenue sharing. In *Revenue sharing and the city*, ed. H.S. Perloff and R.P. Nathan. Baltimore: Johns Hopkins University Press.

1969. *Monetary vs. Fiscal Policy* (a dialogue with Milton Friedman). New York: W.W. Norton.

1975. What's right with economics? *American Economic Review* 65: 1–26.

1976. *The economy: Old myths and new realities*. New York: W.W. Norton.

1982. Kennedy economics revisited. In *Economics in the public service: Papers in honor of Walter W. Heller*, ed. J.A. Pechman and N.J. Smiler. New York: W.W. Norton.

Helvetius, Claude Adrien (1715–1771)

J. Wolff

There are, for Helvetius, a certain number of fundamental points: the individual is led, spontaneously, to seek pleasure and to avoid pain, and this engenders self-esteem; having realized what his needs are, the search for objects able to satisfy them determines his behaviour; personal interest governs his decisions, and these vary as interests do according to individuals, the social environment and the era; education, custom and environment form the whole man; men would be equally happy if they could fill all the different moments of their lives agreeably.

The question to ask is if, and how, one can guarantee general happiness. To maintain universal contentment there has to be a reciprocal dependence between all the members of society; that is to say, they should all be ‘equally’ occupied, or work should be ‘equally’ divided amongst them. For this to be the case, there must not be too large an inequality of wealth, condemning some to deprivation and excessive work whilst others are corrupted by luxury. This is all the more true today, as people almost everywhere are divided into two classes, one of which lacks necessities whilst the other has too much and consequently grows bored.

The only way to proceed is greatly to increase the number of landowners and therefore to redistribute land. This is always a difficult step to take as it constitutes a violation of a sacred right, the right of ownership.

It is the government which, to a large extent, is responsible for the happiness of the individual. It can and must ‘mould’ men and take every possible measure to secure for them the equality of happiness which is their right. It must endeavour to reduce the wealth of some and increase that of others, ensure that the poor have property and combat concentrations of wealth by means of taxation and laws of succession. This would only be possible by making very gradual changes. Moreover, the legislator could, by means of a wise education, show men that they can be happy without being equally rich.

Helvetius was Farmer-General from 1738 to 1751, that is, one of the financiers entrusted by the monarchy with the task of collecting tax by means of outright payment. He was also one of the Encyclopédists, a group which included Diderot, d’Alembert and d’Holbach.

He was influenced by Locke. He preached the right to rational criticism in all matters. For him nothing is innate, everything is acquired. The individual is the integral product of his environment and circumstances, which is a sort of rudimentary materialism.

His first book, *De l’ésprit,* was condemned and burnt in 1759. Thereafter it was re-edited several times in London and Amsterdam, and was illegally brought into France, where it was widely read. Helvetius’ ideas had an extremely important influence on Bentham and on the formation of utilitarianism; he was also to influence J.S. Mill and Beccaria in Italy. He was translated into German and read in Russia, and praised by Marx for having emphasized the determining role of social conditions in the development of humanity. Curiously, it could be said that his thought has been forgotten during the last sixty years.

Selected Works

1758. *De l’ésprit.* Paris.

1795. *De l’homme, de ses facultés intellectuelles et de son éducation.* (Posthumous.)

The complete works of Helvetius were published by Editions Didot, Paris, in 1795 and by Editions Lepetit, Paris, in 1818.

Bibliography

Garaudy, R. 1948. *Les sources françaises du socialisme scientifique*. Paris: Editions Sociales.

Horowitz, I.L. 1954. *C. Helvetius*. New York: Paine, Whitman.

Keim, H. 1907. *Helvetius – sa vie et son oeuvre*. Paris: Alcan.

Lichtenberger, H. 1895. *Le socialisme au XVIIIème siècle*. Paris: Alcan.

Henderson, Alexander (1914–1954)

Alan Peacock

Abstract

Described by J.R. Hicks as one of the most brilliant students in Cambridge in the 1930s, Henderson's professional career was cut short by a long spell of war service (1940–45) and by his early death at the age of 39. He held professorial appointments at the University of Manchester (1949–50) and Carnegie Institute of Technology, Pittsburgh (1951–4). His major journal articles were in the field of microeconomics, the best known being a note in the *Review of Economic Studies* (1941) which markedly influenced Hicks's well known exposition of the meaning and measurement of consumer's surplus. Of more lasting interest, perhaps, is his development of public utility pricing theory in respect of the case where marginal cost pricing theory would require a public enterprise to make a loss. He argued that as a loss would have to be covered by a tax, the problem became one of choosing the 'best' tax or combination of taxes. Taxes were labelled 'good' if (a) they ensured that once an investment in a public enterprise had taken place it would be used by all who would be willing to pay the marginal cost; (b) they ensured that an investment would not be undertaken if its cost exceeded consumers' surplus; and (c) they would place the burden where political preferences would wish it to be put, meaning that if the distribution of income was optimal before the investment were undertaken, any tax should be levied on the users of the product of the investment. Applying these criteria he was able to make some trenchant criticisms of established views of the financing of public enterprise investment, notably concerning the two-part tariff system. He wrote on population problems, international trade and took part in the somewhat arid debate on the welfare effects of direct versus indirect taxes during the 1940s. He also co-authored with Charnes and Cooper (1953) one of the best known earlier texts on the application of linear programming to economic problems.

Described by J.R. Hicks as one of the most brilliant students in Cambridge in the 1930s, Henderson's professional career was cut short by a long spell of war service (1940–45) and by his early death at the age of 39. He held professorial appointments at the University of Manchester (1949–50) and Carnegie Institute of Technology, Pittsburgh (1951–4). His major journal articles were in the field of microeconomics, the best known being a note in the *Review of Economic Studies* (1941) which markedly influenced Hicks's well known exposition of the meaning and measurement of consumer's surplus. Of more lasting interest, perhaps, is his development of public utility pricing theory in respect of the case where marginal cost pricing theory would require a public enterprise to make a loss. He argued that as a loss would have to be covered by a tax, the problem became one of choosing the 'best' tax or combination of taxes. Taxes were labelled 'good' if (a) they ensured that once an investment in a public enterprise had taken place it would be used by all who would be willing to pay the marginal cost; (b) they ensured that an investment would not be undertaken if its cost exceeded consumers' surplus; and (c) they would place the burden where political preferences would wish it to be put, meaning that if the distribution of income was optimal before the investment were undertaken, any tax should be levied on the users of the product of the investment. Applying these criteria he was able to make some trenchant criticisms of established views of the financing of public enterprise investment, notably concerning the two-part tariff system. He wrote on population problems, international trade and took part in the somewhat arid debate on the welfare effects of direct versus indirect taxes during the 1940s. He also co-authored with Charnes and Cooper (1953) one of the best known earlier texts on the application of linear programming to economic problems.

Selected Works

1941. Consumer's surplus and the compensating variation. *Review of Economic Studies* 8: 117–121.

1947. The pricing of public utility undertakings. *Manchester School of Economics and Social Studies* 15: 223–250.

Bibliography

Charnes, A., W.W. Cooper, and A. Henderson. 1953. *An introduction to linear programming*. New York: Wiley.

Henderson, Hubert Douglas (Later Sir Hubert) (1890–1952)

E. A. G. Robinson

Henderson was born of a Scottish family. Educated at Rugby School and Cambridge, he began his university studies as a not very successful mathematician but then changed over to economics and at once found his metier. He was placed in the first class with Dennis Robertson and two others in 1912, at a time when Cambridge economics had become a very lively school, very much in the hands of a younger generation, with Pigou as a very young professor and Maynard Keynes, Walter Layton and Ryle Fay as active young lecturers.

Like most of them, Henderson was drawn off into wartime activities. Unfit for military service, he was first in the Board of Trade and subsequently in the Cotton Control Board, whose history he later wrote. After the war he retired to Cambridge with a fellowship at Clare College, lecturing ostensibly on monetary problems but in practice, to the enjoyment of my own generation of undergraduates, on the economic problems of the moment. In this period he wrote the small book *Supply and Demand*, which for thousands of English students during the following 30 years was their first introduction to economics. But he was never by choice an economic theorist and in later life apt to be out of touch with the latest theoretical developments.

In the Cambridge of the 1920s Henderson, with Keynes and others, was in the thick of the re-thinking of Liberal economic policies with Lloyd George as figurehead. When a group of Liberals acquired in 1923 the weekly *Nation and Athenaeum* Henderson, with Keynes as his chairman, became its editor. For the next seven years the *Nation* under his editorship was compulsive reading for every political economist. He might discuss with Keynes, but it was always Henderson who wrote. This, it seems clear in retrospect, was the peak of his career and the job he did best.

In 1930 Henderson was persuaded to give up the *Nation* to become the chief economist of the Economic Advisory Council, then newly created by Ramsay MacDonald's Labour government. He was faced by an impossible task at an impossible time. Britain, saddled by Winston Churchill's decision when Chancellor of the Exchequer to return to the prewar gold standard, was struggling with the hopelessly inconsistent tasks of deflating to achieve that and simultaneously expanding to overcome a mountain of unemployment. It was not Henderson's fault that despite the ingenuities of Keynes and the debates of countless committees they failed to do so. It was the fault of a generation of politicians who could not be persuaded to grasp the nettle. But these years of frustration left Henderson a different man. He was no longer the crusading optimist. He had become the eternal critic, with a duty to ensure that no one should ever overlook any possible difficulties of any proposed source of action.

The outbreak of war in 1939 found him a member with Lord Stamp and Henry Clay of a committee to examine the war plans of government departments and more generally the problems of the war economy. When, soon after Churchill became Prime Minister in 1940 this came to an end, Henderson was absorbed into the Treasury with no very specific responsibility. For his period there he was the arch-critic, always engaged, as has been said, in detecting difficulties, and something of a discouragement to those who were trying to design policies for a better world. Administration, the achievement of consensus around the best practicable answer, was not his role.

In 1944 he was offered and gladly accepted a special research fellowship at All Souls College, Oxford; a year later he was elected to the long-established Drummond Professorship of Political Economy in the University of Oxford. He was back in the atmosphere in which he was completely happy. He could forget the problems of consensus. He could be right in a minority of one. He had enthusiastic undergraduate audiences to hear his views on the interwar years. He was by now out of touch with the theories, not only of his very able younger Oxford colleagues but also of Keynes and his own contemporaries. But in the vigorous argument of an Oxford common room he had few equals. Shortly before his death early in 1952 he had been elected Warden of All Souls. He did not live to take up the office.

Hennipman, Pieter (1911–1994)

Arnold Heertje

Keywords

Austrian economics; Chamberlin, E. H; Hennipman, P; *Homo economicus*; Interpersonal utility comparisons; Monopoly; Pareto optimal redistribution; Pierson, N. G; Robbins, L. C; Tinbergen, J; Welfare economics; Unanimity; Wicksell, J. G. K

JEL Classifications

B31

Dutch economist born in Leiden, 12 September 1911, who died in Amsterdam on 3 July 1994. Hennipman belongs to the three most important economists of the Netherlands, the two others being Nicolaas Gerard Pierson (1839–1909) and Jan Tinbergen (1903–1994). He studied at the Faculty of Economics of the University of Amsterdam, and was taught economic theory by H. Frijda and economic history by N.W. Posthumus. He took his Master's degree in 1934, and in 1938 became reader in economics at the University of Amsterdam, next to his beloved teacher Frijda. He continued his work on his dissertation and received his doctorate in July 1940, in time to enable Frijda, who soon after had to flee from the Nazis, to act as his director of his thesis. Of Hennipman's impressive work on economic motive and economic principle a much-enlarged edition appeared after the Second World War in 1945. The book presents a detailed historical-critical survey of the manifold varieties of *homo economicus*, concluding that the scope of economics is not restricted to the behaviour of such an animal. It is argued that the concept of economic welfare is subjective and devoid of specific content and that economics cannot be normative. His work shows the influence of the Austrian subjectivist way of thinking and Lionel Robbins' *Essay* (1932). In 1945 Hennipman became Professor of Economics at the University of Amsterdam.

From 1945 to 1972 Hennipman was managing editor of the Dutch Journal *De Economist*, nowadays published in English. Many articles have appeared which reveal evidence of his vast knowledge of the literature and demonstrate his ability to encourage authors to improve their manuscripts by his constructive and well-founded comments. In 1951, invited by E.H. Chamberlin, Hennipman participated in a conference held by the International Economic Association on monopoly, competition and their regulation. Hennipman's paper 'Monopoly: Impediment or Stimulus to Economic Progress?' received great praise at the conference from J.M. Clark, G. Haberler, F.H. Knight and F. Machlup (Hennipman 1954). In 1962 he published his essay on the theory of economic policy, of which a shortened version in English is published in Hennipman's book *Welfare*

Economics and the Theory of Economic Policy (Hennipman 1995). The analysis builds further on his dissertation by applying to the theory of economic policy the principles set out in his work on economic motive and economic principle. This essay is without doubt one of the highlights of non-mathematical economic literature. He contributed to the publication of the Walras correspondence, edited by W. Jaffé in 1965 (Jaffé 1965).

Following his retirement in 1973 Hennipman was very active on methodology, the history of economic thought and, in particular, welfare economics. Publications during the last decade of his life mainly concern welfare economics: for example, a pair of articles exploring the historical and analytical relations between Pareto optimality and Wicksellian unanimity. A major theme in Hennipman's work is the contention that welfare economics is a non-normative theory, as he convincingly spelled out in major debates with Ezra Mishan and Mark Blaug (Hennipman 1995). Interpersonal comparisons of utility and Pareto optimal redistribution are discussed by Hennipman from this point of view.

It is only due to his incredible and miraculous modesty that the international audience of economists had to wait until after his death for an accessible publication of his work in English. This event also explains that his influence on the development of international economics literature fell short of what would have been justified by the high quality and relevance of his contributions, which are innocent of mathematics and do not reflect empirical research. He influenced both students and professors by allowing them access to his vast knowledge of almost all areas of economic theory and his analytical insights. There is no doubt that he became the leading Dutch economist, in particular since the war, albeit still in the shadow of Jan Tinbergen, who had built his international reputation during the years of the Great Depression of the 1930s.

See Also

▶ Chamberlin, Edward Hastings (1899–1967)
▶ Pierson, Nicolaas Gerard (1839–1909)
▶ Robbins, Lionel Charles (1898–1984)
▶ Tinbergen, Jan (1903–1994)

Selected Works

1945. *Economisch Motief en Economisch Principe*. Amsterdam: North-Holland.

1954. Monopoly: Impediment or stimulus to economic progress? In *Monopoly and competition and their regulation*, ed. E.H. Chamberlin. London: Macmillan.

1995. *Welfare economics and the theory of economic policy*, ed. Walker, D., Heertje, A., and H. van den Doel, with an introduction by D. Walker. Aldershot: Edward Elgar.

Bibliography

Jaffé, W. 1965. *Correspondence of Léon Walras and related papers*, vol. 3 vols. Amsterdam: North-Holland.

Robbins, L.C. 1932. *An essay on the nature and significance of economic science*. London: Macmillan.

Herfindahl Index

William G. Shepherd

This is one form within the species of 'comprehensive' indices of market structure, and is used in industrial economics to suggest the degree of monopoly power. The Hirschman–Herfindahl Index (HHI) is the sum of the squared values of all firms' market shares in a given market. If shares are measured from 0 to 1.0, the HHI ranges from minimal to 1. If the shares are taken as per cent values from 0 to 100, then the HHI ranges from minimal to 10,000.

The index first acquired the name of Orris C. Herfindahl (an energy economist) in the 1950s, but Albert O. Hirschman used the index earlier in assessing foreign trade patterns, hence the dual name. Its users note that it is comprehensive, while the standard concentration ratio covers only the leading firms. The ratio gained a certain technical vogue in the 1980s, but has not displaced concentration ratios as the mainstream basis for estimating the degree of market power.

The HHI presents three problems. First, as a pure number it lacks content. Users must translate it into equivalent 'real' concentration ratios, in order to convey its possible meaning. Thus a 1000–2000 HHI range has no intrinsic meaning. It is (*very*) roughly comparable to four-firm concentration ratios of 50 to 80, and that is the way in which the ratios have come to be evaluated. Second, the HHI's data requirements are heavy. If market shares are known for individual firms, those details are the key facts to use, rather than to submerge them in a single index. Finally, the weighting of shares by an exponent of 2 (or any other specific value) has no basis in theory or empirical patterns. As one result, the upper ranges of market shares give very high HHI values (e.g. a firm with 70 per cent of the market has, by itself, an HHI of 4900). Such high numbers may correctly reflect an extreme degree of monopoly power held by dominant firms, but the issue has not been researched. HHI users have preferred to look only at oligopoly patterns in the lower HHI ranges of 1000 and 2500.

Other comprehensive indexes ('entropy', 'numbers equivalent', etc.) offer variations on the HHI in the hopeful quest for a single 'best' index. All of these technical variations suffer from problems of lack of content, burdensome data needs, and debatable weighting. None of them is likely to displace the standard concentration ratio for mainstream analytical purposes.

See Also

- ► Concentration Ratios
- ► Index Numbers
- ► Market Share

Bibliography

Hirschman, A.O. 1964. The paternity of an index. *American Economic Review* 54: 761.

Scherer, F.M. 1970. *Industrial market structure and economic performance*, 2nd ed. Chicago: Rand-McNally, 1980.

Hermann, Friedrich Benedict Wilhelm von (1795–1868)

Mark Blaug

Keywords

Cameralism; German economics in the early 19th century; Life tables; Supply and demand

JEL Classifications

B31

Hermann was born in Dinkelsbuhl, Germany. His career spanned the half-century or more in which German economics came to terms with English classical political economy, first welcoming it and then rejecting it, particularly in its Ricardian variety. After teaching mathematics in a secondary school, Hermann was appointed to the chair in what was still called *Kameralwissenschaften* [Cameralism] – an old title soon to be discarded – at the University of Munich in 1827. He made his reputation with *Staatswirthschaftliche Untersuchungen* [Investigations into Political Economy] (1832), a book which owed much to *The Wealth of Nations* but little to the writings of either Malthus or Ricardo. The book was organized around the simple but appealing idea that all economic variables are the outcome of the forces of demand and supply, so that economic analysis consists essentially of an investigation of the factors lying behind demand and supply. The book revelled in endless definitions and classifications of types of goods, wants, costs, capitals, and so on, but did not clutter the analysis with endless attacks on the deductive method of the English school. Together with Rau (1792–1870), Hermann thereby laid the foundations on which Mangoldt (1824–68) and Thünen (1783–1850) were soon to build a German brand of classical economics. No wonder Marshall much admired 'Hermann's brilliant genius' and frequently quoted Hermann's treatise in his own *Principles of Economics* (1890).

Hermann became a Director of the Bavarian Statistical Bureau in 1839 and organized the first official life table covering an entire German state. As a member of the Frankfurt Parliament in 1848, he advocated the unification of all German states.

Herskovits, Melville Jean (1895–1963)

George Dalton

Keywords

Conventional economics; Economic anthropology; Formalism; Herskovits, M. J.; Knight, F. H.; Marxism; Non-market economies; Polanyi, K.; Substantivism

JEL Classifications

B31

Herskovits was born in Bellefontaine, Ohio, and died in Evanston, Illinois. He studied history at the University of Chicago (BA, 1920) and anthropology at Columbia University (Ph.D., 1923) as a student of Franz Boas. He taught at Columbia and Howard universities before going to Northwestern in 1927, where he spent the rest of his academic career. Herskovits did anthropological fieldwork in West Africa, the Caribbean and Brazil, and was among the first American anthropologists to specialize in African societies as well as blacks in the Caribbean and the United States. He 3started the first Program of African Studies in the United States, at Northwestern.

Herskovits was an early contributor to the field of study now established as economic anthropology. The first edition of his book on this topic was called *The Economic Life of Primitive Peoples* (1940), the revised edition being *Economic Anthropology* (1952).

Herskovits is best remembered by economic anthropologists for his views on a theoretical issue of importance that arose in his controversy with Frank Knight, who reviewed the 1940 edition of Herskovits's book. In the 1940 edition, Herskovits criticized the conventional economics of Marshallian microtheory for its uselessness to anthropologists trying to understand the underlying principles which explain the working of primitive economies – such as African tribal economies not yet changed by European colonial rule – primitive economies lacking capitalism's core attributes of machine technology, modern money, and market organization for the transaction of inputs and outputs. In his book review, Frank Knight criticized Herskovits for misunderstanding the 'abstract' and 'intuitive' nature of economic theory. (I doubt that Knight's portrayal of economics, as stated there, would be shared today by many economists.) Knight's review, together with a rejoinder by Herskovits, are reprinted in *Economic Anthropology* (1952).

The relevance of conventional economic theory to the analysis of pre-industrial, non-capitalist economies remains an unresolved issue to this day. It is an issue much more important today because of the much greater interest now in the study of early and primitive economies, and in the study of the large, diverse set of developing economies in the Third World. This inability to agree on the relevance of conventional economics to the analysis of non-market economies finds expression in economic anthropology's literature of acrimonious theoretical dispute and in the existence side by side of three radically different theoretical systems all employed by archaeologists, anthropologists and historians to analyse non-capitalist economies: formalism (that is, conventional microeconomic theory); Marxism; and substantivism (that is, Karl Polanyi's system of analysis described in his *Trade and Market in the Early Empires,* 1957, and his *Primitive, Archaic, and Modern Economies,* 1971).

See Also

- ▶ Economic Anthropology
- ▶ Knight, Frank Hyneman (1885–1962)
- ▶ Polanyi, Karl (1886–1964)

Selected Works

1940. Anthropology and economics. In *The economic life of primitive peoples.* New York: Knopf.

1941. Economics and anthropology: A rejoinder. *Journal of Political Economy* 49: 269–278.

1952. *Economic anthropology.* New York: Knopf.

Bibliography

Dalton, G. 1961. Economic theory and primitive society. *American Anthropologist* 63: 1–25.

Dalton, G. 1969. Theoretical issues in economic anthropology. *Current Anthropology* 10: 63–102.

Knight, F. 1941. Anthropology and economics. *Journal of Political Economy* 49: 247–268.

Martin, J., and K. Knapp (eds.). 1967. *The teaching of development economics*. Chicago: Aldine Press.

Simpson, G.E. 1973. *Melville J. Herskovits*. New York: Columbia University Press.

Heterodox Economics

Frederic S. Lee

Abstract

Although ‘heterodox economics’ is a widely used term, precisely what it means is debated. I argue that heterodox economics refers to a body of economic theories that holds an alternative position vis-à-vis mainstream economics; to a community of heterodox economists who identify themselves as such and embrace a pluralistic attitude towards heterodox theories without rejecting contestability and incommensurability among heterodox theories; and to the development of a coherent economic theory that draws upon various theoretical contributions by heterodox approaches which stand in contrast to mainstream theory.

Keywords

Association for Heterodox Economics; Austrian economics; Capitalism; Class; Evolutionary economics; Feminist economics; Heterodox economics; Institutional economics; Interdependencies; Marxian economics; Post Keynesian economics; Radical economics; Scarcity; Social economics; Sraffian economics; Pluralism; Political economy; Mainstream economics; Social provisioning

JEL Classifications

B5

‘Heterodox economics’ refers to economic theories and communities of economists that are in various ways an alternative to mainstream economics. It is a multi-level term that refers to a body of economic theories developed by economists who hold an irreverent position vis-à-vis mainstream economics and are typically rejected out of hand by the latter; to a community of heterodox economists who identify themselves as such and embrace a pluralistic attitude towards heterodox theories without rejecting contestability and incommensurability among heterodox theories; and to the development of a coherent economic theory that draws upon various theoretical contributions by heterodox approaches which stand in contrast to mainstream theory. Thus, the article is organized as follows. The first section outlines the emergence of ‘heterodox economics’ in the sense of a body of heterodox theories; the second section deals with heterodox economics as a pluralist community of heterodox economists; the third section situates heterodox economics relative to mainstream economics; and the fourth section delineates heterodox economics in terms of theory and policy.

Heterodox Economics as a Group of Heterodox Theories

Heterodox as an identifier of an economic theory and/or economist that stands in some form of dissent relative to mainstream economics was used within the Institutionalist literature from the 1930s to the 1980s. Then, in 1987, Allan Gruchy used ‘heterodox economics’ to identify Institutional as well as Marxian and Post Keynesian theories as ones that stood in contrast to mainstream theory. By the 1990s, it became obvious that there were a number of theoretical approaches that stood, to some degree, in opposition to mainstream theory. These heterodox approaches included Austrian economics, feminist economics, Institutional evolutionary economics, Marxian-radical economics, Post Keynesian and Sraffian economics, and social economics; however, none of the names of the various heterodox approaches were suitable as a general term that could represent them collectively. While terms such as 'non-traditional', 'non-orthodox', 'non-neoclassical' and 'non-mainstream' were used to collectively represent them, they did not have the right intellectual feel or a positive ring. Moreover, some thought that 'political economy' (or 'heterodox political economy') could be used as the collective term, but its history of being another name for Marxian-radical economics (and its current reference to public choice theory) made this untenable. Therefore, to capture the commonality of the various theoretical approaches in a positive light without prejudicially favouring any one approach, a descriptive term that had a pluralist 'bigtent feel' combined with being unattached to a particular approach was needed. Hence, 'heterodox' became increasingly used throughout the 1990s in contexts where it implicitly and/or explicitly referred to a collective of alternative theories vis-à- vis mainstream theory and to the economists who engaged with those theories.

The final stage in the general acceptance of heterodox economics as the 'official' collective term for the various heterodox theories began c. 1999. First there was the publication of Philip O'Hara's comprehensive *Encyclopedia of Political Economy* (1999), which explicitly brought together the various heterodox approaches. At the same time, in October 1998, Fred Lee established the Association for Heterodox Economics (AHE); and to publicize the conference and other activities of the AHE as well as heterodox activities around the world, he also developed from 1999 an informal 'newsletter' that eventually became (in September 2004) the *Heterodox Economics Newsletter*, now received by over 3,800 economists worldwide (see http://heterodoxnews.com). These twin developments served to establish 'heterodox economics' as the preferred terminology by which these groups of economists referred to themselves.

Heterodox Economics as a Community of Heterodox Economists

'Heterodox economics' also denotes a community of heterodox economists, which implies that the members are not segregated along professional and theoretical lines. The segregation of professional engagement has not existed among heterodox associations, with the exception of two instances in the mid-1970s. For example, from their formation in 1965–70, the three principal heterodox associations in the United States, AFEE, ASE, and URPE (see Table 1 for full names), opened their conferences to Institutionalist, social economics, radical-Marxian, and Post Keynesian papers and sessions; appointed and/or elected heterodox economists to the editorial boards of their journals and to their governing bodies who also were members of other heterodox associations or engaged with Post Keynesian economics; and had members who held memberships in other heterodox associations, engaged with Post Keynesian economics, and subscribed to more than one heterodox economics journal. Moreover, a number of heterodox associations formed since 1988, such as AHE, EAEPE, ICAPE, SDAE and SHE, have adopted an explicitly pluralistic approach towards their name, membership and conference participation: for a list of heterodox associations, dates formed and primary country or region of activity, see Table 1. Finally, the informal and explicit editorial policies of heterodox journals have, from their formation, accepted papers for publication that engage with the full range of heterodox approaches; and this tendency strengthened since the mid-1990s as heterodox economics became more accepted. To illustrate this point, from 1993 to 2003 the eight principal English-language generalist heterodox journals – *Cambridge Journal of Economics*, *Capital and Class*, *Feminist Economics*, *Journal*

Heterodox Economics, Table 1 Heterodox economics associations (currently active)

Name	Date established	Country or region of primary activity
Association for Evolutionary Economics (AFEE)	1965	United States
Association for Heterodox Economics (AHE)	1998	United Kingdom & Ireland
Association for Institutionalist Thought (AFIT)	1979	United States
Association for Social Economics (ASE)	1970	United States
Association pour le Développement des Etudes Keynesiennes	2000	France
Brazilian Keynesian Association (BKA)	2008	Brazil
Conference of Socialist Economists (CSE)	1970	United Kingdom
European Association for Evolutionary Political Economy (EAEPE)	1988	Europe
French Association of Political Economy (FAPE)	2009	France
International Association for Feminist Economics (IAFFE)	1992	World
International Confederation of Associations For Pluralism in Economics (ICAPE)	1993	United States/World
International Initiative for Promoting Political Economy (IIPPE)	2006	Europe
Japan Association for Evolutionary Economics (JAFEE)	1996	Japan
Japan Society of Political Economy (JSPE)	1959	Japan
Korean Social and Economic Studies Association	1987	Korea
L'Association d'Economie Politique	1980	Canada
Progressive Economics Forum (PEF)	1998	Canada
Society for the Advancement of Socio-Economics (SABE)	1989	United States
Society for the Development of Austrian Economics (SDAE)	1996	United States
Society for Heterodox Economics (SHE)	2002	Australia
Union for Radical Political Economics (URPE)	1968	United States
US Society for Ecological Economics (USSEE)	2000	United States

of Economic Issues, *Journal of Post Keynesian Economics*, *Review of Political Economy*, *Review of Radical Political Economics*, and *Review of Social Economy* – cited each other so extensively that no single journal or subset of journals was isolated; hence they form an interdependent body of literature where all heterodox approaches have direct and indirect connections with each other. Thus, in terms of professional engagement since the mid-1990s, the heterodox community is a pluralistic integrative whole.

Theoretical segregation involves the isolation of a particular theoretical approach and its adherents from all other approaches and their adherents; that is to say, theoretical segregation occurs when there is no engagement across different theoretical approaches. However, it does not exist within heterodox economics currently, nor has it existed in the past among the various heterodox approaches. From the 1960s to the 1980s heterodox economists engaged, integrated or synthesized Institutional, Post Keynesian and Marxist-radical approaches, Institutional and Post Keynesian approaches, Post Keynesian and Marxian-radical approaches, Post Keynesian and Austrian, Austrian and Institutional, feminist and Marxist-radical approaches, Institutional and Marxist-radical approaches, Institutional and social economics, ecological and Marxian-radical approaches, and social and Marxian economics. Thus by 1990 many heterodox economists could no longer see distinct boundaries between the various approaches. Moreover, from the 1990s to the present day heterodox economics has continued the past integration efforts of engaging across the various heterodox approaches. Hence, it is clear that the heterodox community is not segregated along theoretical lines, but rather there is cross-approach engagement to such an extent that the boundaries of the various approaches do not simply overlap – they are, in some cases, not there at all. The ensuing theoretical messiness of cross-approach engagement is evidence, to detractors, of the theoretical incoherence of heterodox economics, whereas to supporters of progress it is evidence of a more theoretically coherent heterodox economics – a glass half-empty of coherence as opposed to a glass half-full of coherence.

Heterodox Critique of Mainstream Economics

Mainstream economics is a clearly defined theoretical story about how the economy works; but this story is theoretically incoherent. That is, mainstream theory is comprised of a core set of propositions – such as scarcity, equilibrium, rationality, preferences, and methodological individualism and derivative beliefs, vocabulary, symbols and parables – while there is a range of heterogeneous theoretical developments beyond the core that do not call into question the core itself in totality. As a result, critiques of the theory vary in that they can deal with the internal coherence and/or empirical grounding of the theory; they can be directed at the theory at a particular point in time or at specific components of theory (such as methodology, concepts qua vocabulary, parables qua stories and symbols); and they can be initiated from a particular heterodox approach. What emerges is a varied but concatenation of particular and extensive critiques that generate an emergent encompassing rejection of mainstream theory, although any one particular critique may not go that far.

Although the internal critiques and critiques of models that tell theoretical stories show that the theory is incoherent, they do not by themselves differentiate mainstream from heterodox theory. This, however, can be dealt with in terms of specific critiques of the core propositions. That is, each of the heterodox approaches has produced critiques of particular core propositions of the theory, while each core proposition has been subject to more than one critique; in addition, the multiple heterodox critiques of a single proposition overlap in argumentation. To illustrate this point, consider the critiques of the concept of scarcity. The Post Keynesians argue that produced means of production within a circular production process cannot be characterized as scarce and that production is a social process, while Institutionalists reject the view that natural resources are not 'produced' or socially created to enter into the production process, and the Marxists argue that the concept is a mystification and misspecification of the economic problem – that it is not the relation of the isolated individual to given resources, but the social relationships that underpin the social provisioning process. The three critiques are complementary and integrative and generate the common conclusion that the concept of scarcity must be rejected as well as the mainstream definition of economics as the science of the nonsocial provisioning process analysed through the allocation of scarce resources among competing ends given unlimited asocial wants of asocial individuals. Other critiques of the core propositions exist and arrive at similar conclusions. Together, the three critiques – internal, story qua model and core propositions – form a concatenated structured heterodox critique that rejects and denies the truth and value of mainstream theory.

Heterodox Economics: Theory and Policy

Since the intellectual roots of heterodox economics are located in traditions that emphasize the wealth of nations, accumulation, justice, social relationships in terms of class, gender, and race, full employment, and economic and social reproduction, the discipline of economics, from its perspective, is concerned, not with prediction per se, but with explaining the actual process that provides the flow of goods and services required by society to meet the needs of those who participate in its activities. That is, economics is the science of the *social* provisioning process, and this is the general research agenda of heterodox economists. The explanation involves human agency in a cultural context and social processes in historical time affecting resources, consumption patterns, production and reproduction, and the meaning (or ideology) of market, state, and non-market/state activities engaged in social provisioning. Thus heterodox economics has two interdependent parts: theory and policy. *Heterodox economic theory* is an empirically grounded theoretical explanation of the historical process of social provisioning within the context of a capitalist economy. Therefore it is concerned with explaining those factors that are part of the process of social provisioning, including the structure and use of resources, the structure and change of social wants, structure of production and the reproduction of the business enterprise, family, state, and other relevant institutions and organizations, and distribution. In addition, heterodox economists extend their theory to examining issues associated with the process of social provisioning, such as racism, gender and ideologies and myths. Because their economics involves issues of ethical values and social philosophy and the historical aspects of human existence, heterodox economists make ethically based *economic policy* recommendations to improve human dignity, that is, recommending ameliorative and/or radical, social and economic policies to improve the social provisioning and hence well-being for all members of society and especially the disadvantaged members. To do this properly, their economic policy recommendations must be connected to heterodox theory which provides an accurate historical and theoretical picture of how the economy actually works – a picture that includes class and hierarchical domination, inequalities, and social-economic discontent.

Given the definition of economics as the science of the social provisioning process and the structure of the explanation of the process combined with the pluralistic and integrative proclivities of heterodox economists, there has emerged a number of elements that have come to constitute the provisional theoretical and methodological core of heterodox theory. Some elements are clearly associated with particular heterodox approaches, as noted by O'Hara (2002, p. 611):

> The main thing that social economists bring to the study [of heterodox economics] is an emphasis on ethics, morals and justice situated in an institutional setting. Institutionalists bring a pragmatic approach with a series of concepts of change and normative theory of progress, along with a commitment to policy. Marxists bring a set of theories of class and the economic surplus. Feminists bring a holistic account of the ongoing relationships between gender, class and ethnicity in a context of difference . . . And post-Keynesians contribute through an analysis of institutions set in real time, with the emphasis on effective demand, uncertainty and a monetary theory of production linked closely with policy recommendations.

However, other provisional elements, such as critical realism, non-equilibrium or historical modelling, the gendering and emotionalizing agency, the socially embedded economy, and circular and cumulative change, emerged from a synthesis of arguments that are associated only in part with particular heterodox approaches.

The core methodological elements establish the basis for constructing heterodox theory. In particular, the methodology emphasizes realism, structure, feminist and uncertain agency qua individual, history, and empirical grounding in the construction of heterodox theory, which is a historical narrative of how capitalism works. The theory qua historical narrative does not simply recount or superficially describe actual economic events, such as the exploitation of workers; it does more in that it analytically explains the internal workings of the historical economic process that, say, generates the exploitation of workers. Moreover, because of its historical nature, the narrative is not necessarily organized around the concepts of equilibrium/long period positions and tendencies towards them. Because the narrative provides an accurate picture of how capitalism actually works and changes in a circular and cumulative fashion, economists use their theory to suggest alternative paths that future economic events might take and propose relevant economic policies to deal with them. In constructing the narrative, they have at the same time created a particular social-economic-political picture of capitalism.

The core theoretical elements generate a three-component structure–organization–agency economic theory. The first component of the theory consists of three overlapping interdependencies that delineate the structure of a real capitalist economy. The first interdependency is that the production of goods and services requires goods and services to be used as inputs. Hence, with regard to production, the overall economy (which includes both market and non-market production) is represented as an input–output matrix of material goods combined with different types of labour skills to produce an array of goods and services as outputs. Many of the outputs replace the goods and services used up in production and the rest constitute a physical surplus to be used for social provisioning, that is for consumption, private investment, government usage and exports. A second interdependency is the relation between the wages of workers, profits of enterprises, and taxes of government and expenditures on consumption, investment, and government goods as well as non-market social provisioning activities. The last interdependency consists of the overlay of the flow of funds or money accompanying the production and exchange of the goods and services. Together, these three interdependencies produce a monetary input–output structure of the economy where transactions in each market are a monetary transaction; where a change in price of a good or the method by which a good is produced in any one market will have an indirect or direct impact on the entire economy; and where the amount of private investment, government expenditure on real goods and services, and the excess of exports over imports determines the amount of market and non-market economic activity, the level of market employment and non-market labouring activities, and consumer expenditures on market and non-market goods and services. These elements of course have parallels in non-heterodox economics, but the ideas are developed differently.

The second component of heterodox theory consists of three broad categories of economic organization that are embedded in the monetary input–output structure of the economy. The first category is micro market-oriented, hence particular to a set of markets and products. It consists of the business enterprise, private and public market organizations that regulate competition in product and service markets and the organizations and institutions that regulate the wages of workers. The second is macro market-oriented and hence is spread across markets and products, or is not particular to any market or product. It includes the state and various subsidiary organizations as well as particular financial organizations, that is, those organizations that make decisions about government expenditures and taxation, and the interest rate. Finally, the third category consists of non-market organizations that promote social reproduction and include the family and state and private organizations that contribute to and support the family. The significance of organizations is that they are the social embeddedness of agency qua the individual, the third component of heterodox theory. That is, agency, which are decisions made by individuals concerning the social provisioning process and social well-being, takes place through these organizations. And because the organizations are embedded in both instrumental and ceremonial institutions, such as gender, class, ethnicity, justice, marriage, ideology, and hierarchy qua authority, agency qua the individual acting through organizations affect both positively and negatively but never optimally the social provisioning process.

Conclusion

If mainstream economics suddenly disappeared, heterodox economics would be largely unaffected. It would still include the various heterodox traditions; there would still be an integrated professional and theoretical community of heterodox economists; and its heterodox research agenda would still be directed at explaining the social provisioning process in capitalist economies and argue for economic policies that would enhance social well-being. In this regard, heterodox economics is not out to reform mainstream economics. Rather, it is an alternative to mainstream economics: an alternative in terms of explaining the social provisioning process and suggesting economic policies to promote social well-being. Since the mid-1990s the community of heterodox economics has grown, diversified and integrated. The previously isolated are now part of a community, heterodox associations exist in countries where previously no heterodox associations had existed, and developments in heterodox theory and policy are occurring at breakneck speed. In short, heterodox economics is now an established feature on the disciplinary landscape and the progressive future of economics.

See Also

▶ Pluralism in Economics

Bibliography

Bortis, H. 1997. *Institutions, behaviour and economic theory*. Cambridge: Cambridge University Press.

Dow, S.C. 2000. Prospects for the progress of heterodox economics. *Journal of the History of Economic Thought* 22: 157–170.

Granovetter, M. 1985. Economic action and social structure: The problem of embeddedness. *American Journal of Sociology* 91: 481–510.

Gruchy, A.G. 1987. *The reconstruction of economics: An analysis of the fundamentals of institutional economics*. New York: Greenwood Press.

Keen, S. 2001. *Debunking economics: The naked emperor of the social sciences*. New York City: St Martin's Press.

Lavoie, M. 2006. Do heterodox theories have anything in common? A Post-Keynesian point of view. *International Journal of Ecodynamics* 3: 87–112.

Lawson, T. 1997. *Economics and reality*. London: Routledge.

Lee, F.S. 2001. Conference of socialist economists and the emergence of heterodox economics in post-war Britain. *Capital & Class* 75: 15–39.

Lee, F.S. 2002. The association for heterodox economics: Past, present, and future. *Journal of Australian Political Economy* 50: 29–43.

Lee, F.S. 2004. To be a heterodox economist: the contested landscape of American economics, 1960s and 1970s. *Journal of Economic Issues* 38: 747–763.

Lee, F.S., and S. Keen. 2004. The incoherent emperor: A heterodox critique of neoclassical microeconomic theory. *Review of Social Economy* 62: 169–199.

Lee, F.S., S. Cohen, G. Schneider, and P. Quick. 2005. *Informational directory for heterodox economists: Journals, book series, websites, and graduate and undergraduate programs*. 2nd ed. Kansas City: Department of Economics, University of Missouri-Kansas City . Online. Available at http://l.web.umkc.edu/leefs/htnf/HeterodoxDirectory.pdf. Accessed 12 Jan 2007.

Matthaei, J. 1984. Rethinking scarcity: Neoclassicism, neoMalthusianism, and neoMarxism. *Review of Radical Political Economics* 16: 81–94.

O'Hara, P.A., eds. 1999. *Encyclopedia of political economy*. London: Routledge.

O'Hara, P.A. 2000. *Marx, Veblen, and contemporary institutional political economy: Principles and unstable dynamics of capitalism*. Cheltenham: Edward Elgar.

O'Hara, P.A. 2002. The role of institutions and the current crises of capitalism: A reply to Howard Sherman and John Henry. *Review of Social Economy* 60: 609–618.

Power, M. 2004. Social provisioning as a starting point for feminist economics. *Feminist Economics* 10: 3–19.

White, G. 2004. Capital, distribution and macroeconomics: 'Core' beliefs and theoretical foundations. *Cambridge Journal of Economics* 28: 527–547.

Wrenn, M.V. 2004. What is heterodox economics? Ph.D. thesis, Colorado State University.

Heteroskedasticity

J. Kmenta

One of the basic assumptions of the classical regression model

$$Y_i = \beta_1 + \beta_2 X_{i2} + \cdots + \beta_k X_{iK} + \varepsilon_i \quad (i = 1, 2, \ldots, n)$$

is that the variance of the regression disturbance ε_i is constant for all observations, that is, that $\mathrm{Var}(\varepsilon_i) = \sigma^2$ for all i. This feature of ε_i is known as *homoskedasticity* and its absence is called *heteroskedasticity*. The homoskedasticity assumption is quite reasonable for observations on aggregates over time, since the values are of a similar order of magnitude for all observations. It is, however, implausible with respect to observations on microeconomic units such as households or firms included in a survey, since there are likely to be substantial differences in magnitude of the observed values. For example, in the case of survey data on household income and consumption, we would expect less variation in consumption of low-income households, whose average level of consumption is low, than in consumption of high-income households, whose average level of consumption is high. Empirical evidence suggests that this expectation is in accord with actual behaviour. Heteroskedasticity also arises when the data are in the form of group averages and the groups are of unequal size.

Heteroskedasticity has two important consequences for estimation: (1) The least squares estimators of the regression coefficients are no longer efficient or asymptotically efficient. (2) The estimated variances of the least squares estimators are, in general, biased, and the conventionally calculated confidence intervals and tests of significance are invalid. The second of these consequences is more serious than the first since inefficiency of estimation can be compensated for by a large number of observations.

The deficiencies of the least squares estimation can be remedied by adopting a *weighted* (or *generalized*) least squares procedure. This method involves weighting each observation by the reciprocal of the respective standard deviation of the disturbance, and then applying the least squares method to the transformed equation

$$(Y_i/\sigma_i) = \beta_1(1/\sigma_i) + \beta_2(X_{i2}/\sigma_i) + \cdots + \beta_K(X_{iK}/\sigma_i) + u_i,$$

where

$$\sigma_i = \sqrt{\mathrm{Var}(\varepsilon_i)} \qquad \text{and} \qquad u_i = \varepsilon_i/\sigma_i.$$

The difficulty with the weighted least squares method is that its implementation requires knowledge of σ_i, which is rarely available. This difficulty is usually overcome by making certain assumptions about σ_i or, when possible, by estimating σ_i. The assumptions typically involve associating σ_i with some variable Z_i, normally represented by one of the explanatory variables of the regression equation. For instance, in a microconsumption function the variance of the disturbance is frequently positively associated with income. In general, two forms of association between σ and Z have been proposed in the literature and applied in practice: a *multiplicative* and an *additive* form. Multiplicative heteroskedasticity – which is more common – can be described as

$$\sigma_i^2 = \sigma^2 Z_1^{\delta},$$

where σ and δ are parameters to be estimated. A frequent representation of additive heteroskedasticity is

$$\sigma_i^2 = a + bZ_i + cZ_i^2,$$

where a, b, and c are parameters to be estimated. Estimation of the parameters involved in the specification of σ_i can be carried out simultaneously with the estimation of the regression coefficients by using the method of maximum likelihood. No assumptions about the form of heteroskedasticity are necessary where σ_i can be estimated from replicated data which, unfortunately, are rather rare in applied economic research.

The presence or absence of heteroskedasticity may be subjected to a test. Several suitable tests, some developed only recently, are available and are described in recent econometric texts.

The problem of heteroskedasticity and its consequences was brought to the attention of applied economists by two seminal research monographs, Stone (1954) and Prais and Houthakker (1955). The subject has been further developed by a number of econometricians and is now standard fare in all introductory courses of econometrics; see, for example Kmenta (1986).

See Also

- ▶ Least Squares
- ▶ Regression and Correlation Analysis
- ▶ Residuals

Bibliography

Kmenta, J. 1986. *Elements of econometrics*, 2nd ed. New York: Macmillan.

Prais, S.J., and H.S. Houthakker. 1955. *The analysis of family budgets*. Cambridge: Cambridge University Press.

Stone, J.R.N. 1954. *The measurements of consumers' expenditure and behaviour in the United Kingdom, 1920–1938*, vol. I. Cambridge: Cambridge University Press.

Heteroskedasticity and Autocorrelation Corrections

Kenneth D. West

Abstract

Many time series studies, including in particular those estimated by generalized method of moments, involve disturbances that are serially correlated and, possibly, conditionally heteroskedastic. The serial correlation and heteroskedasticity often are of unknown form. Corrections for serial correlation and heteroskedasticity are required for inference and efficient estimation. This article surveys procedures to implement such corrections.

Keywords

Autocovariances; Bartlett kernel; Generalized method of moments; Heteroskedasticity; Heteroskedasticity and autocorrelation consistent (HAC) covariance matrix estimation; Kernel weights; Long-run variance; Moving average processes; Newey–West estimator; Quadratic spectral kernel; Serial correlation; Spectral density estimation; Truncated estimators; Vector autoregressions

JEL Classifications

C1

Heteroskedasticity and autocorrelation consistent (HAC) covariance matrix estimation refers to calculation of covariance matrices that account for conditional heteroskedasticity of regression disturbances and serial correlation of cross products of instruments and regression disturbances. The heteroskedasticity and serial correlation may be of unknown form. HAC estimation is integral to empirical research using generalized method of moments (GMM) estimation (Hansen 1982). In this article I summarize results relating to HAC estimation, with emphasis on practical rather than theoretical aspects.

The central issue is consistent and efficient estimation of what is called a 'long-run variance', subject to the constraint that the estimator is positive semidefinite in finite samples. Positive semidefiniteness is desirable since the estimator will be used to compute standard errors and test statistics. To fix notation, let h_t be a $q \times 1$ stationary mean zero random vector. Let Γ_j denote the $q \times q$ autocovariance of h_t at lag q, $\Gamma_j \equiv E h_t h_{t-j}{}'$; of course, $\Gamma_j = \Gamma_{-j}{}'$. The long run variance of h_t is the $q \times q$ matrix

$$S = \sum_{j=-\infty}^{\infty} \Gamma_j = \Gamma_0 + \sum_{j=1}^{\infty} \left(\Gamma_j + \Gamma_0{}'\right). \qquad (1)$$

Apart from a factor of 2π, the symmetric matrix S, which I assume to be positive definite, is the spectral density of h_t at frequency zero. As discussed below, techniques for spectral density estimation are central to HAC estimation. (For an arbitrary stationary process, the sum in the right-hand side of (1) may not converge, and may not be positive definite even if it does converge. But here and throughout I assume unstated regularity conditions. As well, I use formulas that allow for relatively simple notation, for example assuming covariance stationarity even when that assumption can be relaxed. The cited papers may be referenced for generalizations and for technical conditions.)

To illustrate how estimation of S figures into covariance matrix estimation, consider the following simple example. As in Hansen and Hodrick (1980), let us suppose that we wish to test the 'rationality' of a scalar variable x_t as an n period ahead predictor of a variable y_{t+n+1}, for $n \geq 0$: the null hypothesis is $E_t y_{t+n+1} = x_t$, where E_t denotes expectations conditional on the information set used by market participants. The variable x_t might be the expectation of y_{t+n+1} reported by a survey, or it might be a market determined forward rate. Let u_t denote the expectational error: $u_t = y_{t+n+1} - E_t y_{t+n+1} = y_{t+n+1} - x_t$. (The expectational error u_t, which is not realized until period $t + n + 1$, is dated t to simplify notation.)

One can test one implication of the hypothesis that x_t is the expectation of y_{t+n+1} by regressing y_{t+n+1} on a constant and x_t, and checking whether the coefficient on the constant term is zero and that on x_t is 1:

$$y_{t+n+1} = \beta_0 + \beta_1 x_t + u_t \equiv X_t'\beta + u_t;\ H_0 : \beta = (0, 1)'. \qquad (2)$$

Under the null, $EX_t u_t = 0$, so least squares is a consistent estimator. As well, $X_t u_t$ follows a moving average process of order n. Thus the asymptotic variance of the least squares estimator of β is $(EX_t X_t')^{-1} S (EX_t X_t')^{-1}$, where $S = \Gamma_0 + \sum_{j=1}^{n} \left(\Gamma_j + \Gamma_j{}'\right)$, $\Gamma_j \equiv EX_t u_t (X_{t-j} u_{t-j})'$. This example maps into the notation used in (1) with $h_t = X_t u_t$, $q = 2$ and a known upper bound to the number of non-zero autocovariances of h_t. Clearly one needs to estimate EX_tX_t' and S to conduct inference. A sample average of X_tX_t' can be used to estimate EX_tX_t'. If $n = 0$, so that h_t is serially uncorrelated, $S = EX_tu_t\ (X_tu_t)'$ and estimation of S is equally straightforward; White's (1980) heteroskedasticity consistent estimator can be used. The subject at hand considers ways to estimate S when h_t is serially correlated. I note in passing that one cannot sidestep estimation of S by applying generalized least squares. In this example and more generally, generalized least squares is inconsistent. See Hansen and West (2002).

To discuss estimation of S, let us describe a more general set-up. In GMM estimation, h_t is a $q \times 1$ orthogonality condition used to identify a k-dimensional parameter vector β. The orthogonality condition takes the form

$$h_t = Z_tu_t \tag{3}$$

for a $q \times \ell$ matrix of instruments Z_t and an $\ell \times 1$ vector of unobservable regression disturbances u_t. The vector of regression disturbances depends on observable data through β, $u_t = u_t(\beta)$. In the example just given, $q = 2$, $\ell = 1$, $Z_t = X_t$, $u_t\ (\beta) = y_{t+n+1}\ {-X_t}'\ \beta$. The example just given is overly simple in that the list of instruments typically will not be identical to right-hand side variables, and the model may be nonlinear. For a suitable $k \times q$ matrix D, the asymptotic variance of the GMM estimator of β takes the form DSD' (for example, $D = (EX_t\,Xt\,')^{-1}$ in the example just given). In an overidentified model (that is, in models in which the dimension of the orthogonality condition q is greater than the number of parameters k) the form D takes depends on a certain weighting matrix. Let $h_{t\beta}$ be the $q \times k$ matrix $\delta h_t = \delta\beta$. When the weighting matrix is chosen optimally, $D = (Eh_{t\beta}\ 'S^{-1}Eh_{t\beta})^{-1}\ Eh_{t\beta}\ 'S^{-1}$ and the asymptotic variance DSD' simplifies to $(Eh_{t\beta}'S^{-1}Eh_{t\beta})^{-1}$. The optimal weighting matrix is one that converges in probability to S, and thus the results about to be presented are relevant to efficient estimation as well as to hypothesis testing. In any event, the matrix $Eh_{t\beta}$ typically is straightforward to estimate; the question is how to estimate S. This will be the focus of the remainder of the discussion.

We have sample of size T and sample counterparts to u_t and h_t, call them $\hat{u}_t = u_t\left(\hat{\beta}\right)$ and $\hat{h} = h_t\left(\hat{\beta}\right)$. Here, $\hat{\beta}$ is a consistent estimate of β. In the least squares example given above, $\hat{u}_t$ is the least squares residual, $\hat{u}_t = y_{t+n+1}\ - X_t'\hat{\beta}$, and $\hat{h}_t = X_t\hat{u}_t = X_t\left(y_{t+n+1}\ - X_t'\hat{\beta}\right)$. One path to consistent estimation of S involves consistent estimation of the autocovariances of h_t. The natural estimator is a sample average,

$$\hat{\Gamma}_j = T^{-1}\sum\nolimits_{t=j+1}^{T} \hat{h}_t\hat{h}_{t-j}{}' \text{ for } j \geq 0. \tag{4}$$

For given j, (4) is a consistent ($T \rightarrow \infty$) estimator of Γ_j.

I now discuss in turn several possible estimators, or classes of estimators, of S: (1) the truncated estimator; (2) estimators applicable only when h_t follows a moving average (MA) process of known order; (3) an autoregressive spectral estimator; (4) estimators that smooth autocovariances; (5) some recent work, on estimators that might be described as extensions or modifications of ones the estimators described in (4).

The Truncated Estimator

Suppose first that it is known a priori that the autocovariances of h_t are zero after lag n, as is the case in the empirical example above. A natural estimator of S is one that replaces population objects in (1) with sample analogues. This is the *truncated* estimator:

$$\hat{S}_{TR} = \hat{\Gamma}_0 + \sum_{j=1}^{n}\left(\hat{\Gamma}_j + \hat{\Gamma}_j{}'\right). \tag{5}$$

In the more general case in which $\Gamma_j \neq 0$ for all j, the truncated estimator is consistent if the truncation point $n \rightarrow \infty$ at a suitable rate. Depending on exact technical conditions, the rate may be $n/T^{1/2} \rightarrow 0$ or $n/T^{1/4} \rightarrow 0$ (Newey and West 1987). The truncated estimator need not, however, yield a positive semidefinite estimate. With certain plausible data generating processes, simulations indicate that it will not be p.s.d. in a large fraction of samples (West 1997). Hence this estimator is not used much in practice.

Estimators Applicable only When h_t Follows an MA Process of Known Order n

Such a process for h_t holds in studies of rationality (as illustrated above) and in the first order conditions from many rational expectations models (for example, Hansen and Singleton 1982).

Write the Wold representation of h_t as $h_t = e_t + \Theta_1 e_{t-1} + \ldots + \Theta_n e_{t-n}$. Here, e_t is the $q \times 1$ innovation in h_t. Let Ω denote the $q \times q$ variance covariance matrix of e_t. Then it is well known (for example, Hamilton 1994, p. 276) that

$$S = (I + \Theta_1 + \ldots + \Theta_n)\Omega(I + \Theta_1 + \ldots + \Theta_n)' \tag{6}$$

Suppose that one fits an MA(n) process to $\hat{h}_t$, and plugs the resulting estimates of the Θ_i and Ω into the formula for S. Clearly the resulting estimator is $T^{1/2}$ consistent and positive semidefinite. Nevertheless, to my knowledge this estimator has not been used, presumably because of numerical difficulties in estimating multivariate moving average processes.

Two related estimators have been proposed that impose a smaller computational burden. Hodrick (1992) and West (1997) suggest an estimator that requires fitting an MA(n) to the vector of regression residuals $\hat{u}_t$(or, in Hodrick's 1992, application, using MA coefficients that are known a priori). The computational burden of such MA estimation will typically be considerably less than that of MA estimation of the h_t process, because the dimension of u_t is usually much smaller than that of h_t. For example, $\hat{u}_t$ will be a scalar in a single equation application, regardless of the number of orthogonality conditions captured in h_t. Write the estimated MA process for $\hat{u}_t$ as $\hat{u}_t = \hat{\in}_t + \hat{\psi}_1 \hat{\in}_{t-1} + \cdots + \hat{\psi}_n \hat{\in}_{t-n}$, where the $\hat{\psi}_j$ are $\ell \times \ell$. (Note that $\in_t$, the $\ell \times 1$ innovation in u_t, is not the same as e_t, the $q \times 1$ innovation in h_t.) Then a $T^{1/2}$ consistent and positive semidefinite estimator of S is

$$\begin{aligned}\hat{S}_{MA-\ell} &= T^{-1}\sum_{t=1}^{T-n} \hat{d}_{t+n}\hat{d}_{t+n}{}', \hat{d}_{t+n}\\ &= \left(Z_t + Z_{t+1}\hat{\psi}_1 + \cdots + Z_{t+n}\hat{\psi}_n\right)\hat{\in}_t,\end{aligned} \tag{7}$$

where, again, Z_t is the $q \times \ell$ matrix of instruments (see Eq. (3)).

Eichenbaum et al. (1988) and Cumby et al. (1983) propose a different strategy that avoids the need to estimate a moving average process for either u_t or h_t. They suggest estimating the parameters of $\hat{h}_t$'s autoregressive representation, and inverting the autoregressive weights to obtain moving average weights. Call the results $\hat{\Theta}_1, \ldots, \hat{\Theta}_n$, with $\hat{\Omega}$ the estimate of the innovation variance–covariance matrix. The resulting estimator $\hat{S} = \left(I + \hat{\Theta}_1 + \cdots + \hat{\Theta}_n\right)\hat{\Omega}\left(I + \hat{\Theta}_1 + \cdots \ +\hat{\Theta}_n\right)'$ is positive semidefinite by construction. The rate at which it converges to S depends on the rate at which the order of the autoregression is increased.

Autoregressive Estimators

Den Haan and Levin (1997) propose and evaluate an autoregressive spectral estimator. Suppose that h_t follows a (possibly) infinite-order vector autoregression (VAR)

$$h_t = \sum_{j=1}^{\infty} \Phi_j h_{t-j} + e_t, \quad E e_t e_t' = \Omega. \tag{8}$$

Then (Hamilton 1994, p. 237)

$$S = \left(I - \sum_{j=1}^{\infty} \Phi_j\right)^{-1} \Omega \left(I - \sum_{j=1}^{\infty} \Phi_j\right)^{-1'}.$$

The idea is to approximate this quantity via estimates from a finite-order VAR in $\hat{h}_t$. Write the estimate of a VAR in $\hat{h}_t$ of order p as

$$\begin{aligned}\hat{h}_t &= \hat{\Phi}_1 \hat{h}_{t-1} + \cdots + \hat{\Phi}_p \hat{h}_{t-p} + \hat{e}_t,\\ \hat{\Omega} &= T^{-1} \sum_{t=p+1}^{T} \hat{e}_t \hat{e}_t'.\end{aligned} \tag{10}$$

Then the estimator of $\hat{S}$ is

$$\hat{S}_{AR} = \left(I - \sum_{j=1}^{p} \hat{\Phi}_j\right)^{-1} \hat{\Omega} \left(I - \sum_{j=1}^{p} \hat{\Phi}_j\right)^{-1'}. \tag{11}$$

Den Haan and Levin (1997, Section 3.5) conclude that if p is chosen by BIC, and some other technical conditions hold, then this estimator converges at a rate very near $T^{1/2}$ (the exact rate depends on certain characteristics of the data). A possible problem in practice with this estimator (as well as with the estimator described in the final paragraph of Section 2, which also requires estimates of a VAR in $\hat{h}_t$) is that it may require estimation of many parameters and inversion of a large matrix. Den Haan and Levin therefore suggest judiciously parametrizing the autoregressive process, for example by using the BIC criterion equation-by-equation for each of the q elements of $\hat{h}_t$.

Estimators that Smooth Autocovariances

In practice, the most widely used class of estimators is one that relies on smoothing of autocovariances. Andrews (1991), building on the literature on estimation of spectral densities, established a general framework for analysis. Andrews considers estimators that can be written

$$\hat{S} = \hat{\Gamma}_0 + \sum_{j=1}^{T-1} k_j \left(\hat{\Gamma}_j + \hat{\Gamma}_j'\right) \tag{12}$$

for a series of kernel weights $\{k_j\}$ that obey certain properties. For example, to obtain a consistent estimator, we need k_j near zero (or perhaps identically zero) for values of j near $T - 1$, since autocovariances at large lags are estimated imprecisely, while $k_j \rightarrow 1$ for each j is desirable for consistency. We would also like the choice of k_j to ensure positive definiteness.

The two most commonly used formulas for the kernel weights are: Bartlett : for some $m \geq 0$: $k_j = 1 - [j/(m + 1)]$ for $j \leq m$, $kj = 0$ for $j > m$. (13a)

$$\text{Quadraticspectral (QS) : for some } m > 0, \text{and with} x_j = j/m : \\ k_j = \left[25/12\pi^2 x_j^2\right] \times \left\{\left[\sin\left(6\pi x_j/5\right)/\left(6\pi x_j/5\right)\right] - \cos\left(6\pi x_j/5\right)\right\}. \tag{13b}$$

If we let $z_j = 6\pi x_j/5$, the QS formula for k_j can be written in more compact form as $\left(3/z_j^2\right)\left\{\left[\sin\left(z_j\right)/z_j\right] - \cos\left(z_j\right)\right\}$. Call the resulting estimators $\hat{S}_{BT}$ and $\hat{S}_{QS}$. For example,

$$\hat{S}_{BT} = \hat{\Gamma}_0 + \sum_{j=1}^{m} [1 - j/(m+1)]\left(\hat{\Gamma}_j + \hat{\Gamma}_j'\right) \tag{14}$$

The vast literature on spectral density estimation suggests many other possible kernel weights. For conciseness, I consider only the Bartlett and QS kernels.

To operationalize these estimators, one needs to choose the lag truncation parameter or bandwidth m. I note that for both kernels, consistency requires $m \rightarrow \infty$ as $T \rightarrow \infty$, even if h_t follows an MA process of known finite order, as in the example given above. Thus one should not set m to be the number of non-zero autocovariances. Subject to possible problems with positive definiteness, setting $m = n$ is fine for the truncated estimator (5) but not for estimators that use nontrivial weights $\{k_j\}$.

Andrews shows that maximizing the rate at which $\hat{S}$ converges to S requires that m increase as a suitable function of sample size, with the ‘suitable function’ varying with kernel. For the Bartlett and QS, the maximal rates of convergence are realized when

$$\text{Bartlett}: m = \gamma T^{1/3}\left(\text{or } m = \left(\text{integer part of } \gamma T^{1/3}\right)\right) \text{ for some } \gamma \neq 0, \text{QS}: m = \gamma T^{1/5} \text{ for some } \gamma \neq 0, \quad (15)$$

in which case $\hat{S}_{BT}$ converges to S at rate $T^{1/3}$ and the mean squared error in estimation of S goes to zero at rate $T^{2/3}$; the comparable figures for QS are $T^{2/5}$ and $T^{4/5}$. Since both estimators are nonparametric, they converge at rates slower than $T^{1/2}$; since faster convergence is better, the QS rate is preferable to that of the Bartlett. Indeed, Andrews (1991), drawing on Priestley (1981), shows that for a certain class of kernel weights $\{k_j\}$, the mean squared error of QS rate is optimal in the following sense: a $T^{4/5}$ rate on the asymptotic mean squared error is the fastest that can be achieved if one wants to ensure a positive definite $\hat{S}$, and within the class of kernels that achieve the $T^{4/5}$ rate, the QS has the smallest possible asymptotic mean squared error.

As a practical matter, the formulas in (15) have merely pushed the question of choice of m to one of choice of γ; putting arbitrary γ in (15) yields convergence that is as fast as possible, but different choices of γ lead to different asymptotic mean squared errors. The choice of γ that is optimal from the point of view of asymptotic mean squared error is a function of the data (Hannan 1970, p. 286). Let $S^{(0)} = \sum_{j=-\infty}^{\infty} \Omega_j (= S); S^{(1)} = \sum_{j=-\infty}^{\infty} |j|\Omega_j; S^{(2)} = \sum_{j=-\infty}^{\infty} j^2\Omega_j$. For scalar ($q = 1$) S optimal choices are:

$$\text{Bartlett}: \gamma = 1.1447\left[S^{(1)}/S^{(0)}\right]^{2/3}; \text{QS}: \gamma = 1.3221\left[S^{(2)}/S^{(0)}\right]^{2/5}. \quad (16)$$

(See Andrews 1991, for the derivation of these formulas.)

Andrews (1991), Andrews and Monahan (1992) and Newey and West (1994) proposed feasible data dependent to procedures to estimate γ, for vector as well as scalar h_t. Rather than exposit the general case, I will describe two 'cookbook' procedures that have been offered as reasonable starting points in empirical work. One procedure relies on Andrews (1991) and Andrews and Monahan (1992), and assumes the QS kernel and estimation of γ via parametric models. The second relies on Newey and West (1994), and assumes a Bartlett kernel and nonparametric estimation of γ. I emphasize that both papers present more general results than are presented here; both allow the researcher to (for example) use any one of a wide range of kernels.

Let there be a $q \times 1$ vector of weights $w = (w_1, w_2,..., w_q)'$ whose elements tells us how to weight the various elements of S with respect to mean squared error. The weights might be sample dependent, and den Haan and Levin (1997) argue that there are benefits to certain sample-dependent weights, but a simple choice proposed by both papers is: $w_i = 0$ if the corresponding element of h_t is a cross product of a constant term and a regression disturbance, otherwise $w_i = 1$. Andrews's loss function is the normalized expectation of $\sum_{i=1}^{q} w_i \left(S_{ii} - \hat{S}_{ii}\right)^2$, while Newey and West's loss function is the normalized expectation of $\left[w'\left(\hat{S} - S\right)w\right]^2$; the normalization is $T^{4/5}$ for QS and $T^{2/3}$ for Bartlett.

Both procedures begin with using a vector autoregression to prewhiten, and end with re-colouring. The basic justification for prewhitening and re-colouring is that simulation evidence indicates that this improves finite sample performance.

1. Prewhitening: Estimate a vector autoregression in $\hat{h}_t$, most likely of order 1. Call the residuals $\hat{h}_t^{\dagger}$

$$\hat{h}_t = \hat{A}\hat{h}_{t-1} + \hat{h}_t^{\dagger}, \hat{A} = \sum_{t=2}^{T} \hat{h}_t\hat{h}_{t-1}' \left(\sum_{t=2}^{T} \hat{h}_{t-1}\hat{h}_{t-1}'\right)^{-1}. \quad (17)$$

2. Let $\hat{\Gamma}_j^{\dagger}$ denote the jth autocovariance of the VAR residual $\hat{\Gamma}_j^{\dagger}, \hat{\Gamma}_j^{\dagger} = (T-1)^{-1}\sum_{t=2+j}^{T} \hat{h}_t^{\dagger} \hat{h}^{\dagger}{}_{t-j'}$. Using $\left\{\hat{\Gamma}_j^{\dagger}\right\}$ (rather than $\left\{\hat{\Gamma}_j\right\}$ [the autocovariances of $\hat{h}_t$]), and choosing m optimally as described in steps 2a or 2b below, construct an estimate of the long run variance of the residual of the VAR just estimated. Call the result $\hat{S}^{\dagger}$.

2a. Andrews and Monahan (1992): Fit a univariate AR(1) to each of the q elements of $\hat{h}^{\dagger}$. Call the resulting estimate of the AR coefficient and variance of the residual $\hat{\rho}_i$ and $\hat{\sigma}_i^2$. Compute

$$\hat{s}_2 = \sum_{i=1}^{q} w_i \left(4\hat{\rho}_i^2\hat{\sigma}_i^4\right)/(1-\hat{\rho}_i)^8, \hat{s}_0 = \sum_{i=1}^{q} w_i\hat{\sigma}_i^4/(1-\hat{\rho}_i)^4, \hat{\gamma}_{QS} = 1.3221[\hat{s}_2/\hat{s}_0]^{1/5}, \hat{m}_{QS} = \hat{\gamma}_{QS}T^{1/5}. \tag{18}$$

Then plug $\hat{m}_{QS}$ into formula (13b). Call the result $\hat{k}_j$. Compute $\hat{S}^{\dagger} = \hat{\Gamma}_0^{\dagger} + \sum_{j=1}^{T-1} \hat{k}_j \left(\hat{\Gamma}_j^{\dagger} + \hat{\Gamma}_{j'}^{\dagger}\right)$.

2b. Newey and West (1994): Set n = integer part of $12(T/100)^{2/9}$. Compute

$$\hat{s}^{(1)} = w'\hat{\Gamma}_0 w + 2\sum_{i=1}^{n} i w'\hat{\Gamma}_i^{\dagger} w, \hat{s}^{(0)} = 2\sum_{i=1}^{n} w'\hat{\Gamma}_i^{\dagger} w, \hat{\gamma}_{BT} = 1.1447\left[\hat{s}^{(1)}/\hat{s}^{(0)}\right]^{2/3}, \hat{m}_{BT} = \text{integer part of } \hat{\gamma}_{BT}T^{1/3}. \tag{19}$$

Then compute $\hat{S}^{\dagger}$ according to (14), using $\hat{m}_{BT}$.

3. Re-colouring: compute $\hat{S} = \left(I - \hat{A}\right)^{-1}\hat{S}^{\dagger}\left(I - \hat{A}\right)^{-1'}$.

These two recipes for estimates of S can serve as a starting point for experimentation for alternative choices of m and alternative kernels.

What is the simulation evidence on behaviour of these and other proposed estimators? In answering this question, I focus on sizing of test statistics and accuracy of confidence interval coverage: accuracy in estimation of S is desirable mainly insofar as it leads to accuracy of inference using the relevant variance–covariance matrix. The simulations in papers cited in this article suggest the following. First, no one estimator dominates others. This means in particular that the rate of convergence is not a sufficient statistic for performance in finite samples. The truncated estimator often and the autoregressive estimator sometimes perform more poorly than the slower converging QS estimator, which in turn sometimes performs more poorly than the still slower converging Bartlett estimator. Second, given that one decides to use QS or Bartlett, performance generally though not always is improved if one prewhitens and uses a data-dependent bandwidth as described in the recipes above. Third, the QS and Bartlett estimators tend to reject too much in the presence of positive serial correlation in h_t, and have what I read as a DGP dependent rejection rate (sometimes over-reject, sometimes under-reject) in the presence of negative serial correlation in h_t. The truncated estimator is much likelier to fail to be positive semidefinite in the presence of negative than positive serial correlation. Finally, the performance of all estimators leaves much to be desired. Plausible data-generating processes and sample sizes can lead to serious mis-sizing of any given estimator. Nominal 0.05 tests can have empirical size as low as 0.01 and higher than 0.25.

Some Recent Work

Because simulation studies have yielded disappointing performance, ongoing research aims to develop better estimators. I close by summarizing a few of many recently published papers.

1. I motivated my topic by observing that consistent estimation of S is a natural element of consistent estimation of the variance–covariance matrix of a GMM estimator. Typically we estimate the variance–covariance matrix because

we wish to construct confidence intervals or conduct hypothesis tests. A recent literature has evaluated inconsistent estimators that lead to well-defined test statistics, albeit statistics with non-standard critical values. These estimators set lag truncation (or bandwidth) equal to sample size. For example, for the Bartlett estimator, these estimators set $m = T - 1$ (see Kiefer et al. 2000; Kiefer and Vogelsang 2002). Simulation evidence indicates that the non-standard statistics may be better behaved than standard statistics. Jansson (2004) provides a theoretical rationale for improved performance in a special case, with more general results in Kiefer and Vogelsang (2005). Phillips et al. (2006, 2007) propose a related approach, which under some assumptions will yield statistics with standard critical values.

2. Politis and Romano (1995) propose what they call a 'trapezoidal' kernel. A trapezoidal kernel is a combination of the truncated and Bartlett kernels. For given truncation lag m, let $x_j = j/(m{+}1)$. Then for some c, $0 < c < 1$, the trapezoidal weights satisfy: $k_j = 1$ if $0 \le x_j \le c$, $k_j = (x_j - 1)/(c - 1)$ for $c < x_j \le 1$. Thus for $0 \le j \le c(m)$ 1), the autocovariances receive equal weight, as in the truncated kernel; for $c(m + 1) < j \le m + 1$, the weights on the autocovariances decline linearly to zero, as in the Bartlett kernel. Such kernels have the advantage that, like the truncated kernel, their convergence is rapid (near $T^{1/2}$). They share with the truncated kernel the possibility of not being positive semidefinite. The authors argue, however, that these kernels are better behaved in finite samples than is the truncated kernel.
3. Xiao and Linton (2002) propose 'twicing' kernels. Operationally, one first computes an estimate such as one of those described in Section 4. One also constructs a multiplicative bias correction by smoothing periodogram ordinates via a 'twiced' kernel. For a properly chosen bandwidth and kernel, the mean squared error of the estimator is of order $T^{8/9}$ (versus $T^{4/5}$ for the QS and $T^{2/3}$ for the Bartlett, absent any corrections). As well, Hirukawa's (2006) version of the Xiao and Linton estimator is positive semidefinite by construction. (The rate results for this estimator and that described in the previous paragraph do not contradict Andrews's 1991, optimality result for the QS kernel, because these procedures fall outside the class considered by Andrews.)

See Also

- ▶ Euler Equations
- ▶ Generalized Method of Moments Estimation
- ▶ Rational Expectations Models, Estimation of
- ▶ Spectral Analysis
- ▶ Time Series Analysis

Acknowledgment I thank Steven Durlauf, Masayuki Hirukawa and Tim Vogelsgang for helpful comments, and the National Science Foundation for financial support.

Bibliography

Andrews, D.W.K. 1991. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* 59: 817–858.

Andrews, D.W.K., and J.C. Monahan. 1992. An improved heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 60: 953–966.

Cumby, R.E., J. Huizanga, and M. Obstfeld. 1983. Two step, two-stage least squares estimation in models with rational expectations. *Journal of Econometrics* 21: 333–355.

den Haan, W.J., and A.T. Levin. 1997. A practitioner's guide to robust covariance matrix estimation. In *Handbook of statistics: Robust inference*, ed. G. Maddala and C. Rao, Vol. 15. New York: Elsevier.

Eichenbaum, M.S., L.P. Hansen, and K.J. Singleton. 1988. A time series analysis of representative agent models of consumption and leisure choice under uncertainty. *Quarterly Journal of Economics* 103: 51–78.

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

Hannan, E.J. 1970. *Multiple time series*. New York: Wiley.

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

Hansen, L.P., and R.J. Hodrick. 1980. Forward exchange rates as optimal predictors of future spot rates: An econometric analysis. *Journal of Political Economy* 96: 829–853.

Hansen, L.P., and K.J. Singleton. 1982. Generalized instrumental variables estimation of nonlinear rational expectations models. *Econometrica* 50: 1269–1286.

Hansen, B.E., and K.D. West. 2002. Generalized method of moments and macroeconomics. *Journal of Business and Economic Statistics* 20: 460–469.

Hirukawa, M. 2006. A modified nonparametric prewhitened covariance estimator. *Journal of Time Series Analysis* 27: 441–476.

Hodrick, R.J. 1992. Dividend yields and expected stock returns: Alternative procedures for inference and measurement. *Review of Financial Studies* 5: 357–386.

Jansson, M. 2004. The error in rejection probability of simple autocorrelation robust tests. *Econometrica* 72: 937–946.

Kiefer, N.M., and T.J. Vogelsang. 2002. Heteroskedasticity-autocorrelation robust standard errors using the Bartlett kernel without truncation. *Econometrica* 70: 2093–2095.

Kiefer, N.M., and T.J. Vogelsang. 2005. A new asymptotic theory for heteroskedasticity-autocorrelation robust tests. *Econometric Theory* 21: 1130–1164.

Kiefer, N.M., T.J. Vogelsang, and H. Bunzel. 2000. Simple robust testing of regression hypotheses. *Econometrica* 68: 695–714.

Newey, W.K., and K.D. West. 1987. A simple, positive semidefinite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55: 703–708.

Newey, W.K., and K.D. West. 1994. Automatic lag selection in covariance matrix estimation. *Review of Economic Studies* 61: 631–654.

Phillips, P.C.B., Y. Sun, and S. Jin. 2006. Spectral density estimation and robust hypothesis testing using steep origin kernels without truncation. *International Economic Review* 47: 837–894.

Phillips, P.C.B., Y. Sun, and S. Jin. 2007. Long run variance estimation and robust regression testing using sharp origin kernels with no truncation. *Journal of Statistical Planning and Inference* 137: 985–1023.

Politis, D.N., and J.P. Romano. 1995. Bias-corrected nonparametric spectral estimation. *Journal of Time Series Analysis* 16: 67–103.

Priestley, M.B. 1981. *Spectral Analysis and Time Series*. New York: Academic Press.

West, K.D. 1997. Another heteroskedasticity and autocorrelation consistent covariance matrix estimator. *Journal of Econometrics* 76: 171–191.

White, H. 1980. A heteroskedasticity consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 48: 817–838.

Xiao, Z., and O. Linton. 2002. A nonparametric prewhitened covariance estimator. *Journal of Time Series Analysis* 23: 215–250.

Hicks, John Richard (1904–1989)

Christopher Bliss

Abstract

This biographical review of the life and works of John Hicks covers his contributions to numerous fields, and in each case assesses the particular contributions for which he was responsible. The fields concerned are The Theory of Wages, Value Theory, Welfare Economics, The Keynesian Revolution, Monetary Theory, Growth and Capital Theory, and Other Topics. An extensive bibliography of Hicks's writings is provided. Two points that are stressed are the unusual departure point for Hicks's thought in the general equilibrium ideas of European economists, and the radical effect on Hicks of Keynes's ideas.

Keywords

Austrian economics; Capital measurement; Cardinal utility; Causality in economics; Chamberlin, E. H.; Compensation test; Consumer surplus; Consumption function; Economic dynamics; Economic history; Elasticity of expectations; Elasticity of substitution; Excess demand; Firm, theory of the; Fixprice versus flexiprice economy; General equilibrium; Growth theory; Hicks, J. R.; Hicks, U. K.; Hicksian stability condition; Hicksian substitution effect; Income measurement; Indirect taxation; Induction; Inflation; Intertemporal equilibrium; IS–LM model; Kaldor, N.; Keynes, J. M.; Marginal and average cost pricing; Marginal efficiency of capital; Methodology of econometrics; Monetary theory; Multiplier–accelerator interaction; Neoclassical synthesis; New welfare economics; Nominal income; Ordinal utility; Price theory; Probability; Public expenditure; Robinson, J. V.; Samuelson, P. A.; Scitovsky, T.; Simultaneous differential equation system; Slutsky, E.; Social accounting; Social income; Social welfare function; Temporary equilibrium; Traverse; Underconsumptionism; Value judgements; Value theory; Von Neumann model of capital accumulation causality; Wage flexibility; Wages theory; Waiting; Welfare economics

JEL Classifications
B31

Biography and Intellectual Development

Hicks was born in Warwick. He studied at Oxford (1922–1926) and taught at the London School of Economics (1926–1935). He was Professor at Manchester University (1935–1946), from where he moved to Oxford, first as Fellow of Nuffield College, and from 1952 until he retired, from teaching but not from writing, as Drummond Professor of Political Economy and Fellow of All Souls College. In 1935 he married Ursula Webb, a distinguished public finance specialist, and he collaborated with her in the preparation of numerous works on public finance, its theory and its application to various countries. Ursula Hicks, as she was subsequently known, died in 1985. John Hicks was a member of the Royal Commission on the Taxation of Profits and Income in 1951. He became a Fellow of the British Academy in 1942, a Knight in 1964, and was awarded the Nobel Prize in Economics (jointly with Kenneth J. Arrow) in 1972. He died in 1989.

Hicks was the product of a generation which was the last to produce in abundance all round economic theorists – economists who could turn their minds to almost any theoretical problem. Its leading lights, among whom Hicks is certainly to be counted, left their marks on most of the major new branches and issues of economics as these in turn attracted the interest of themselves and their contemporaries. Hicks's powerful and original mind first made itself felt in what is now called microeconomics, particularly in *The Theory of Wages* (1932, 2nd edition 1963) and with R.G.D. Allen, 'A Reconsideration of the Theory of Value' (*Economica*, 1934) and in welfare economics. However his best-known work, *Value and Capital* (1939), goes beyond microeconomics to offer an economic dynamics and discussion of monetary theory which reaches into the new macroeconomics.

Before Keynes's *General Theory* fundamentally altered the way in which economists viewed their subject, the theory of value, including the theory of the firm, shared the field with monetary theory. Hicks was first a value theorist, but he never neglected monetary theory, and it was an area to which he was frequently to return. It was a value theorist with an interest in monetary economics who provided in 'Mr Keynes and the "Classics" ' (*Econometrica*, 1937) an exposition of Keynes's *General Theory* that was probably more directly influential than the original. There followed work on the trade cycle, *A Contribution to the Theory of the Trade Cycle* (1950); on growth, *Capital and Growth* (1965); and an unusual approach to capital theory, *Capital and Time: A Neo-Austrian Theory* (1973).

Each decade of Hicks's life seemed to find him more eclectic and innovative than the last. Indeed, his willingness to speculate about and write on areas in which he had not seeped himself as a specialist was a notable feature of his later writing. Striking examples are *A Theory of Economic History* (1969), in which Hicks undertook the risks inherent in proposing a grand theory of economic history, and *Causality in Economics* (1979), in which he entered ground normally reserved for philosophers and statisticians. These works can be criticized, but as their author always commands a well-provisioned base camp in the economics which is his own, they are never merely amateurish. Hicks is an economist of outstanding breadth and erudition.

With hindsight it is remarkable that the author of such a formidable theoretical corpus should write ('Commentary' in the 1963 edition of *The Theory of Wages*, p. 306): '... at first I regarded myself as a labour economist, not a theoretical economist at all'. Lionel Robbins is given the credit for interesting Hicks in theory: '... he moved me from Cassel to Walras and Pareto, to Edgeworth and Taussig to Wicksell and the

Austrians – with all of whom I was more at home at that stage than I was with Marshall and Pigou' (p. 306). It would be foolish to attempt to explain why Hicks became the distinctive economist that he was to become. However the above snatches of autobiography probably go some way to explaining why *Value and Capital* turned out to be a book like no other that an English economist had written before.

Hicks's huge output (for the papers see the three-volume *Collected Essays on Economic Theory*, 1981–1983) is all the more remarkable when one considers that he seldom simply reacted to the work of others. There are no papers by Hicks pointing out mistakes by other writers, and none which embody minor changes to or extensions of existing models. Naturally Hicks produced work which follows paths opened up by others. However when he did so, as in *A Revision of Demand Theory* (1956), or with the famous IS–LM model, his approach was so distinctive that the commentary is recognizably a contribution of Hicks. Other writers feature mainly in footnotes and even such a powerful contribution as Samuelson's treatment of Walrasian stability earns no more than two pages in the Second Edition of *Value and Capital*. There is a streak of self-centredness and parochialism in Hicks which mirrors that to be found in other English economists of his generation and those before. It would be insufferable in an economist less gifted and genuinely self-critical.

The Theory of Wages

Writing later (1963) of the first edition of *The Theory of Wages* its author remarks that '... there has been no date this century to which the theory that I was putting out could have been more inappropriate.' However, Hicks was careful not to attribute the shortcomings of his first book to the misfortune of publishing in the worst year of the depression and a few years ahead of the reassessment of the theory of the firm brought about by the writings of Chamberlin and Joan Robinson and, worse fortune still, ahead of the *General Theory*. In this he was right. *The Theory of Wages* set out to examine the determination of wages under supply and demand in a competitive market. This admittedly limited task is important, and had it been perfectly accomplished it would not be sensible to criticize the resulting work for not solving other problems, such as wages under imperfect competition or the consequences of nominal wage bargaining, weighty though those problems might be. However the truth is that there were shortcomings in Hicks's treatment even given its chosen emphasis. It was not as good a book as Hicks was later to show that he could write, though it was surely a better book than the later Hicks's embarrassment at its shortcomings allowed him to admit.

G.F. Shove (whose fairly hostile review Hicks reprinted in the Second Edition) identified a number of the shortcomings. Notable among these is the relatively weak treatment of the supply side of labour markets and the consequently limited ability to treat unemployment. Shove also seems to accuse Hicks of failing to provide a treatment of the general equilibrium of many labour markets, which must be counted a rather common failing among labour economists. Shove, not surprisingly, was clear on minimum cost and the adding-up problem where Hicks's account needed improvement – it was after all Shove's bread and butter at the time. A point which Shove missed is that Hicks always discussed differences in the productivity of different workers as equivalent to differences in the quantity of effective labour provided per hour of work. In other words, like Marx before him, he fudged the problem of aggregating different types of labour.

These legitimate criticisms apart, there were very considerable merits. By concentrating on the long-run determinants of wage rates Hicks was able to examine some of the most interesting influences at work. He saw changes in the demand for labour as consisting of two components quite analogous to the income and substitution effects in demand that he was to investigate later. A lower wage rate leads to an expansion of output, because the cost curve has fallen, which induces a higher demand for labour. In addition a lower wage rate induces the adoption of more labour intensive methods of production, which increases the demand for labour for a given output. The analysis of this last effect lead to the discovery of the new concept of the elasticity of substitution, not quite as neat in Hicks's formulation as in Joan Robinson's later presentation, but this was the original. In general, Hicks's definition of the elasticity of substitution is different from Joan Robinson's, but the two are equivalent in the two-factor case. Many topics discussed only briefly and not deeply analysed were far ahead of their time. There is the idea that because capital tends to accumulate faster than labour, technical progress tends to be labour saving – the induced bias of technical progress as we would now say. There is the first ever attempt to model a labour dispute which may culminate in a strike, and more besides.

In a passing discussion in *The Theory of Wages* its author records a fascinating fact. Many wage rates in inter-War Britain were tied to the value of the output concerned, and for that reason were automatically flexible. Once account is taken of such arrangements, the remaining pure flexibility of money wages is exceedingly small. This provided an opportunity, not taken, to bring Hicks's analysis to bear on an event that must have impressed itself on the young Oxford undergraduate: the 1926 miners' strike that lead to the failed General Strike. Britain restored Sterling convertibility in 1925 at the pre-war rate of $4 to the pound. The resulting over-valuation of Sterling made much British industrial activity internationally uncompetitive. At the time the world price of coal in dollars had fallen sharply, with the consequence that British coal was worth less in dollars, and even less in over-valued Sterling. The coal miners' contracts required sharp cuts in their wages, for which reason they went on strike. Tying miners' wages to the price of coal implied too much wage flexibility in these circumstances. Britain's coal-mining sector needed to contract, which should have raised the marginal product of labour in terms of coal, where the existing contracts held that number constant.

Value Theory

This area and welfare economics are fields to which Hicks contributed the writings that would have made him a great economist if he had done nothing else. In making the 1972 Nobel Prize award to Hicks jointly with Arrow the Committee mentioned 'general equilibrium and welfare economics'. The reference in Hicks's case was clearly to *Value and Capital* on the one hand, and to the various papers which established the Kaldor-Hicks criterion in welfare economics on the other.

Hicks's paper with R.G.D. Allen, 'A Reconsideration of the Theory of Value' (1934) was written when both authors were at the London School of Economics, but its pedigree goes back to Slutsky, who had discovered the income and substitution effects in demand as early as 1915. However Slutsky's work was almost entirely unknown to economists in the West, and this included, as Hicks informs us, himself and Allen ('... I never saw Slutsky's work until my own was very far advanced, and some time after the substance of these chapters had been published in *Economica* by R.G.D, Allen and myself' (1939, p. 19).

Value and Capital is a work so rich in ideas that a short account of it cannot hope to do it justice. It showed that the basic results of consumer theory could be obtained from ordinal utility; it expounded what became known as the 'Hicksian substitution effect', obtained by varying income as relative prices changed so as to maintain an index of utility constant; it developed the parallel results for production theory; and it popularized among English speaking economists the notion of a general equilibrium of markets. Unlike Arrow, his fellow Nobel laureate, Hicks did not take the existence argument beyond equation and variable counting. There was about the Walrasian approach, Hicks concluded, '... a certain sterility' (1939, p. 60). The way to overcome this was to consider the 'laws of change' of a general equilibrium system. This lead Hicks to the first ever attempt to analyse the stability of a system of multiple exchange.

It is fascinating that both Hicks and Samuelson, working entirely independently, both came up with the idea that dynamics might rescue general equilibrium theory from emptiness. Paul Samuelson in various papers of the 1940s and in his *Foundations of Economic Analysis* (1947) adopted an entirely different approach from that of Hicks. Consider a system of *M* markets with prices p_1, p_2, ..., p_M and excess demands for the goods X_1, X_2, ..., X_M. Making the dependence of excess demands on all prices explicit, this system can be written as:

$$\begin{aligned} X_1(p_1,p_2,\ldots,p_M) &= 0 \\ X_2(p_1,p_2,\ldots,p_M) &= 0 \\ X_M(p_1,p_2,\ldots,p_M) &= 0 \end{aligned} \qquad (1)$$

In equilibrium prices are such that all excess demands are zero. Now consider one good, which may be taken without loss of generality to be good 1. Select any value for p_1 and suppose that there are unique values of the remaining prices such that the excess demands for goods 2 to M are zero. If the excess demands for the other goods are always maintained at zero by changes in their prices, all other prices become implicit functions of p_1. The Hicks stability condition is then the one that would be required of a single market – X_1 should decrease with p_1. Full stability requires that this condition should be satisfied for each good in turn.

At first sight the condition appears to be asymmetrical but as the condition must be satisfied by all goods, there is no genuine asymmetry involved. However each test does involve a certain kind of asymmetry, and this is what Samuelson objected to.

When we look at good 1 we implicitly assume that prices in other markets react more rapidly to disequilibrium than does the price of good 1. When we look at good 2 we make the same implicit assumption for the price of good 2, and so on. What Samuelson did was to make the time rate of change of each price a function of the excess demand in its own market hence arriving at the system of simultaneous differential equations:

$$\begin{aligned} dp_1/dt &= X_1(p_1,p_2,\ldots,p_M) \\ dp_2/dt &= X_2(p_1,p_2,\ldots,p_M) \\ &\cdots \\ dp_M/dt &= X_M(p_1,p_2,\ldots,p_M) \end{aligned} \qquad (2)$$

The Hicksian stability condition can be shown to be neither necessary nor sufficient for the stability of (2). Hicks however defended his own approach, on the ground that it answers a different but interesting question, in the Second Edition of *Value and Capital* (Additional note C).

Parts III and IV of *Value and Capital* record the effect of a road-to-Damascus- like change of vision by Hicks. It seems that while preparing his great work on price theory, Hicks read Keynes, and, to borrow a modern term, it blew his mind. He could no longer find any real satisfaction in the static formalism of Walrasian equilibrium theory, and what he then did shows the full extent of his originality. In these later Parts of the book that eventually resulted he adapted the static theory of the earlier parts to create an economic dynamics which borrowed equally from the Marshallian-Keynesian tradition of the short period and the Walras-Wicksell tradition of long-period equilibrium. The key idea was the concept of temporary equilibrium – an equilibrium of current markets in which future markets make their influence felt indirectly, through the expectations held by agents, which influence their behaviour in current markets. From this emerged the concept of the elasticity of expectations, an idea which proved to be crucial in much later work on macroeconomic theory.

Welfare Economics

Hicks's writings on welfare economics are largely accounted for by work on four closely connected fields of interest: the foundations of welfare economics, including the famous compensation test; the valuation of social income; the definition and measurement of consumer surplus; and, lastly, the measurement of capital.

Hicks was one of the pioneers of the 'new welfare economics', an approach which owed its inception to Kaldor's 'Welfare propositions in economics and interpersonal comparisons of utility' (Economic Journal 1939). The problem at issue is inescapable and fundamental to the justification of the recommendations of economists. By the time the debate arose, cardinal utility was no longer generally accepted and the need was felt to differentiate between 'scientific' propositions and 'value judgements'. The notion of a 'Pareto improvement' – a change that would make no individual worse off, and at least one better off – was familiar but was seen to be limited as a basis for recommendations, as nearly all actual changes made at least one person or group worse off. In Robbins's telling example, economists could not state scientifically that the abolition of the Corn Laws was a good thing because this reform made landlords worse off.

Hicks's suggested solution to the difficulty was the same as that proposed by Kaldor – a compensation test. A reform should be counted an improvement if the gainers could afford to compensate the losers and still be better off. In 'The Foundations of Welfare Economics' (*Economic Journal*, 1939), Hicks discussed the question of whether compensation must be paid for the improvement to count without a sense of how crucial this question was to prove to be. It was of course central to the issue posed by the Scitovsky example, which showed that the Kaldor-Hicks rule could lead to contradictory recommendations if compensation were not paid. A well-argued solution to this problem was proposed by I.M.-D. Little (1950), but this required explicit value judgements concerning whether income distribution had improved or not in a movement from one position to another, hence negating the original intention of the exercise, which had been to remove value judgements from welfare economics.

Hicks seemed to see these developments as fairly unimportant qualifications to the original idea. In 'The Measurement of Real Income' (1958), he writes of the 'new welfare economists'; 'They were indeed over-confident in their belief that they had found a means of direct comparison which will always work. But I still maintain that they did find a means of direct comparison which will often work' (reprinted in *Collected Essays*, Vol. I, p. 168). For a statement of Hicks's mature views on these questions see 'The scope and status of welfare economics' (1975). Perhaps the most interesting thing to notice about Hicks's long involvement with the foundations of welfare economics is that he seems never to have wholly accepted the conclusion upon which the majority of economists have been willing to settle. Briefly put, this view says that value judgements are an inescapable element in welfare evaluations and this should be accepted and the judgements made explicit. Hence the design of policy by the means of the maximization of an explicit social welfare function – the welfare weights of cost-benefit analysis – never engaged Hicks's interest.

It is evident that the problem of the measurement of income is closely allied to the issue of welfare improvements and Hicks, as would be expected, contributed to this area as well. Hicks discussed social accounting in his text book *The Social Framework* (1942), and the valuation of social income in a paper of that title in *Economica* (1940).

Hicks concluded that the measurement of income could mean measurement in terms of utility or measurement in terms of cost, and that the two measures were in general different. The most interesting issue to which this gave rise was the problem of how to treat indirect taxation and government expenditure on goods and services in the valuation of social income. This led Hicks into controversy with Kuznets (*Economica*, 1948; see also Essay 7 in Volume I of the *Collected Essays*). The usual practice is to measure prices at factor cost and to value public services at cost.

Hicks's original position may be briefly summarized as follows:

> (i) As there is no market test where public goods are concerned the taxation which pays for them is not a reliable measure of their value to the consumer; and (ii) even if consumers were to be regarded as implicitly choosing public expenditure exactly as they choose private expenditure, the appropriate price weights would not be average costs but marginal costs. For a mature statement, see the Addendum to Essay 7 in Volume I of the *Collected Essays*.

Between 1941 and 1946 Hicks published a number of papers on consumer surplus in the *Review of Economic Studies* that did much to revive interest in a concept which had seemed to lose its validity when measurable utility went out of fashion. His most important contribution to the controversial question of the measurement of capital, significantly entitled 'Measurement of Capital in Relation to the Measurement of Other Economic Aggregates', is in F.A. Lutz and D.C. Hague (1961).

The Keynesian Revolution and the Theory of Money

Hicks's first response to the *General Theory* is described in detail in 'Recollections and documents' (*Economica*, 1973, included in *Economic Perspectives*, 1977).

However the response for which he is best known was an expository piece 'Mr Keynes and the "Classics" ' (1937) that perfectly fulfilled the innate demand for a more readily accessible account of the essentials of Keynes's argument. It is important to make clear that what was provided was more than an haut vulgarization of Keynes, because the paper has been widely criticized for vulgarization and still more for seriously misrepresenting what the General Theory is about. This case has never been rigorously argued and it is hard to see how it could succeed. Hicks reproduced rather faithfully Keynes's various specifications, but by working with a two-sector model produced a framework which resulted in a simple diagram – the IS–LM diagram – which became to macroeconomic textbooks what the benzene ring diagram is to textbooks of organic chemistry. It is no surprise therefore that Keynes on reading the paper wrote to Hicks that he had '... next to nothing to offer by way of criticism'. Certainly there is more in the General Theory than just the IS–LM model. In particular there is the idea of a long-term under-consumption problem, no less worrying for being loosely formalized. Nevertheless, the IS–LM framework is there, as is what Samuelson later called the neoclassical synthesis, however much Keynes's latter-day disciples may dislike it.

In fact Hicks's way of presenting the argument is in some ways superior to that adopted in the *General Theory* because the original IS–LM model brings out very clearly how the relative price of capital and consumption goods enters into the determination of the solution – a point which is somewhat obscure in Keynes. How ironic therefore that one of the arguments later advanced against the IS–LM model, admittedly with simpler versions than Hicks's in mind, was that it omitted an essential feature of Keynes – relative prices of capital and other goods.

Hicks's IS curve is based on the striking observation that if the capital stocks in the two sectors of the economy are given, and if the money wage is known, then outputs in the two sectors depend on the nominal prices of their products through short-term profit maximization conditions. Given these outputs and prices, the value of nominal total income follows. The output of the investment sector depends on the rate of interest through the marginal efficiency of capital relation. Then, given the rate of interest, the nominal price of the investment good follows and the part of income generated in that sector. Now choose an arbitrary value, which can be thought of as a guess at the level of total nominal income. As the part of nominal income generated in the investment goods sector is known, given the rate of interest, the guess implies a certain level of nominal income to be generated in the consumption good sector. We now have a value of total income and a value of total consumption, both in nominal terms. If these values are consistent with the consumption function our guess for the value of total income was correct and we have discovered the level of income on the IS curve for the rate of interest with which we were working.

We have discussed only the IS curve but the LM curve is relatively uncomplicated – there is less going on behind it. The beauty of this elegant and lucid way of expounding Keynes's model is that it brings out clearly the vital role played in the model by aggregation assumptions which have the effect that the model decomposes, so that parts of it can be dealt with in partial isolation from the complete system. The simple specifications of the determinants of investment and the consumption function produce this result. The role played by income and working in terms of nominal values – which are equivalent to wage units, as the nominal wage has been taken as given – are all brought out clearly.

In the hands of others the IS–LM model often became merely a model of an economy with all prices fixed and was often misused, as when it was applied to long-run questions for which it is not suitable. However it made the *General Theory* intelligible to a whole generation, not because it left out the subtleties, it was never intended to substitute for the text, but because it perfectly captured the part of Keynes's message which is most amenable to formalization.

A Contribution to the Theory of the Trade Cycle (1950) provides an example of the type of model that explains cycles as the outcome of the interaction between the multiplier and the accelerator. These systems are linear in their simplest formulations when they lead to cycles which are almost certainly either damped or anti-damped. Three different ideas have been proposed to yield an outcome in conformity to the stylized model of a capitalist economy with regular cycles of constant amplitude.

The underlying solution may be anti-damped and buffers, in the form of a floor on or a ceiling to the level of economic activity, may be added to keep the solution within bounds. The system may be made non-linear, which is equivalent to buffers which make their influence felt continuously rather than abruptly. Finally, the underlying solution may be damped, in which case the cycle will have to be kept alive by the frequent intervention of random shocks. Hicks's main model embodies the last type of approach.

From 1937 Hicks continued to write regularly on questions of macroeconomics. Volume II of his *Collected Essays* contains a selection of his best work in this vein. Essay 18, 'Methods of dynamic analysis', proposes the distinction between the fixprice and the flexiprice economy which was to be developed in *Capital and Growth*. In his Yrjö Jahnsson lectures, *The Crisis in Keynesian Economics* (1974), Hicks offers reflections on the Keynesian theory and particularly on the impact of inflation on a Keynesian model.

Hicks never remained far from monetary theory. *Critical Essays in Monetary Theory* (1967), shows the richness of his early writings on monetary economics, while Essay 19 in Volume II of the *Collected Essays* gives a good indication of his later work. It is tempting to say that if Hicks had written nothing but his work on monetary economics he would be counted a considerable economist. However the truth is that he could not have written on monetary economics as he did write had he not been the broad economic theorist that he was. Hicks always placed monetary theory centrally in equilibrium theory. This was the distinctive idea of his first paper on the subject, 'A Suggestion for Simplifying the Theory of Money' (*Economica*, 1935), and it is a theme which he was to carry through all his later work.

Growth and Capital Theory

Hicks's two other books with 'Capital' in their titles, *Capital and Growth* (1965) and *Capital and Time: A Neo-Austrian theory* (1973a), have little else in common. *Capital and Growth* was Hicks's response to the frantic interest in growth theory which infected the 1960s. It was a characteristically personal response in which Hicks tried to apply the framework for dynamic analysis that he had developed in *Value and Capital* to the construction of a growth model.

The analogue of the static problem of Part I of *Value and Capital* was now the steady state growth path, but once again Hicks found the most interesting question to be the dynamic adjustment to equilibrium, and once again he attacked this problem with an approach which was all his own. The 'traverse' was the history of the movement of an economy from one steady state to another. This approach to growth theory was not very influential and the reason was not so much that the new interest in growth had extinguished interest in equilibrium theory. Rather it was that equilibrium theory and its sister economic dynamics had moved on a great deal since *Value and Capital*. Hicks, who had taught a generation how to do general equilibrium economics, was no longer talking a language that most economic theorists found congenial.

Capital and Time was not the product of the latest fashion in economic theory but was surely the result of long meditation starting from that wonderfully fruitful comparative ignorance of Marshall and Pigou as against the Austrians and other continentals, noted above. Hicks always conceded a place to the old classical idea that capital accumulation means more 'waiting'. In *Value and Capital* (1939a, pp. 197–8) however, he pointed out that the conclusion that the rate of interest is the marginal product of waiting is a special case of more general rules which apply to an intertemporal equilibrium. This conclusion, that Austrian models of capital are special cases of the more general von Neumann model of capital accumulation, remains valid. However special cases permit of special results, and Hicks's analysis of the Austrian model was remarkably successful in showing how that framework permits some strong and definite conclusions to be drawn.

Other Topics

We consider only *A Theory of Economic History* (1969) and *Causality in Economics* (1979), as these constitute the most audacious of Hicks's expeditions far from the mainstream of economic theory. A longer review of Hicks's work would have to find space to discuss his writings on economic policy (for a sample of which see *Essays in World Economics*, 1959) and on the history of economic thought (for some of which see Volume III of the Collected Essays), but here we merely note that these are serious omissions from the present survey.

We begin with *Causality in Economics*. This book was not the eventual product of long years of mental rumination, but the result, its author tells us frankly, of dissatisfaction with the 1974 International Economic Association conference on 'The Micro-foundations of Macroeconomics' which Hicks attended. It is book of interesting ideas on economics which are reluctantly regimented by a Sergeant-Major called 'causality'. This gentleman turns out to be only remotely related to the 'causation' of Aristotle or Kant. Hicks's definition of causality is reminiscent of Hume, but without the idea that the validity of induction is importantly involved.

Causation is seen as conjunction of events, possibly in a complex form. This idea is an old one and was very effectively criticized by the Cartesians but their contribution is not considered. As an essay in philosophy *Causality in Economics* cannot be taken seriously. The economics of course is of a higher standard. The last chapter provides a statement of Hicks's views on the meaning of probability and on econometric methodology. These are *obiter dicta*, not the fruits of profound investigation.

A Theory of Economic History is as ambitious a sortie into foreign territory as *Causality in Economics*, but is the product of more thought and reading and must be regarded as much more successful. The main idea, that economic history is tied up with the development of the market, is one that few would question. However most historians would be tempted to take cover behind a safe position according to which developments of ideas, knowledge, social institutions, etc., would all be seen as progressing in parallel with the development of the market, which consequently would enjoy no special status as a motive force. Put simply, Hicks's account gives a much more leading role to the market, although he does not of course go so far as to argue that the market drives history.

Such a strong argument could not fail to attract criticism, particularly from professional historians. A long book would have done the same but a very short book was a particularly provocative target. As the argument gave a lot of attention to the ancient world this proved to be a contested area. However while *A Theory of Economic History* was criticized it received respectful criticism. It may be only a way of looking at economic history but it was generally judged to be a good way. Hicks's reply to his critics may be found in *Economic Perspectives* (1977, pp. 181–4).

Retrospect

Schumpeter argued that the ideas of a great economist are more or less in place by the age of 40 – the rest is nurturing and polishing. At first glance Hicks appears to be an exception. He was 65, for example, when his theory of economic history was announced to the world. Yet probably on closer examination he will be seen to conform to the Schumpeter pattern. In the case of the *A Theory of Economic History* he tells us in the foreword that he had nursed the idea for years. There is indeed a powerful sense of direction to Hicks's intellectual journey. He often returns to old themes and new themes are examined from older perspectives. Probably after 40 Hicks was only nurturing and polishing, but it is no contradiction of that claim to say that the second half of his life produced some of his most creative work.

It remains to mention some particular qualities of Hicks the man. First, he wrote beautifully, in a style that is very correct from the formal point of view, yet almost conversational in its flow and ease. Secondly, his greatness justified a little vanity, and he was not wholly free of that minor vice. That said, he was always approachable, and he never attempted to win an argument by pulling rank or flaunting his formidable distinction.

Selected Works

1932. *The theory of wages*. 2nd edn. London: Macmillan, 1963. 1934 (With R.G.D. Allen.) A reconsideration of the theory of value. Pts I–II. *Economica* I, Pt I, February, 52–76; Pt II, May, 196–219.

1935. A suggestion for simplifying the theory of money. *Economica* 2 February, 1–19.

1937. Mr. Keynes and the 'classics'. *Econometrica* 5 April, 147–59. 1939a. *Value and capital*. Oxford: Clarendon Press.

1939b. The foundations of welfare economics. *Economic Journal* 49, 696–712. 1940. The valuation of the social income. *Economica* 7 May, 105–24.

1942. *The social framework*. Oxford: Clarendon Press.

1948. The valuation of the social income: A comment on Professor Kuznets' reflections. *Economica* 15 August, 163–72.

1950. *A contribution to the theory of the trade cycle*. Oxford: Clarendon Press. 1956. *A revision of demand theory.* Oxford: Clarendon Press.

1958. The measurement of real income. *Oxford Economic Papers* 10 June, 125–62. 1959. *Essays in world economics*. Oxford: Clarendon Press.

1961. Measurement of capital in relation to the measurement of other economic aggregates. In *The theory of capital*, ed. F.A. Lutz and D.C. Hague. London/New York: Macmillan/ St Martin's Press.

1965. *Capital and growth*. Oxford: Clarendon Press.

1967. *Critical essays in monetary theory.* Oxford: Clarendon Press. 1969. *A theory of economic history.* Oxford: Clarendon Press.

1973a. *Capital and time: A Neo-Austrian theory.* Oxford: Clarendon Press. 1973b. Recollections and documents. *Economica* 40 February, 2–11.

1974. *The crisis in Keynesian economics*. Oxford: Basil Blackwell.

1975. The scope and status of welfare economics. *Oxford Economic Papers* 27, 307–26.

1977. *Economic perspectives*. Oxford: Clarendon Press. 1979. *Causality in economics*. Oxford: Basil Blackwell.

1981–3. *Collected essays on economic theory.* Oxford: Basil Blackwell.

Bibliography

Kaldor, N. 1939. Welfare propositions in economics and interpersonal comparisons of utility. *Economic Journal* 49: 549–552.

Little, I.M.D. 1950. *A critique of welfare economics*. Oxford: Clarendon Press.

Samuelson, P.A. 1947. *Foundations of economic analysis*. Cambridge, MA: Harvard University Press.

Hicks, Ursula Kathleen (1896–1985)

Alan Peacock

Keywords

Direct and indirect taxation; Federal finance; Hicks, U. K.; Local public finance; Public finance; Tax incidence; Taxation of expenditure; Taxation of income

JEL Classifications

B31

An Irish-born economist specializing in public finance, Lady Hicks's long career spanned teaching and research at the London School of Economics and Political Science, University of Liverpool, and the University of Oxford (latterly as Foundation Fellow of Linacre College), as well as the holding of many visiting posts in foreign universities and service as a member of advisory missions on fiscal matters, notably in the Caribbean, India and Africa.

She made three significant contributions to her specialism, the theory and practice of public finance. Her paper 'The Terminology of Tax Analysis' (1946) questioned the usefulness of the distinction between direct and indirect taxes and argued persuasively for distinguishing between taxes on income and taxes on expenditure (outlay), the dichotomy now used in national accounting. She also explored the difference between the formal incidence of taxes (the liability to pay taxes) and the effective incidence (the determination of tax burdens). Second, in collaboration with her husband, Sir John Hicks, she endeavoured to produce coherence between what the aims of government should be and how fiscal institutions should be organized to achieve them (1947, Part 3). Third, she applied a unique knowledge of fiscal systems to the study of federal and local finance particularly in developing countries (for example, 1961).

No account of her contribution would be complete without mentioning her immense influence as a teacher of students of public finance from all parts of the world and her part in the foundation of the *Review of Economic Studies* (together with Abba Lerner and Paul Sweezy), of which she was Managing Editor from 1933 to 1961.

Selected Works

1946. The terminology of tax analysis. *Economic Journal* 56: 38–50.

1947. *Public finance*. London/Cambridge: Nisbet & Co./Cambridge University Press.

1961. *Development from below*. Oxford: Clarendon Press.

Bibliography

David, W.L. 1973. Introduction. In *Public finance, planning and economic development essays in honour of Ursula Hicks*, ed. W.L. David. London: Macmillan.

Hicksian and Marshallian Demands

Eugene Silberberg

Abstract

Soon after the presentation of demand in Alfred Marshall's *Principles of Economics* in 1890, a debate ensued concerning whether money income or some sort of real income should be held constant as the price of the good changed. By the mid-20th century, these two conceptions of a demand function became known as the Marshallian and Hicksian functions, respectively. The issue is critical to the interpretation of the area to the left of the demand curve between two prices as some sort of consumer surplus, that is, the gain from purchasing a good at the lower price.

Keywords

Compensated demand; Constrained maximum problem; Consumer surplus; Edgeworth, F.; Envelope theorem; Hicksian and Marshallian demand curves; Homogeneity; Income effect; Law of demand; Marginal utility of income; Marginal utility of money; Palgrave, R.; Pollack, R.; Roy's identity; Samuelson, P.; Shephard's lemma; Slutsky equation; Substitution effect; Uncompensated demand

JEL Classifications

D11

Although earlier writers had formulated the concept of a (downward sloping) demand curve, the analysis took on much great refinement with the publication of Alfred Marshall's *Principles of Economics* in 1890 (and continuing until 1920 with the eighth edition). In Chapter III, Marshall derived the law of demand from a postulate of diminishing marginal (cardinal) utility. He measured utility in terms of money, constantly reminding us, however, of the necessity to assume that the marginal utility of money remained constant. Although it would be reasonable to conclude that the demand function he had in mind is the standard formulation $x_i = x_i^*(p_1, \ldots, p_n, M)$, where the p_i's are the prices of the n goods and M is money income, Marshall never wrote out an expression such as the above. Although he was very clear that the demand curve represented diminishing marginal values of the good to an individual, he never specified the *ceteris paribus* conditions we are now familiar with.

However, as early as 1894 in the original *Dictionary of Political Economy* edited by R. Palgrave, Edgeworth gave the now current interpretation, stated above. (See the interesting footnote 5 in Friedman 1949). However, Marshall's suggestion that individuals purchase additional quantities only if the additional utility they gain is at least as great as the price paid suggests that the demand price represents the maximum the individual will pay for an additional unit. In that case, it would be utility, rather than money income, that was being held constant along the demand curve. To obscure things further, in Chapter VI of the same book, 'Consumer's surplus', Marshall insists that the adding up of demand prices to generate the consumer surplus, or net benefits of all the units purchased, is valid only when the marginal utility of income is constant, or the same across individuals within a market demand curve. These remarks about marginal utility spawned an industry of economists working on consumer surplus for the first 75 years of the 20th century. The matter finally came to an end in the 1970s when the derivations of the demand functions with either money income or utility as an argument were clarified.

The Marshallian Demand Functions

There are two main threads motivating the entire literature on Hicksian and Marshallian demands: first and foremost, consumer's surplus, and second, providing a rigorous discussion of the pure substitution term in the Slutsky equation. For convenience I limit the discussion to the case of two goods. The Marshallian demand functions are the solutions to the constrained maximum problem

$$\text{maximize } \; U = U(x_1, x_2)$$

$$\text{subject to } p_1 x_1 + p_2 x_2 = M$$

where, of course, x_1 and x_2 are two goods; their prices p_1 and p_2 and income M are assumed exogenous. The Lagrangian for this model is $L = U(x_1, x_2) + \lambda(M - p_1 x_1 - p_2 x_2)$. Differentiating partially with respect to x_1, x_2 and λ yields the necessary first-order conditions (NFOC)

$$L_1 = U_1(x_1, x_2) - \lambda p_1 = 0 \tag{1a}$$

$$L_2 = U_2(x_1, x_2) - \lambda p_2 = 0 \tag{1b}$$

$$L_\lambda = M - p_1 x_1 - p_2 x_2 = 0 \tag{1c}$$

The sufficient second-order condition (SSOC) is that the bordered Hessian determinant of L be positive:

$$H = \begin{vmatrix} U_{11} & U_{12} & -p_1 \\ U_{21} & U_{22} & -p_2 \\ -p_1 & -p_2 & 0 \end{vmatrix} > 0 \tag{2}$$

(In the case of n goods, the border-preserving principal minors of H alternate in sign. See, for example, Silberberg and Suen 2000.) On the assumption that the SSOC holds, the NFOC can be solved simultaneously for the demand functions with money income as a argument, now universally termed the Marshallian demand functions

$$x_1 = x_1^M(p_1, p_2, M) \tag{3a}$$

$$x_2 = x_2^M(p_1, p_2, M) \tag{3b}$$

and the Lagrange multiplier

$$\lambda = \lambda^M(p_1, p_2, M) \tag{3c}$$

In the parlance of intermediate microeconomics texts, when a price changes ‘money income is held constant’. But this is just an imprecise way of stating that the demand for any good is a function of the price of that good, the prices of all other relevant goods, and, in particular, money income.

Substituting the Marshallian demand functions (3a) and (3b) into the utility function yields the maximum utility for given prices and money income, $U^*(p_1, p_2, M)$. This is the indirect objective (utility) function for this model. By the envelope theorem (see Silberberg and Suen 2000)

$$U^*_{pi} = L_{pi} = -\lambda^M x_i^M i = 1, 2 \tag{4a}$$

$$U^*_M = L_M = \lambda^M \tag{4b}$$

Equation (4b) reveals that the Lagrange multiplier is the marginal utility of income. On the assumption that the constraint is preventing the consumer from gaining a higher utility, $\lambda^M > 0$. Equation (4a), known as Roy’s identity, shows that (maximum) utility varies inversely with price, as previously indicated, since consumption levels are assumed positive.

The traditional comparative statics of this model proceeds by substituting Eqs. (3) into the NFOC and differentiating with respect to M and, say, p_1. Since p_1 enters *two* of the first-order equations, two terms are produced in the expression for $\partial x_1^M / \partial p_1$. In 1916, Slutsky identified these terms as a substitution effect (which is always negative) and an income term. Rather than replicate these somewhat tedious calculations, we proceed to the more modern analysis.

The Hicksian Demand Functions

Consider now an alternative formulation of consumer behaviour, that of minimizing the expenditure needed to achieve a specified utility level at give prices:

$$\text{minimize } M = p_1 x_1 + p_2 x_2$$

$$\text{subject to } \; U(x_1, x_2) = U^0$$

The Lagrangian for this model is $L = p_1x_1 + p_2x_2 + \lambda(U^0 - U(x_1, x_2))$. Differentiating with respect to x_1, x_2 and λ as before produces the following NFOC and SSOC:

$$L_1 = p_1 - \lambda U_1(x_1, x_2) = 0 \tag{5a}$$

$$L_2 = p_2 - \lambda U_2(x_1, x_2) = 0 \tag{5b}$$

$$L_\lambda = U^0 - U(x_1, x_2) = 0 \tag{5c}$$

$$H = \begin{vmatrix} -\lambda U_{11} & -\lambda U_{12} & -U_1 \\ -\lambda U_{21} & -\lambda U_{22} & -U_2 \\ -U_1 & -U_2 & 0 \end{vmatrix} < 0 \tag{6}$$

On the assumption that the SSOC holds, Eqs. (5) can be solved simultaneously for the Hicksian demand functions

$$x_1 = x_1^U(p_1, p_2, U^0) \tag{7a}$$

$$x_2 = x_2^U(p_1, p_2, U^0) \tag{7b}$$

and the Lagrange multiplier

$$\lambda = \lambda^U(p_1, p_2, U^0) \tag{7c}$$

Eliminating λ from Eqs. (5a) and (5b) produces the same ‘tangency’ relation as eliminating λ from (1a) and (1b):

$$\frac{U_1}{U_2} = \frac{p_1}{p_2} \tag{8}$$

In both models, the consumer chooses the point on an indifference curve where the budget or expenditure line has the same slope as the indifference curve.

Alternatively, the consumer chooses a mix of goods such that

$$\frac{U_1}{p_1} = \frac{U_2}{p_2} \tag{9}$$

That is, the individual consumes each good until the marginal benefit (utility) per dollar is the same across all commodities. *At the margin*, all goods consumed are perfect substitutes. Given an increment of income, the consumer would be indifferent as to how to spend it, since he or she has already equalized the marginal utility of a dollar across all goods.

However, the comparative statics of these two models are not the same. For the Hicksian demand functions, when a price changes utility is held constant. That is, the consumer is constrained to slide along the same indifference curve. Thus, the responses to changes in prices are, by definition, the pure substitution effects. In the Marshallian case, as a price changes utility changes also, in the opposite direction.

Substituting the Hicksian functions (7a) and (7b) into the objective function, we obtain the expenditure function $M = M^*(p_1, p_2, U^0)$. This is the indirect objective function in this model; it gives the minimum expenditure needed to achieve utility level U^0 at prices p_1 and p_2. Since $M^*(p_1, p_2, U^0)$ is a minimum expenditure, by definition, $M^* \leq M$, so that the function $F = p_1x_1 + p_2x_2 - M^*(p_1, p_2, U^0)$ has a (constrained) minimum (of zero) with respect to not only x_1 and x_2, but also p_1, p_2, and U^0. The Lagrangian for this 'primal-dual' problem is

$$L = p_1x_1 + p_2x_2 - M^*\left(p_1, p_2, U^0\right) \\ + \lambda\left(U^0 - U(x_1, x_2)\right)$$

The first-order equations with respect to x_1, x_2 and λ are just Eqs. (5); with respect to p_1, p_2 and U^0 we have

$$L_{p_1} = F_{p_1} = x_1 = M^*_{P_1} = 0 \qquad (10a)$$

$$L_{p_2} = F_{p_2} = x_2 = M^*_{P_2} = 0 \qquad (10b)$$

$$L_{U^0} = M^*_{U^0} - \lambda = 0 \qquad (10c)$$

Equations (10a) and (10b) are sometimes referred to as Shephard's lemma (Shephard 1970): the Hicksian demand functions are the first partials of the expenditure function. Moreover, since p_1 and p_2 do not enter the constraint, F has an *unconstrained* minimum in p_1 and p_2. The second-order conditions include

$$F_{p_ip_i} = -M^*_{p_ip_i} \geq 0 \quad i = 1, 2 \qquad (11)$$

so that the expenditure function is concave in prices. However, since $M^*_{p_i} \equiv x_i^U\left(p_1, p_2, U^0\right)$,

$$\frac{\partial x_i^U}{\partial p_i} = M^*_{p_ip_i} \leq 0 \qquad (12)$$

The pure substitution effects, which are the ordinary slopes of the Hicksian demand curves, are negative. No such sign is implied for the Marshallian demand functions, since in the associated primal–dual problem, the prices enter the constraint, eliminating any implications about slope of the demand functions based on the curvature properties of the indirect utility function. Additionally, for the Hicksian demands, we find the reciprocity condition

$$M^*_{p_1p_2} = \partial x_1^U/\partial p_2 = \partial x_2^U/\partial p_1 = M^*_{p_2p_1} \qquad (13)$$

Homogeneity

The Marshallian demands are the solutions to Eq. (8) and the budget constraint (1c). Suppose there is a proportionate increase in prices and money income. That is, $p_1 \to tp_1$, $p_2 \to tp_2$ and $M \to tM$ where $t > 0$. But Eqs. (8) and (1c) are unchanged by this proportionate increase in parameters; hence their solutions must be identical. Thus, $x_i^M(tp_1, tp_2, tM) \equiv x_i^M(p_1, p_2, M)$; the Marshallian demand functions are homogeneous of degree zero in prices and money income. Consumers respond to changes in relative prices, not absolute prices. Similarly, the Hicksian demands are the solutions to Eqs. (8) and the constant utility constraint (5c). If $p_1 \to tp_1$ and $p_2 \to tp_2$, these equations are unchanged, and therefore $x_i^U\left(tp_1, tp_2, U^0\right) \equiv x_i^U\left(p_1, p_2, U^0\right)$. With the use of these properties, the indirect utility function must also be homogeneous of degree zero in prices and income, while the expenditure function is homogeneous of degree one in all prices.

The Slutsky Equation

Evgeny Slutsky (1916) published perhaps the seminal work in the economic theory of the consumer, in which he showed that a consumer's response to a change in price could be partitioned into two parts: a pure substitution effect which was always negative (that is, in the opposite direction to the price) and an income effect, whose sign was indeterminate. When a price, say p_1, decreases, the budget line pivots outward, intersecting the x_1 axis at a greater amount. Slutsky decreased the consumer's income by shifting the new flatter budget line back until it went through the original equilibrium. By such a 'compensation', Slutsky isolated the substitution effect. At various schools, particularly the University of Chicago, the Marshallian and Hicksian demand curves were referred to respectively as uncompensated and compensated demand curves. By the 1930s and 1940s, with the publications of John Hicks's *Value and Capital* (1939) and Paul Samuelson's *Foundations of Economic Analysis* (1947), the 'pure' substitution effect had become defined as the response to a price change when the budget line was shifted (parallel to itself) back to *the original indifference curve*, producing a response in consumption holding utility constant. (It was shown by J. Mosak and A. Wald, 1942, that, if p_1 changed by an amount Δp_1, the differences between the Slutsky and Hicks demands were of second-order smallness, that is, functions of powers of Δp_1 order two and higher. See also Silberberg and Suen 2000, p. 304.)

By the 1940s, the profession had largely settled on the Hicksian interpretation of the pure substitution effect (though Friedman 1949, stressed the operational advantage of Slutsky's measure). However, it was not until the 1970s that the demand functions (7a) and (7b) derived from constrained expenditure minimization came to be recognized as the analytical basis for the pure substitution effects. At that point, economists realized that the Slutsky equation showed the relationship between the Hicksian and Marshallian demand functions. We now demonstrate this using the concept of 'conditional demands,' first developed by Robert Pollak (1969).

The Hicksian demand function is obtained from the Marshallian function by adjusting money income, when a price changes, to the minimum amount necessary to keep the consumer on the same utility level. Stated mathematically, this is the identity (for x_1, say)

$$x_1^U(p_1, p_2, U^0) \equiv x_1^M(p_1, p_2, M^*(p_1, p_2, U^0)) \tag{14}$$

Differentiating this identity with respect to p_1,

$$\frac{\partial x_1^U}{\partial p_1} \equiv \frac{\partial x_1^M}{\partial p_1} + \frac{\partial x_1^M}{\partial M}\frac{\partial M^*}{\partial p_1} \tag{15}$$

Applying the envelope theorem (10a) and rearranging, we get the classic Slutsky equation

$$\frac{\partial x_1^M}{\partial p_1} \equiv \frac{\partial x_1^U}{\partial p_1} - x_1\frac{\partial x_1^M}{\partial M} \tag{16}$$

The slope of the Marshallian demand equals the slope of the Hicksian demand (which is always negative in its own price) and an indeterminate income effect. If, however, the good is non-inferior, that is, the income effect is non-negative, then the Marshallian demand curve is necessarily downward sloping.

Consumer's Surplus

Most of the interest in the distinction between Hicksian and Marshallian demand functions was generated by the analysis of consumer's surplus. Marshall developed the concept as follows. (I use hamburgers and dollars in place of Marshall's quainter example of tea and shillings.) Suppose, at a price of $10, a consumer will buy only one hamburger a month. At a price of $9, he will buy two; at $8, he will buy three, at $7, four, and so on. Since these prices measure the marginal values of hamburgers to this consumer at these consumption levels, we interpret these numbers as the maximum the consumer would pay for an additional hamburger. In that case, the amount the consumer would pay to consume four hamburgers rather than none would be \$10 + \$9 + \$8 + \$7 = \$34. Marshall thus interpreted the area under the demand curve as the all-or-nothing value of that quantity of a good, or the total utility of those units, measured in units of money income. Additionally, at a price of \$7, say, the consumer would spend \$28 on the good, leaving the area to the right of the demand curve above the price, \$6, as the consumer's surplus. This is the additional amount the consumer would have been willing to pay to consume the four units at a price of \$7.

The question is: when can we interpret the area to the left of a demand curve in this fashion? Consider a decrease in the price of x_1 from ${p_1}^0$ to ${p_1}^1$. Mathematically,

$$CS = -\int_{p_1^0}^{p_1^1} x_1 dp_1 \tag{17}$$

For both the Hicksian and Marshallian demand curves, we can calculate this area, which has the units of dollars – not utility – being price times quantity. However, the important question is: when we calculate this amount, what question, if any, does it answer? In the case of the Hicksian demands, the answer is clear. If we use the envelope relation (10a),

$$\begin{aligned} CS &= -\int_{p_1^0}^{p_1^1} x_1^U dp_1 = -\int_{p_1^0}^{p_1^1} \frac{\partial M^*}{\partial p_1} dp_1 \\ &= M^*\left(p_1^0, p_2, U^0\right) - M^*\left(p_1^1, p_2, U^0\right) \end{aligned} \tag{18}$$

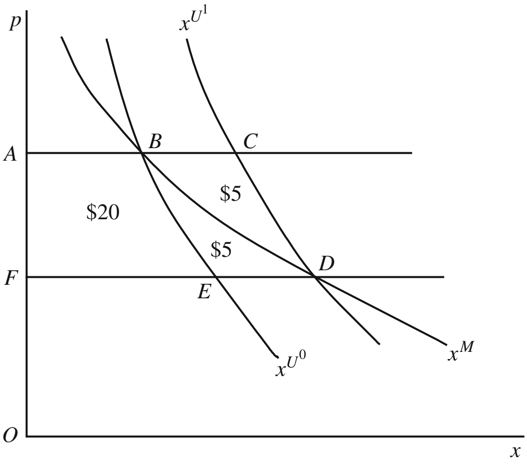
Because the Hicksian demands are the first partials of the expenditure function, the integral is simply the savings in expenditure the consumer enjoys when the price is lowered (or, likewise, the extra expenditure the consumer must make if the price increases). Thus the area to the left of a Hicksian demand curve is the amount consumers would be willing to pay, or have to be paid, to face the new price. Moreover, suppose two prices change. That is, suppose the price of x_1 changes from \$8 to \$4, and we calculate a CS_1 (\$18 if we use the above linear demand curve). The demand curve for x_2 will have now shifted. Suppose we now lower the price of x_2 from \$7 to \$3, generating CS_2 = \$15, say, producing $CS = CS_1 + CS_2$ = \$33. Suppose we did the experiment in the reverse order – lowering the price of x_2 first and then x_1. Would we get the same answer for total *CS*? Indeed we would:

$$\begin{aligned} CS &= -\int_{p^0}^{p^1} \sum x_i^U dp_i = -\int_{p_1^0}^{p_1^1} \sum \frac{\partial M^*}{\partial p_i} dp_i \\ &= -\int_{p^0}^{p^1} dM^* = M^*\left(p_1^0, p_2^0, U^0\right) \\ &\quad - M^*\left(p_1^1, p_2^1, U^0\right) \end{aligned} \tag{19}$$

Because of the reciprocity condition (13), this integral is exact; the path of price changes does not affect the value of the integral.

In the case of the Marshallian demands, no such interpretations are possible (Silberberg 1972). The Marshallian demands x_i^M are not the first partials of any function, so the area to the left of the demand curve given by (17) has no easy interpretation. Moreover, since for the Marshallian demands $\partial x_1^M / \partial p_2 \neq \partial x_2^M / \partial p_1$ (unless the utility function is homothetic) the integral corresponding to (19) for the Marshallian demand functions is path dependent. That is, depending on which price changes first, a different answer emerges, even if all the initial and final prices are identical in the two experiments. There is simply no unique measure of a change in utility in terms of income, except for some famous special cases. (See Silberberg and Suen, 2000).

Consider Fig. 1, for some good x. At the initial price *OA*, the consumer purchases *AB*. If the price decreases to *OF*, she moves along her Marshallian demand curve x^M, and consumes *FD*. If, however, we were to keep her on the same initial indifference curve U^0 as p decreased, she would move along the Hicksian demand curve to point *E*. Point *E* would be to the left of *D* if the good is normal (non-inferior), since we are eliminating this (positive) income effect. If however, the consumer



Hicksian and Marshallian Demands, Fig. 1

started at the lower price *OF* and the price were raised to *OA*, *and* we now held her on the higher level of utility U^1 she achieved at point *D*, she would move up along the Hicksian demand curve associated with U^1, leading her to point *C*. Suppose the area to the left of the Hicksian demand curve *BE* is $20, the area to the left of the Marshallian demand curve *BD* is $25, and the area to the left of the Hicksian demand curve *CD* is $30. What questions, if any, do these numbers answer? The amount the consumer would pay to face price *OF* instead of *OA* is $20. If the price were initially *OF*, the amount she would have to be paid to voluntarily face *OA* instead is $30. It seems odd, but it is true nonetheless that, for non-inferior goods, the amount one must be paid to face a higher price is greater than the amount we would pay to get the lower price. Lastly, there is simply no operational question for which $25 is the answer. However, Robert Willig (1976) investigated the actual empirical differences one would be likely to encounter; not too surprisingly, they turn out to be small.

See Also

- ▶ Envelope Theorem
- ▶ Le Chatelier Principle
- ▶ Marginal Utility of Money

Bibliography

Edgeworth, F. 1894. Demand curves. In *Dictionary of political economy*, ed. R. Palgrave. London: Macmillan.

Friedman, M. 1949. The Marshallian demand curve. *Journal of Political Economy* 57: 464–474.

Hicks, J. 1939. *Value and capital*. London: Oxford University Press.

Marshall, A. 1920. *Principles of economics*. 8 ed. London: Macmillan.

Mosak, J. 1942. On the interpretation of the fundamental equation of value theory. In *Studies in mathematical economics and econometrics*, ed. O. Lange, F. McIntyre, and T. Entema, 69–74 . Chicago: University of Chicago Press.esp. 73–74, n. 5, which contains a rigorous proof of this statement by A. Wald

Pollak, R. 1969. Conditional demand functions and consumption theory. *Quarterly Journal of Economics* 83: 60–78.

Samuelson, P. 1947. *Foundations of economic analysis*. Cambridge, MA: Harvard University Press.

Shephard, R. 1970. *The theory of cost and production functions*. Princeton: Princeton University Press.

Silberberg, E. 1972. Duality and the many consumer surpluses. *American Economic Review* 62: 942–956.

Silberberg, E., and W. Suen. 2000. *The structure of economics*. 3 ed. New York: Irwin/McGraw-Hill.

Slutsky, E. 1916. On the theory of the budget of the consumer. In *Readings in price theory*, ed. G. Stigler and K. Boulding, 1952. Homewood, IL: Richard D. Irwin, Inc..

Willig, R. 1976. Consumer's surplus without apology. *American Economic Review* 66: 589–597.

Hidden Actions, Moral Hazard and Contract Theory

Roger Guesnerie

'Moral hazard' in the literal sense refers to the adverse effects, from the insurance company's point of view, that insurance may have on the insuree's behaviour. As an extreme but standard example, a fire insurance holder may burn the property in order to obtain the insured sums. Although the expression can be found in earlier literature, its extensive use in economics can be dated from Arrow's *Essays in the Theory of Risk-bearing* (1971), which had a decisive influence in popularizing the term as well as in stimulating a systematic study both of the subject itself and of related phenomena. Arrow stresses that the complete set of markets required for first best efficiency often cannot be organized. The (so-called) Arrow–Debreu contracts which are needed would have to be contingent on states of nature. This term, 'states of nature', has to be taken in its meaning in decision theory where it refers to random events whose realization reflects an exogenous choice by 'Nature', and not an endogenous choice by agents. However, states of nature may not be observable either directly or indirectly, so that real contracts have to rely upon imperfect proxies. Take the overly simple fire insurance example: Arrow–Debreu contingent contracts would make indemnification conditional only on the occurrence of those natural events that can cause fire, such as thunderstorms, whereas actual real-world contracts make it dependent upon the occurrence of fire itself, whether due to an unusual exogenous event, or to a more normal exogenous event coupled with insufficient care.

Following Arrow, modern economic terminology has come to use 'moral hazard' to mean the unobservability of contingencies, about which information is needed in order to design first-best efficient contracts. Considering now a general framework of contracts, it is normally the case that the relevant contractual information can be obtained through observation of actions and outcomes, the latter themselves dependent on states of nature. Assuming that outcomes are always observed, moral hazard is therefore restricted to mean that some actions of one or more of the parties are not publicly observable (i.e. by all parties to the contract). With the more suggestive terminology of Arrow, moral hazard is thus associated with the existence of *hidden actions* in a contractual relationship.

This definition deserves three comments.

(i) 'Moral hazard' has unfortunate ethical connotations. Given that parties to contracts are usually modelled as standard maximizers of utility, it seems preferable to employ the term *hidden actions*.

(ii) Recent literature on contracts distinguishes between the observability and the verifiability of actions. A variable can be observable by all the parties to a contract, but not to outsiders to the contract. In particular, it may provide no evidence for a court of law. It is then said to be non-verifiable. Then *hidden actions* conveys the right idea but not the right nuance, and we should rather speak of *unverifiable actions*.

(iii) Difficulties in organizing a contractual relationship arise not only from actions that some parties can hide but also from the limited accessibility of the information that some parties use before taking actions. This may be private information of one party about itself (an agent usually knows his own characteristics better than do his partners in the contract), or information on some relevant states of nature which can influence the outcome of the relationship. Such difficulties are thus due to *hidden knowledge* as well as to hidden actions. Consider again an example drawn from insurance. Insurance companies (life insurance, car insurance) face both good risks and bad risks, i.e. agents who for a given level of care or prevention have to be assigned different probabilities of injury. This distinction thus refers to privately known characteristics of the insurees themselves rather than to the actions they take. Hidden knowledge generates opportunism. Faced with a set of contracts, high risk and low risk people will select different contracts; this is self-selection or, from the company's point of view, adverse selection. The distinction between hidden actions and hidden knowledge seems more suggestive than the more usual distinction in contract theory between moral hazard and adverse selection. Although we will examine some problems in which hidden actions and hidden knowledge are mixed, the main subject of this article is the analysis of contractual problems raised by hidden actions. Attention will be focused primarily on an abstract hidden action model, rather on the subject-specific discussions which generated the main building blocks of that model.

The Basic Hidden Action Model of a Bilateral Relationship

The prototype model considered in this section owes much to the pioneering work of Ross (1973), Mirrlees (1974, 1976) and Holmstrom (1979), and its presentation here draws heavily on the syntheses of Grossman and Hart (1983a). It is a principal–agent model with one principal and one agent. The agent chooses from among available actions one which together with random events (states of nature) determines a measurable result, which most of the time is a money payment to the principal. The principal is interested in the results as well as in the money remuneration he gives to the agent. The agent has a utility function depending upon the action taken and on the money transfer he receives from the principal. Some actions are more costly or involve higher 'effort'. Indeed, in many specific models the action variable is a loosely defined effort level: effort of the manager when the principal consists of shareholders, effort of firms when the principal is a bank.

In the simpler version of the model considered here, each utility function is separable, and risk-neutrality vis-á-vis income obtains, with a utility linear in income. It is assumed that the agents' actions are not observable but that the results are verifiable. A contract between the principal and the agent then consists of a reward schedule which associates a money transfer to any possible result. Analytically an optimal contract for the principal is a solution (if any) to a programme which maximizes over the set of all reward schedules, under a constraint of individual rationality for the agent.

The solution just sketched calls for preliminary comments:

(i) In the degenerate case where there is no choice of action – the principal–agent problem reduces to a pure risk-sharing question, whose solution depends on the risk-aversion of the parties concerned. In particular, a risk-neutral agent bears all income fluctuations and provides full insurance to his risk-averse partners. An optimal contract between a risk-neutral firm and risk-averse workers leads to utility profiles of workers constant across states of nature and a constant wage in states where workers are employed. This latter remark is at the core of the theory of implicit contracts initiated by Azariadis (1975), Bailey (1974) and Gordon (1974).

(ii) With a non-degenerate set of actions, but with an observability assumption, the optimal contract trades off between efficiency and risk-sharing considerations. Following the usual terminology, the corresponding contract is referred to as first best. When actions are not observable, the reward scheme has to be based on results only. It is generally impossible to reward actions indirectly in a way which mimics the first best contract. We then have to determine a second best contract.

First insights into the model are obtained when the reward schedule is restricted to be an affine function of the money outcome. Then, when the principal is risk-neutral and the agent is risk-averse, the optimal contract trades off between incentives and risk-sharing requirements in a way which confirms intuition. A positive fixed fee has to be combined with a linear schedule the slope of which is, however, smaller than the marginal value of the performance for the principal.

The derivation of the optimal non-linear second best contract leads to a serious analytical difficulty, which has been of primary concern to analysts. In the context of moral hazard this difficulty was initially stressed by Mirrlees (1975), and was independently discovered and analysed in the context of a general equilibrium second best problem by Guesnerie and Laffont (1978). It can be described as follows: For a given reward schedule, the agent's utility as an indirect function of actions is not generally quasi-concave. Hence, when the parameters of the reward schedule are modified the optimal response of the agent may jump. Although this jump only occurs for exceptional values of the parameters, it may still be the case that the optimal contract systematically selects such exceptional values (this is really the essence of the point made by Mirrlees and Guesnerie–Laffont). Then the local description of the agent's local behaviour from the first order conditions of utility maximization – which is analytically very convenient – becomes invalid. This failure of the so-called 'first-order approach' has generated contributions which are decisive for a rigorous analysis of the problem (see e.g. Rogerson 1985).

The research has led us to a much more thorough understanding of an optimal schedule. In particular, it has made clear that the reward associated to a given result reflects the Bayesian statistical inference made by the principal from this result, although this convenient interpretation should not hide the fact that the principal does not ignore the agent's action! However, the results on the shape of the optimal schedule are somewhat deceptive. As the statistical inference argument suggests, few restrictions on it can be deduced from general theory. Even monotonicity – higher rewards for higher results – cannot be guaranteed, without strong assumptions on the distribution of results conditional on actions. For example, monotonicity obtains with the monotone likelihood ratio property introduced in this problem by Milgrom (1981) and the concavity of distribution function condition (see Grossman and Hart 1983a). Non-monotonicity is hardly surprising; imagine that the most desirable actions from the principal's point of view give rise to high results and to low results with smaller probability but never to intermediate results. Conceivably, intermediate results will thus be less rewarded than low results.

In this rather disappointing picture, a result of general relevance does emerge. Although weak, it is remarkably robust. All variables that are correlated with the noise carry useful information for the design of optimal contracts. New information is redundant only when existing variables are sufficient statistics (see Holmstrom 1979; Gjesdal 1982).

To complete the picture, cases where the first best is implementable have to be stressed.

(i) If the agent is risk-neutral, a reward schedule which gives him the money result up to some constant provides correct incentives (such a reward schedule is reminiscent of the Groves scheme in an adverse selection problem). The agent then acts as a residual claimant and chooses the first best action.

(ii) Suppose that one result signals for sure that some non-optimal action has been chosen (i.e. this result has probability zero when the optimal action is taken). Then, if a high penalty is associated with this result, the agent will be deterred from choosing any action for which this result can occur with non-zero probability. It follows that the first-best action will be chosen if there is a subset of highly penalized results that are reached with probability zero when the optimal action is taken, and with positive probability when any non-optimal action is taken. In particular, if the result is a noisy estimate of the action, the first best is implementable when the noise is additive and has compact support. The power of high penalties, at least in some contexts, is a striking feature of moral hazard problems. We will come back to this point later.

Further General Considerations on Hidden Action Models

We will examine briefly four directions of development for the basic hidden action model described in section "The Basic Hidden Action Model of a Bilateral Relationship".

The Complexity of the Optimal Reward Schedule

The results described in the previous section suggest contractual arrangements which are more complex than those observed in real situations. Several explanations have been suggested: for example, bounded rationality of the parties is a plausible argument for the use of unsophisticated reward schemes. Another possible explanation might be found in the inadequacy of the modelling options described in section "The Basic Hidden Action Model of a Bilateral Relationship". This is a subject of current research and an interesting point has recently been made by Holmstrom and Milgrom (1985). They modify the basic model by assuming that the agent has progressive information on the occurrence of the outcome so that he can continuously adapt his action (here, his effort) in the time interval where the relationship takes place. They show that the optimal reward schedule, which in the standard version of these problems is highly non-linear, becomes linear. Although this conclusion relies on special assumptions concerning the agent's utility and the noise, it suggests that the enrichment of the action space of the agent leads to simpler reward schedules.

Mixing Adverse Selection and Moral Hazard

It has been argued above that hidden action and hidden knowledge determine two polar cases in the theory of contracts – in fact many contract problems involve both hidden action and hidden knowledge. In the mixed case the non-linear reward scheme thus has three different roles. It should provide correct incentives by limiting the distortion between the value of outcome for the principal and the agent's reward, and should induce adequate risk-sharing; these two functions are already central to the hidden action model. In addition, it should keep control of the self-selection process by inducing satisfactory choices of agents of different characteristics. The determination of the optimal contract in the mixed case then assimilates the analytical difficulties of each of the polar cases (each of these polar cases is reasonably well understood, and for a synthesis on an adverse selection principal-agent problem, in a spirit similar to Grossman and Hart's article on moral hazard, see Guesnerie and Laffont (1984). The understanding of the intricacies of the general case requires further investigation. The analysis of an intermediate case provides a useful benchmark. It is presented now.

Consider a pure hidden knowledge problem when actions of the agent are observable although characteristics are not. Let us introduce the moral hazard ingredient that actions are no longer perfectly observable. Their observation is affected by noise. The new problem calls for two immediate remarks: first, if the parties are risk-averse, the introduction of noise will reduce social welfare (when compared to its pure adverse selection maximum level); second, if the adverse selection problem is degenerate, i.e. the agent's characteristics are known, there is no welfare loss when agents are risk-neutral. This absence of welfare loss can be shown to extend to a non-degenerate hidden knowledge model. For a large class of noises, with risk-neutral agents, the maximum adverse selection welfare, can be at least approximately reached when the observation of actions becomes noisy. In other words, the second best adverse selection welfare can still be implemented with noisy observations. This (quasi) implementation obtains either by using a family of quadratic schedules (see Picard 1987) or by using a single schedule, different from the adverse selection optimal schedule, but obtained from it as the solution of a convolution equation when the noise is additive (see Caillaud et al. 1986) or a Fredholm equation for non-additive noise (see Melumad and Reichelstein 1985). Furthermore, when one of the action variables can be observed, a family of linear schedules may serve for implementation whatever the distribution of noise (it is then a universal family of schedules) or a family of truncated linear schedules may serve for implementation when the noise is small. However, these appealing properties are likely to hold in circumstances which are rather special (see Laffont and Tirole (1986) for one of these special cases, and Caillaud et al. (1986) for a comprehensive analysis of this problem).

Monitoring Devices and High Penalties

We have provided an interpretation of the basic hidden action model where 'results' are an intrinsic and unavoidable outcome of the relationship. There are cases, however, where the principal is only interested in the actions taken by the agents and where the inference on the action is made from observations which are obtained from a special device: examples of such monitoring devices which allow more precise inference of the behaviour of an agent are audits. If the basic frame is easily adapted to the study of such a situation when the monitoring device is given, or even if there are several possible monitoring devices, a basic difficulty occurs when the frequency of use of such a monitoring device is not fixed. The nature of the difficulty is the following: the use of the monitoring device being costly, the principal can economize on expected costs by writing a contract which stipulates that the control device will only be used with probability smaller than one, rather than for sure. But whatever the probability chosen, it is often the case that the principal can reduce it further and modify penalties and rewards accordingly in such a way that the choice of action is unchanged. This argument was made in particular by Polinsky and Shavell (1979; see also Rubinstein 1979) who were considering the substitutability between the probability and the magnitude of legal fines. The fact is that the expected value of the fine may be held constant when the probability of control is decreased and the magnitude of the fine is increased. This argument has proved to be remarkably robust (for extensions, see Nalebuff and Scharfstein 1985). In particular, it does not depend upon the risk-aversion of agents, at least in a bilateral relationship. In our framework, it suggests that the optimal contract, when the use of monitoring devices is costly, may be stochastic and may involve a low probability of control, together with (possibly) high penalties and rewards. Again, real contractual arrangements do suggest neither the use of high penalties, nor the substitution of penalties to control frequencies at least to the extent predicted by the above theory. A more careful analysis of the problems suggests at least three reasons for the first noted discrepancies (and at the same time three directions of improvements for the basic model).

(i) Our argument holds in the special case of a hidden action relationship in which all the elements of the problem are in the language of the theory, 'common knowledge'. For example, it assumes that the agent's risk-aversion is exactly known by the principal or that the distribution functions of the random variables is common knowledge. Giving up one of these assumptions amounts to introducing hidden knowledge into the relationship. The efficiency of high penalties does not seem to be robust to the introduction of these considerations.

(ii) The credibility of the principal's commitment to some probability of control is problematic. It would require the implementation of some kind of public lottery.

(iii) The outcome of control via a monitoring device should be verifiable. If not, the principal would have an incentive to announce results which highly penalize the agent. Some neutral third party is required. But the danger of collusion between this third party and another party increases with the amount of penalties (or reward). (For an analysis of collusion in a three parties relationship, see Tirole 1986).

The Dynamics of Moral Hazard Contracts

Assume that the basic principal–agent relationship is repeated. The one-period game described above is extended to a large number of periods (assuming for simplicity separability between periods). It is intuitively clear from the law of large numbers that time filters out uncertainty and allows a more and more accurate knowledge of the mean action taken by the agent. Repetition should thus alleviate moral hazard problems. The formal analysis confirms and makes precise these findings, at least when parties to the contract put enough weight on the future. If agents are interested only in the average pay-off over an infinity of periods or if both have a (common) discount rate close enough to one, there exist dynamic contracts which allow one to approximate the first best welfare level (see Radner 1981, 1985). It would, however, be premature to conclude from this neat result that moral hazard problems disappear within a long enough relationship. Let us make clear the limits of this result.

(i) The result only holds for discount rates close to one. Even then, it does not provide a characterization of the truly optimal policy (it uses an a priori policy which is shown to be quasioptimal). *A fortiori*, the characteristics of the optimal policy for lower values of the discount rates are not well understood. The study of simple cases such as the one considered by Henriet and Rochet (1984) suggests that the present reward at any period should put more weight on observed performances which are more recent. This is in sharp contrast with what happens in an adverse selection problem, where the observation serves to estimate the value of unobservable characteristics, a case in which the Henriet–Rochet model leads one to base the reward on the mean of observation before the present period.

(ii) The model supposes both that the principal can commit himself to the announced strategy and that the agent is locked-in in the relationship, but is not necessarily needed for the conclusion. In addition, the principal's policy relies on the threat of high penalties, a feature of contracts the adequacy of which has been questioned in the previous subsection. The commitment assumption is subject to the usual objections. The lock-in assumption for the agent is also much debatable. The agent should at least be allowed to smooth his income through time by access to financial markets. Exit of the agent via financial markets is a subject of present research.

(iii) As in the static case, many interesting dynamic problems mix hidden action and hidden knowledge. This leads to more intricate phenomena as demonstrated by the models of Holmstrom (1982b) or Harris and Holmstrom (1982). Assume as in these models, the output of a worker is the product of an unobservable characteristic (say skill) and of an action (say effort). The firms' inference from the sequence of outputs aims at determining both effort and skill. In their turn, workers are induced to over-invest in effort in the first periods to signal high skill and to under-invest when their position has been established. This has some resemblance to real academic life rather than to a pure hidden action dynamic model.

Tournaments and Moral Hazard in a Group

The so-called tournament model focuses attention on a relationship involving one principal and several agents. With several agents, the contracts are not necessarily independent: the reward of one agent can be based not only upon his performance but also upon the performance of the other. One polar case of interdependent contract is the contract associated with a rank-order tournament where actual outputs are ranked and the reward jumps with the rank. With two agents the winner has the highest prize (R&D competition for patents induces a similar structure of rewards: see Guesnerie and Tirole 1985). Let us briefly mention the main direction explored by the tournament literature.

(i) Lazear and Rosen (1981), in a two-agent model, compare the rank-order tournament with special independent contracts, i.e. linear contracts, and discuss the relative merits of both.

(ii) Independent non-linear contracts are dominated by dependent contracts only when the principal can infer more information on the variables faced by the agent (before his decision was made) from the whole set of outcomes than he can infer from any single outcome. In such circumstances Green and Stokey (1983), Holmstrom (1982a), and Nalebuff and Stiglitz (1983) focus attention on situations in which the mean of outcomes is a sufficient statistic for the variables unknown to the principal. Thus, the optimal contract is only dependent upon the mean of outcomes and the individual outcome.

(iii) First best can be approximately implemented from rank order tournaments with high penalties when the number of participants is large enough (see Holmstrom 1982; Nalebuff and Stiglitz 1983). In the different but related context of moral hazard in teams, Holmstrom (1982) has stressed that a team can behave poorly in the solution of moral hazard problems when no agent in the group can act as residual claimant. The group thus cannot commit itself credibly to use a sharing rule which induces efficient effort. The existence of a residual claimant is essential for making credible the threat of destruction.

Conclusion

One of the two obvious omissions in the present review has already been stressed. Applications of the basic ideas to different subjects have not been reviewed. The 'horizontal' presentation adopted here should be complemented by 'vertical' readings which describe the implications of the basic ideas in different fields. The second omission is the fact that the work reviewed is only of partial equilibrium nature. However, the hidden action model is part of the contractual approach to economics which has developed since the 1970s from a recognition of the failure of the impersonal market hypothesis to explain certain phenomena. The corresponding literature had the more or less explicit ambition of assessing the aggregate implications of the existence of contractual arrangements at the micro level. In particular, the study of the general equilibrium implications of moral hazard is an important topic. It has not been presented here, partly from lack of space, and partly because a coherent presentation of existing work is more difficult.

In conclusion, let us briefly mention a number of directions of present research.

First, the nature of competition is affected by the presence of moral hazard at the micro level. Helpman and Laffont (1975), Arnott and Stiglitz (1985), and Hellwig (1987) analyse this problem.

Second, normative economics should take into account the specification of contractual relationships. In particular, one can expect that the contractual approach will favour a better assessment of the informational constraints faced by government action. Also, moral hazard at the micro level is responsible for externalities, the particular features of which are analysed in the case of the labour market by Arnott and Stiglitz (1985).

Finally, the examination of the aggregate consequences of contractual arrangements in the labour market is a subject of intensive research – Shapiro and Stiglitz (1984) argue that in the absence of direct penalties (for reasons discussed above) for breach of labour contracts, unemployment serves as a 'discipline device'. Other work on the general equilibrium consequences of the contractual labour conditions – in case of hidden action – include Malcomson and MacLeod (1986).

See Also

- ▶ Adverse Selection
- ▶ Asymmetric Information
- ▶ Externalities
- ▶ Incomplete Contracts
- ▶ Moral Hazard
- ▶ Principal and Agent (i)
- ▶ Signalling

Bibliography

Arnott, R., and J. Stiglitz. 1985. Labor turnover, wage structures, and moral hazard: The inefficiency of competitive markets. *Journal of Labor Economics* 3: 434–462.

Arrow, K.J. 1964. *Essays on the theory of risk-bearing*. Chicago: Aldine.

Arrow, K.J. 1985. The economics of agency. In *Principals and agents: The structure of business*, ed. J. Pratt and R. Zeckhauser, 37–51. Boston: Harvard Business School Press.

Azariadis, C. 1975. Implicit contracts and underemployment equilibria. *Journal of Political Economy* 83: 1183–1202.

Bailey, M. 1974. Wages and employment under uncertain demand. *Review of Economic Studies* 41: 37–50.

Bester, H. 1985. Screening versus rationing in credit markets with imperfect information. *American Economic Review* 75(4): 850–855.

Bhattacharya, S. 1983. *Tournaments and incentives: Heterogeneity and essentiality*, Research paper no. 695. Stanford: Graduate School of Business, Stanford University.

Caillaud, B., R. Guesnerie, and P. Rey. 1986. *Contracts with adverse selection and moral hazard: The case of risk neutral partners*. New York: Mimeo.

Calvo, G., and S. Wellicz. 1978. Supervision, loss of control and the optimal size of the firm. *Journal of Political Economy* 86(5): 943–952.

Diamond, D. 1984. Financial intermediation and delegated monitoring. *Review of Economic Studies* 51(3): 393–414.

Fama, E. 1980. Agency problems and the theory of the firm. *Journal of Political Economy* 88: 268–307.

Gibbons, R. 1985. Essays on labor markets and internal organization. Unpublished dissertation, Stanford University, July.

Gjesdal, F. 1982. Information and incentives: The agency information problem. *Review of Economic Studies* 49: 373–390.

Gordon, D. 1974. A neo-classical theory of Keynesian unemployment. *Economic Inquiry* 12: 431–459.

Green, J., and N. Stokey. 1983. A comparison of tournaments and contracts. *Journal of Political Economy* 91: 349–364.

Grossman, S., and O. Hart. 1983a. An analysis of the principal-agent problem. *Econometrica* 51: 7–45.

Grossman, S., and O. Hart. 1983b. Implicit contracts under asymmetric information. *Quarterly Journal of Economics*, Supplement, 71: 123–157.

Guesnerie, R., and J.J. Laffont. 1978. Taxing price makers. *Journal of Economic Theory* 19(2): 423–455.

Guesnerie, R., and J.J. Laffont. 1984. A complete solution to a class of principal-agent problem with an application to a self managed firm. *Journal of Public Economics* 25(3): 329–369.

Guesnerie, R., and J. Tirole. 1985. L'économie de la recherche développement. *Revue économique* 5: 843–871.

Harris, M., and B. Holmstrom. 1982. A theory of wage dynamics. *Review of Economic Studies* 49: 315–333.

Hellwig, M. 1987. Some recent developments in the theory of competition in markets with adverse selection. *European Economic Review* 31(1/2): 319–325.

Helpman, E., and J.J. Laffont. 1975. On moral hazard in general equilibrium theory. *Journal of Economic Theory* 10(1): 8–23.

Henriet, D., and J.C. Rochet. 1984. *The logic of bonus-penalty systems in automobile insurance*, Working paper no. A273 0784. Palaiseau: Ecole Polytechnique.

Holmstrom, B. 1979. Moral hazard and observability. *Bell Journal of Economics* 10: 74–91.

Holmstrom, B. 1982a. Moral hazard in teams. *Bell Journal of Economics* 13: 324–340.

Holmstrom, B. 1982b. Managerial incentive problems – A dynamic perspective. In *Essays in economics and management in honor of Lars Wahlbeck*. Helsinki: Swedish School of Economics.

Holmstrom, B. 1983. Equilibrium long-term contracts. *Quarterly Journal of Economics*, Supplement, 98: 23–54.

Holmstrom, B., and P. Milgrom. 1985. Aggregation and linearity in the provision of intertemporal incentives. Cowles discussion paper no. 742, Apr.

Holmstrom, B., and J. Ricart-Costa. 1984. Managerial incentives and capital management. Cowles discussion paper no. 729, Nov.

Joskow, P. 1985. Vertical integration and long-term contracts. *Journal of Law, Economics, and Organization* 1: 33–80.

Laffont, J.J., and J. Tirole. 1986. Using cost observation to regulate firms. *Journal of Political Economy* 94(3), Pt I: 614–641.

Lambert, R. 1986. Executive effort and selection of risky projects. *Rand Journal of Economics* 16: 77–88.

Lazear, E., and S. Rosen. 1981. Rank order tournaments on optimum labour contracts. *Journal of Political Economy* 89(5): 841–864.

Malcomson, J. 1984. Work incentives, hierarchy, and internal labor markets. *Journal of Political Economy* 92(3): 486–507.

Melumad, N., and S. Reichelstein. 1985. *Value of communication in agencies*. New York: Mimeo.

Milgrom, P. 1981. Good news and bad news: Representation theorems and applications. *Bell Journal of Economics* 12: 380–391.

Mirrlees, J. 1974. Notes on welfare economics, information and uncertainty. In *Essays in economic behavior under uncertainty*, ed. M. Balch, D. McFadden, and S. Wu, 243–258. Amsterdam: North Holland.

Mirrlees, J. 1975. *The theory of moral hazard and unobservable behavior – Part I*. New York/Oxford: Mimeo/Nuffield College.

Mirrlees, J. 1976. The optimal structure of authority and incentives within an organization. *Bell Journal of Economics* 7: 105–131.

Mookherjee, D. 1984. Optimal incentives schemes with many agents. *Review of Economic Studies* 51(3): 433–446.

Nalebuff, B., and J. Stiglitz. 1983. Prizes and incentives: Towards a general theory of compensation and competition. *Bell Journal of Economics* 13: 21–43.

Newbery, D., and J. Stiglitz. 1983. *Wage rigidity, implicit contracts and economic efficiency: Are market wages too flexible?* Economic theory discussion paper 68. Cambridge, MA: Cambridge University.

Picard, P. 1987. On the design of incentives schemes under moral hazard and adverse selection. *Journal of Public Economics* 33: 305–331.

Polinsky, A., and S. Shavell. 1979. The optimal tradeoff between the probability and the magnitude of fines. *American Economic Review* 69(5): 880–889.

Radner, R. 1981. Monitoring cooperative agreements in a repeated principal–agent relationship. *Econometrica* 49: 1127–1148.

Radner, R. 1985. Repeated principal–agent games with discounting. *Econometrica* 53: 1173–1198.

Rogerson, W. 1985. The first-order approach to principal-agent problems. *Econometrica* 53: 1357–1368.

Ross, S. 1973. The economic theory of agency: The principal's problem. *American Economic Review* 63: 134–139.

Rubinstein, A. 1979. Offenses that may have been committed by accident – An optimal policy of retribution. In *Applied game theory*, ed. S. Brahms, A. Shotter, and G. Schwödiauer, 406–413. Würtzburg: Physica-Verlag.

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

Shavell, S. 1979. Risk sharing and incentives in the principal and agent relationship. *Bell Journal of Economics* 10: 55–73.

Stiglitz, J., and A. Weiss. 1985. Credit rationing in markets with imperfect information. *American Economic Review* 71(3): 393–410.

Stiglitz, J. 1974. Incentives and risk sharing in sharecropping. *Review of Economic Studies* 41(2): 219–255.

Tirole, J. 1986. Hierarchies and bureaucracies: On the role of collusion in organizations. *Journal of Law, Economics, and Organization* 2(2): 181–214.

Williamson, O. 1985. *The economic institutions of capitalism*. New York: Free Press.

Yaari, M. 1976. A law of large numbers in the theory of consumer's choice under uncertainty. *Journal of Economic Theory* 12: 202–217.

Hierarchical Bayes Models

Siddhartha Chib and Edward Greenberg

Abstract

The standard Bayesian model is defined in terms of an outcome model and the prior density of the parameters. The latter depends on parameters called hyperparameters. A hierarchical Bayes model results when one or more of the hyperparameters are assumed to be random and modelled probabilistically. We discuss canonical versions of these models for the case when both the parameters and the hyperparameters are modelled in groups or blocks, provide relevant examples, and discuss how inference by Markov chain Monte Carlo methods makes even the fitting of complex hierarchical models practical and simple. The problem of model comparisons is also addressed.

Keywords

Bayes' th; Component densities; Exchangeability; Hierarchical Bayes models; Hyperparameters; Marginal likelihood; Markov chain Monte Carlo methods

JEL Classifications

C11

Suppose that y is a univariate random variable or multivariate random vector and θ is a d-dimensional parameter vector that lies in $\mathscr{D}$, a subset of $\mathscr{R}^d$. The standard Bayesian model is then defined in terms of the density of y given θ (the outcome model) and the prior density of θ (the prior model). Specifically, the Bayesian model is specified as

$$y|\theta \sim p(y|\theta) \quad (\text{outcome model : stage } 1) \tag{1}$$

$$\theta|\gamma \sim \pi(\theta|\gamma) \quad (\text{prior model : stage } 2) \tag{2}$$

where γ is the vector of parameters in the prior density. These are called hyperparameters. We can assume that γ is g-dimensional and lies in $\mathscr{G}$, a subset $\mathscr{R}^g$. The labelling of the outcome model as stage 1 and the prior model as stage 2 is arbitrary, and the numbering can be reversed. The outcome model may be called the top or bottom level of the model because this difference in nomenclature has no significance.

Suppose that the researcher is not able to specify one or more of the hyperparameters in γ. In that case, the unknown hyperparameters can be assumed to be random and modelled probabilistically. This modelling of the hyperparameters leads to what is called a Bayesian hierarchical model (Berger 1985; Lehmann and Casella 1998). The simplest version of a Bayesian hierarchical model is defined in terms of the ingredients

$$y|\theta \sim p(y|\theta) \quad (\text{outcome model : stage } 1) \tag{3}$$

$$\theta|\gamma \sim \pi(\theta|\gamma) \quad (\text{prior model : stage } 2) \tag{4}$$

$$\gamma|\lambda \sim \psi(\gamma|\lambda) \quad (\text{hyperparameter model : stage } 3), \tag{5}$$

where $\psi(\gamma|\lambda)$ is the prior density of γ. The hyperparameters λ in the stage 3 model are assumed known. In effect, a hierarchical model is a way of modelling the outcomes and the parameters through a sequence of easily interpretable steps.

In practice, it is often helpful to divide θ into natural groups or blocks (θ_1; θ_2, . . ., θ_p), where, for instance, θ_1 consists of the regression coefficients, θ_2 the scale parameters and θ_p the covariance parameters. Each of these separate blocks may then be modelled independently in terms of prior densities $\pi(\theta_j|\gamma)$. In turn, γ may also be grouped into blocks (γ_1,. . ., γ_q) and, in the third stage, modelled independently through the densities $\psi(\gamma_j|\lambda)$. The resulting three-stage hierarchical model then has the form

$$y|\theta \sim p(y|\theta) \quad (\text{outcome model : stage } 1) \tag{6}$$

$$\theta|\gamma \sim \prod_{j=1}^{p} \pi(\theta_j|\gamma) \quad (\text{prior model : stage } 2) \tag{7}$$

$$\gamma|\lambda \sim \prod_{j=1}^{q} \psi(\gamma_j|\lambda) \text{ (hyperparameter model : stage 3)}. \quad (8)$$

This specification may be considered as the canonical hierarchical Bayes model.

Example 1 (Gaussian linear regression model). *Suppose that $y = (y_1,\ldots,y_n)$ is a vector of observations and θ consists of the two blocks (β, σ^2), where β is a k-vector of regression parameters. Now let*

$$y|\theta \sim N_n(y|X\beta, \sigma^2 I_n)$$
$$\theta|\gamma \sim N_k(\beta|\beta_0, B_0) IG\left(\sigma^2 | \frac{v_0}{2}, \frac{\delta_0}{2}\right),$$

where

$$N_k(\beta|\beta_0, B_0) = (2\pi)^{-k/2} \exp\left\{-\frac{1}{2}(\beta - \beta_0) B_0^{-1}(\beta - \beta_0)\right\}$$

is the k-variate normal density, X is the $n \times k$ matrix of covariates and

$$IG\left(\sigma^2 | \frac{v_0}{2}, \frac{\delta_0}{2}\right) = \frac{(\delta_0/2)^{(v_0/2)}}{\Gamma(v_0/2)} \left(\frac{1}{\sigma^2}\right)^{(v_0/2)+1} \exp\left(-\frac{\delta_0}{2\sigma^2}\right), \sigma^2 > 0$$

is the inverse-gamma density. In this case, the hyperparameters γ consist of the four blocks of parameters $(\beta_0, B_0, v_0, \delta_0)$. The top level of the model is the model of the outcome and the bottom level the model of θ. If it is not possible to fix the value of β_0, for example, one may specify a prior, $\beta_0|\lambda \sim \mathrm{N_k}(\beta_0|\beta_{00}, B_{00})$, where the hyperparameters of the third stage $\lambda = (\beta_{00}, B_{00})$ are pre-specified. Further discussion along these lines is provided by Lindley and Smith (1972).

Since the difficulty of specifying hyperparameters in the second stage model of the model arises in almost all applications, hierarchical Bayes modelling is of special interest and importance in Bayesian analysis. To further fix the ideas, the following example, which we develop further below, is instructive and should be studied carefully.

Example 2 *(Gaussian clustered data model). Clustered data arise when n observations are available for each subject i ($i \leq n$) in the sample. For example, in the panel or longitudinal set-up, there are observations across time for each subject. Let the observations on the ith subject be denoted by $y_i = (y_{i1}, \ldots, y_{in_i})$. Assume that the observations are continuous. Binary or ordinal responses can be dealt with in much the same way by adopting the framework of Albert and Chib (1993). The data for all n subjects are collected in the vector $y = (y_1,\ldots,y_n)$. It is common in this context to allow for unique cluster-specific effects. Let $W_i = (w_{i1}, \ldots, w_{in_i})'$ be a $n_t \times q$ matrix of observations on q covariates w_{ij} whose effect on y is assumed to be cluster-specific. Also suppose that X_{1i} is an additional $n_i \times k_1$ matrix of observations on k_1 covariates whose effect on y is assumed to be non-cluster-specific (fixed effect). Then under the assumption that the observations across clusters are independent, a model for the outcomes is*

$$y|\theta \sim \prod_{i=1}^{n} N_{n_i}(y_i | X_{i1}\beta_1 + W_i \beta_{2i}, \sigma^2 I_{n_i}),$$

where the β_{2i} are the cluster-specific effects. If the numbers of clusters is large, as is usual in practice, it is useful to assume that the effects β_{2i} have some structure. One possibility is to assume that the β_{2i} are drawn from a common distribution

$$\beta_{2i}|\gamma \sim N_q(\beta_2, D)$$

independently across i. This is called the exchangeability assumption since the joint distribution of the β_{2i} is invariant to permutation of the indices. Another possibility is the assumption that the β_{2i} are determined by a set of r cluster-specific covariates a_t:

$$\beta_{2i}|\gamma \sim N_q(A_i\beta_2, D)$$

where

$$A_i = \begin{pmatrix} a_i' & 0' & \dots & \dots & 0' \\ 0' & a_i' & \dots & \dots & 0' \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0' & 0' & \dots & \dots & a_i' \end{pmatrix},$$

$\beta_2 = \beta_{21}, \beta_{22}, \dots, \beta_{2q})$ is a $k_2 = r \times q$-dimensional vector and D, as in the first example, is a $q \times q$ matrix. Writing the second stage model in equivalent form as $\beta_{2i} = A_i\beta_2 + b_i$, where $b_i|D \sim N_q$*(0, D), and substituting this into the outcome model, it follows that the outcome model can be expressed as*

$$y|\theta \sim \prod_{i=1}^{n} N_{n_i}\left(y_i|X_i\beta + W_ib_i, \sigma^2 I_{n_i}\right),$$

where $\theta = (\beta, \sigma^2, b_1, \dots, b_n), X_i = (X_{1i} : W_iA_i)$ is a $n_i \times k$ matrix ($k = k_1 + k_2$) and $\beta = (\beta_1, \beta_2)$. The second stage of the model could now be specified as

$$\theta|\gamma \sim N_k(\beta|\beta_0, B_0) IG\left(\sigma^2|v_0/2, \delta_0/2\right) \prod_{i=1}^{n} N_q(b_i|0, D).$$

Next suppose that there is enough prior information to fix (β_0, B_0, v_0, δ_0), but that D (equivalently D^{-1}) cannot be fixed directly. Then $\gamma = D^{-1}$. A convenient assumption is

$$\gamma|\lambda \sim \text{Wishart}_q\left(D^{-1}|\rho_0, R_0\right),$$

where

$$\text{Wishart}_q\left(D^{-1}|\rho_0, R_0\right) = c\frac{|D^{-1}|^{(\rho_0-q-1)/2}}{|R_0|^{\rho_0/2}} \exp\left\{-\frac{1}{2}\text{trace } \left(R^{-1}D^{-1}\right)\right\}, |D^{-1}| \; > 0,$$

is the q-variate Wishart density,

$$c = \left(2^{\rho_0 q/2}\pi^{q(q-1)/4}\prod_{i=1}^{T}\Gamma\left(\frac{\rho_0 + 1 - i}{2}\right)\right)^{-1}$$

is its normalizing constant, and the stage 3 hyperparameters $\lambda = (\rho_0, R_0)$ are known. Under these assumptions the full model is given by

$$y|\theta \sim \prod_{i=1}^{n} N_{n_i}\left(y_i|X_i\beta + W_ib_i, \sigma^2 I_{n_i}\right) \tag{9}$$

$$\theta|\gamma \sim N_k(\beta|\beta_0, B_0) IG\left(\sigma^2|v_0/2, \delta_0/2\right) \prod_{i=1}^{n} N_q(b_i|0, D) \tag{10}$$

$$\gamma|\lambda \sim \text{Wishart}_q\left(D^{-1}|\rho_0, R_0\right). \tag{11}$$

Putting a prior distribution on the hyperparameters γ in this way has several advantages. For one, it produces a prior distribution on θ that is less dogmatic than a prior based on specified hyperparameters since the resulting prior distribution of θ is averaged over the possible values of γ as dictated by the density $\psi(\gamma|\lambda)$:

$$\pi(\theta|\lambda) = \int \pi(\theta|\gamma)\psi(\gamma|\lambda)\mathrm{d}\gamma.$$

If the hyperparameter γ is a scalar discrete quantity with support on the set $\{\gamma_1, \dots, \gamma_G\}$, where G is potentially infinite, then the mixing density $\psi(\gamma \mid \lambda)$ is a probability mass function of the type $\sum_{j=1}^{G} p_j\delta_{\gamma_j}$, where δ_{γ_j} is the indicator function of γ_j, $0 \leq p_j \leq 1$ and $\sum_{j=1}^{G} p_j = 1$. The resulting conditional density $\pi(\theta|\lambda)$ is then a mixture of densities of the form

$$\pi(\theta|\lambda) = \sum_{j=1}^{G} p_j\pi\left(\theta|\gamma_j\right).$$

In this context, $\pi(\theta|\gamma_j)$ are called the component densities and p_j are the component weights. Such mixtures of component densities provide a simple mechanism for modelling θ in a flexible way.

Of course, one could have started at the outset with the prior $\pi(\theta| \lambda)$ by combining stages 2 and 3, leading to the collapsed model

$$y|\theta \sim p(y|\theta) \tag{12}$$

$$\theta|\lambda \sim \int \pi(\theta|\gamma)\psi(\gamma|\lambda)\ \mathrm{d}\gamma, \tag{13}$$

which has the same structure as the standard 2-stage Bayesian model. This is not done, however, because the density of $\theta|\ \lambda$, even if tractable, is generally less easy to manage.

Example 3 *(Gaussian linear regression model and Student-t prior). Suppose that $y = (y_1, \ldots, y_n)$ is a vector of observations and $\theta = (\beta, \sigma^2)$, where β is a scalar regression parameter. Assume that*

$$\begin{aligned}&y|\theta \sim N_n(y|X\beta, \sigma^2 I_n)\\ &\theta|\gamma \sim N(\beta|\beta_0, B_0)IG\left(\sigma^2|\frac{v_0}{2}, \frac{\delta_0}{2}\right)\\ &B_0^{-1} \sim G\left(B_0^{-1}|\frac{v}{2}, \frac{v}{2}\right)\end{aligned}$$

where $G(\cdot\ |\ \cdot, \cdot)$ is the gamma density and the quantities (β_0, v_0, δ_0) and v are known. Then the density of β marginalized over B_0^{-1} is Student-t, $T(\beta|\beta_0, 1, v)$, with location β_0, dispersion 1 and v degrees of freedom. This Student-t prior density is not conjugate with the outcome model and therefore cumbersome to deal with.

Bayesian hierarchical models can have additional stages. For instance, a further stage can be added by placing a prior density on λ, which leads to the model

$$y|\theta \sim p(y|\theta)\ \ (\text{outcome model : stage 1}) \tag{14}$$

$$\theta|\gamma \sim \prod_{j=1}^{p} \pi(\theta_j|\gamma)\ (\text{prior model : stage 2}) \tag{15}$$

$$\gamma|\lambda \sim \prod_{j=1}^{q} \psi(\gamma_j|\lambda)\ (\text{hyperparameter model : stage 3}). \tag{16}$$

$$\lambda \sim \delta(\lambda)\ \ (\text{hyperparameter model 2 : stage 4}), \tag{17}$$

where δ is the density of λ. Models with more than four stages are rare.

Posterior Distributions

In a Bayesian analysis one is interested in deriving and summarizing the posterior distribution of θ given y. One obvious question concerns the form of this posterior distribution. Another question concerns the posterior distribution of the hyperparameters γ. Consider the canonical three-stage hierarchical model in (6)–(8). By Bayes's theorem,

$$\pi(\theta|y) = \frac{p(y|\theta)\pi(\theta)}{m(y)},$$

where $n(\theta) = \int \pi(\theta|\gamma)\pi(\gamma|\lambda)\ \mathrm{d}\gamma$ and $m(y) = \int p(y|\theta)\pi(\theta)\ \mathrm{d}\theta$, called the marginal likelihood, is the normalizing constant. Similarly, the posterior distribution of γ is

$$\pi(\gamma|y) = \frac{p(y|\gamma)\pi(\gamma|\lambda)}{m(y)},$$

where $p(y|\gamma) = \int p(y|\theta)\pi(\theta|\lambda)\ \mathrm{d}\theta$. Before we discuss the tractability of these distributions we state a general result about how much information the data y supply about θ and γ beyond what is introduced by the prior densities $\pi(\theta)$ and $\pi(\gamma|\lambda)$. To measure this information we can use the Kullback–Leibler (KL) divergence measure, which, for any two densities f and g, is defined as

$$K(f, g) = E^f \log\frac{f}{g},$$

where E^f is the expectation with respect to the density f. The following result was proved by Goel and Degroot (1981). The result and proof can also be found in Lehmann and Casella (1998).

Theorem 1 *For the three-stage hierarchical model,*

$$K[\pi(\gamma|y), \pi(\gamma)] < K[\pi(\theta|y),\ \ \pi(\theta)].$$

This result states that the KL divergence between $\pi(\theta|y)$ and $\pi(\theta)$ is greater than between

$\pi(\gamma|y)$ and $\pi(\gamma)$. In other words, the data supply more information about θ than they do about γ. Equivalently, the prior and the posterior of γ are closer than the prior and the posterior of θ. This implies that less learning is possible about the hyperparameters γ than about the parameters θ.

Much less can be said about the form of the posterior densities. In general, the posterior densities $\pi(\theta|y)$ and $\pi(\gamma|y)$ are not tractable. But if we consider the density of θj given (y,γ) and $\theta_{-j} = (\theta_1,\ldots, \theta_{j-1}, \theta_{j+1},\ldots,\theta_p)$, we have

$$\pi\left(\theta_j|y,\gamma,\theta_{-j}\right) \propto p(y|\theta)\pi\left(\theta_j|\gamma\right),$$

which is in closed form provided the prior density $\pi(\theta_j|\gamma)$ is conjugate with $p(y|\theta)$. The density $\pi(\theta_j|y, \gamma, \theta_{-j})$ is called the full conditional density of θ_j. Of course, the marginal density,

$$\pi\left(\theta_j|y\right) = \int \pi\left(\theta_j|y,\gamma,\theta_{-j}|y\right)\ \mathrm{d}\gamma\ \ \mathrm{d}\theta_{-j},$$

where the mixing distribution is the marginal posterior distribution of (γ, θ_{-j}), is almost never available in closed form.

The same sort of difficulty arises in finding $\pi(\gamma|y)$. The problem is that the prior $\pi(\gamma|\lambda)$ generally does not combine with $p(y|\gamma)$ to produce a recognizable density. Nonetheless, just as in the case of θj, the calculations are easier if one considers the full conditional density of γ_j. To see this, note that

$$\begin{aligned}\pi\left(\gamma_j|y,\theta,\gamma_{-j}\right) &\propto p(y|\theta)\pi(\theta|\gamma)\pi\left(\theta_j|\lambda\right)\\ &\propto \pi(\theta|\gamma)\pi\left(\theta_j|\lambda\right),\end{aligned}$$

where the second line follows from the fact that the outcome model in stage 1 is free of γ. Thus, provided $\pi(\theta_j|\lambda)$ is conjugate with $\pi(\theta|\gamma)$, the full conditional density of γ_j can be derived in closed form.

Example 4 *Consider again the clustered data model given in (9)–(11). The full conditional density of b_i is obtained as*

$$\begin{aligned}&\pi(b_i|y,\theta_{-b_i},\gamma) \propto p(y|\theta)\pi(b_i|\gamma)\\ &\quad\propto N_{n_i}\left(y_i|X_i\beta + W_ib_i, \sigma^2 I_{n_i}\right)N_q(b_i|0,D),\end{aligned}$$

which, by standard Bayesian manipulations, is seen to be a $N_q\left(b_i|\widehat{b}_i, B_i\right)$ density, where

$$\widehat{b}_i = B_i(\sigma^{-2}W_i'(y_i - X_i\beta) \quad \text{and}$$
$$B_i = \left(D^{-1} + \sigma^{-2}W_i'\,W_i\right)^{-1}.$$

Turning now to the full conditional density of D^{-1}, we obtain

$$\begin{aligned}&\pi\left(D^{-1}|y,\theta\right) = \pi\left(D^{-1}|\{b_i\}\right) \propto \pi\left(\{b\}|D^{-1}\right)\\ &\pi\left(D^{-1}|\lambda\right) \propto \prod_{i=1}^{n} Nq(b_i|0,D)\ \ \mathrm{Wishart}_q\left(D^{-1}|\nu_0,R_0\right)\\ &= \mathrm{Wishart}_q\left(D^{-1}|\rho_0 + n, \left(R_0^{-1} + \sum_{i=1}^{n} b_ib_i'\right)^{-1}\right),\end{aligned}$$

where in the first line we have used the fact that the full conditional density of D^{-1} depends neither on y nor on β; in the second line, Bayes's theorem; in the third line, substitutions for the needed densities; and in the fourth line, by observation that the product of the normal and Wishart prior densities is an updated Wishart distribution with the stated parameters.

Computational Issues

Difficulties in the computation of the marginal posterior densities of θ_j and γ_j were previously an impediment to the development and application of hierarchical Bayesian models. These difficulties have largely been resolved through the use of Markov chain Monte Carlo (MCMC) methods. These methods typically proceed by simulating the full conditional distributions, $\pi(\theta_j|y, \gamma, \theta_{-j})$ and $\pi(\gamma_j|y, \theta, \gamma_{-j})$. Under general conditions, the recursive simulation of these distributions produces a Markov chain whose limiting invariant distribution is the posterior density of interest, $\pi(\theta; \gamma\, |y)$.

Although it is not possible in this discussion to provide the theory behind MCMC methods, as outlined in Tierney (1994), and Chib and Greenberg (1995), or the range of hierarchical

Bayes models that have been thus processed, it is useful to illustrate the computations with the help of the simplest MCMC method, the so-called Gibbs sampling algorithm. This algorithm was introduced by Geman and Geman (1984) in the context of image processing, but the papers of Tanner and Wong (1987) and Gelfand and Smith (1990) brought it into the limelight.

Suppose that the various blocks $\{\theta_j\}$ and $\{\gamma_j\}$ are chosen to ensure that the associated set of full conditional densities $\{\pi(\theta_j|y,\ \theta_{-j},\ \gamma)\}$ and $\{\pi(\gamma_j|y,\ \theta,\ \gamma_{-j})\}$ are all tractable. Then one cycle of the Gibbs sampling algorithm is completed by simulating $\{\theta_j\}$ and $\{\gamma_j\}$ from each full conditional distribution, recursively updating the conditioning variables while moving through the set of distributions. The Gibbs sampler in which each block is revised in fixed order is defined as follows.

Algorithm: Gibbs Sampling

1. Specify an initial value $\theta^{(0)} = \left(\theta_1^{(0)}, \ldots, \theta_p^{(0)}\right)$ and $\gamma^{(0)} = \left(\gamma_1^{(0)}, \ldots, \gamma_q^{(0)}\right)$.
2. Repeat for $j = 1,2, \ldots, n_0 + M$:

   Generate $\theta_1^{(j)}$ from $\pi\left(\theta_1|y, \theta_2^{(j-1)}, \ldots, \theta_p^{(j-1)}, \gamma^{(j-1)}\right)$

   Generate $\theta_2^{(j)}$from $\pi\left(\theta_2|y, \theta_1^{(j)}, \theta_3^{(j-1)} \ldots, \theta_p^{(j-1)}, \gamma^{(j-1)}\right)$

   ⋮

   Generate $\theta_p^{(j)}$ from $\pi\left(\theta_p|y, \theta_1^{(j)}, \theta_2^{(j)} \ldots, \theta_{p-1}^{(j)}, \gamma^{(j-1)}\right)$

   Generate $\gamma_1^{(j)}$ from $\pi\left(\gamma_1|y, \theta^{(j)}, \gamma_2^{(j-1)}, \ldots, \gamma_q^{(j-1)}\right)$

   Generate $\gamma_2^{(j)}$ from $\pi\left(\gamma_2|y, \theta^{(j)}, \gamma_1^{(j)}, \gamma_3^{(j-1)} \ldots, \gamma_q^{(j-1)}\right)$

   ⋮

   Generate $\gamma_q^{(j)}$ from $\pi\left(\gamma_q|y, \theta^{(j)}, \gamma_1^{(j)}, \ldots, \gamma_{q-1}^{(j)}\right)$.
3. Return the values $\left\{\theta^{(n_0+1)}, \gamma^{(n_0+1)}, \ \theta^{(n_0+2)}, \gamma^{(n_0+2)}, \ldots, \theta^{(n_0+M)}, \gamma^{(n_0+M)}\right\}$.

Thus, in this algorithm, block θ_j is generated from the full conditional distribution

$$\pi\left(\theta_j|y, \theta_1^{(j)}, \ldots, \theta_{j-1}^{(j)}, \theta_{j+1}^{(j-1)}, \ldots, \theta_p^{(j-1)}, \gamma^{(j-1)}\right),$$

where the conditioning elements for the jth block reflect the fact that the previous $(j - 1)$ blocks of θ have already been updated, but the rest have not been. Note that the output from the first n_0 cycles (the burn-in phase) is ignored to allow the effect of the initial values to wear off. One additional point about MCMC methods is that the blocks must be carefully chosen. Sampling over unnecessary blocks can worsen the quality of the output produced by the algorithm, where quality is measured by how quickly the serial correlations of the sampled draws decline to zero. Chains whose serial correlations decline quickly are preferred because they are closer to the ideal of independent sampling.

Example 5 *Consider again the hierarchical Bayesian model for clustered data given in (9)–(11). The joint distribution of the data and the unknowns is given by*

$$\begin{aligned} p\left(y, \theta, D^{-1}\right) &= \pi\left(\beta, \sigma^2, \{b_i\}, D^{-1}\right) p(y|\theta) \\ &= \pi(\beta)\pi\left(\sigma^2\right)\pi\left(D^{-1}\right)\sum_{i=1}^{n} p(y_i|\theta)\pi(b_i|D). \end{aligned} \tag{18}$$

Wakefield et al. (1994) propose a Gibbs MCMC approach for joint distribution that is based on full blocking (that is, sampling each block of parameters from its full conditional distribution). Chib and Carlin (1999) suggest a number of reduced blocking schemes. One of the simplest proceeds by first sampling β marginalized over $\{b_i\}$ and then sampling $\{b_i\}$ conditioned on β. This reduced blocking is possible because b_i in (18) can be marginalized out leaving a normal distribution that can be combined with the assumed normal prior on β. In particular,

$$p\left(y_i|\beta,\sigma^2,D\right) = \int p(y_i|\theta)\pi(b_i|D)\mathrm{d}b_i \propto |V_i|^{-1/2} \exp\left\{(-1/2)(y_i - X_i\beta)'V_i^{-1}(y_i - X_i\beta)\right\},$$

where $V_i = \sigma^2 I_{n_i} + W_i D W_i'$. The reduced conditional posterior of β is therefore

$$\pi(\beta|y,\sigma^2,D) \propto \pi(\beta)\prod_{i=1}^{n}|V_i|^{-1/2} \exp\left\{-\frac{1}{2}(y_i - X_i\beta)'V_i^{-1}(y_i - X_i\beta)\right\} \propto \exp\left\{-\frac{1}{2}\left(\beta-\widehat{\beta}\right)'B^{-1}\left(\beta-\widehat{\beta}\right)\right\},$$

where

$$\widehat{\beta} = B\left(B_0^{-1}\beta_0 + \sum_{i=1}^{n} X_i'\ V_i^{-1}y_i\right) \quad \text{and}$$

$$B = \left(B_0^{-1} + \sum_{i=1}^{n} X_i'\ V_i^{-1}X_i\right)^{-1}.$$

The rest of the MCMC algorithm follows the steps of Wakefield et al. (1994). *In full, we sequentially sample the following distributions many times:*

$$\beta \sim N_k\left(\widehat{\beta},B\right)$$

$$b_i \sim N_q\left(D_i(\sigma^{-2}W_i'\ (y_i - X_i\beta),\right.$$

$$\left.D_i = \left(D^{-1} + \sigma^{-2}W_i'\ W_i\right)^{-1},\right),\quad i \le n$$

$$\sigma^2 \sim IG\left(\frac{v_0 + \Sigma n_i}{2}, \frac{\delta_0 + \Sigma_{i=1}^{n}||y_i - X_i\beta - W_i b_i||^2}{2}\right)$$

$$D^{-1} \sim \mathrm{Wishart}_q\left\{\rho_0 + n, \left(R_0^{-1} + \sum_{i=1}^{n} b_i b_i'\right)^{-1}\right\},$$

where the second and fourth of these distributions were derived in Example 4.

Model Choice

Another inferential concern in practice is the comparison of several hierarchical Bayesian models in order to judge the extent to which the various models are supported by the data. In the context of a hierarchical model for clustered data, for instance, one may be interested in determining the support for an additional cluster-specific effect or of an additional fixed effect. Questions of this type can be answered via *Bayes factors,* or ratios of *marginal likelihoods.* The marginal likelihood of a particular model $\mathscr{M}$ is the normalizing constant of the posterior density,

$$m(y|\mathscr{M}) = \int p(y|\mathscr{M},\theta)\pi(\theta|\mathscr{M},\gamma)\pi(\gamma|\mathscr{M},\lambda)\mathrm{d}\theta\ \mathrm{d}\gamma, \tag{19}$$

the integral of the first stage outcome density function with respect to the prior density of θ and the prior density of the hyperparameters γ. If there are two models $\mathscr{M}_k$ and $\mathscr{M}_l$, the Bayes factor is the ratio

$$\mathrm{B}_{kl} = \frac{m(y|\mathscr{M}_k)}{m(y|\mathscr{M}_l)}. \tag{20}$$

Because MCMC methods deliver draws from the posterior density and the marginal likelihood is the integral with respect to the prior$\pi(\theta|\mathscr{M},\gamma)$ $\pi(\gamma|\mathscr{M},\lambda)$, MCMC output cannot be used directly to average $p(y|\mathscr{M},\theta)$. Nonetheless, computation is feasible by the method of Chib (1995), a widely used method that we now briefly describe.Chib (1995) begins by noting that $m(y|\mathscr{M},\lambda)$ can be expressed as

$$m(y|\mathscr{M}) = \frac{p(y|\mathscr{M},\theta^*)\pi(\theta^*|\mathscr{M},\gamma^*)\pi(\gamma^*|\mathscr{M},\lambda)}{\pi(\theta^*,\gamma^*|\mathscr{M},y)}, \tag{21}$$

for a given (θ^*, γ^*), usually taken to be a high density point such as the posterior mean. Thus, if we have an estimate $\widehat{\pi}(\theta^*, y^*|\mathscr{M}, y)$ of the posterior ordinate, the marginal likelihood on the log scale can be estimated as

$$\log m(y|\mathscr{M}) = \log p(y|\mathscr{M},\theta^*) + \log \pi(\theta^*|\mathscr{M},\gamma^*) + \log \pi(\gamma^*|\mathscr{M},\lambda) - \log \widehat{\pi}(\theta^*,\ \gamma^*|\mathscr{M},y). \tag{22}$$

It turns out that it is possible to get an efficient estimate of the posterior ordinate. The basic idea is to write the posterior ordinate as

$$\pi(\theta^*,\gamma^*|\mathscr{M},y) = \pi\left(\theta_1^*|\mathscr{M},y\right) \times \cdots \times \pi\left(\theta_p^*|\mathscr{M},y,\theta_1^*,\ \ldots,\theta_{p-1}^*\right) \times \pi\left(\gamma_1^*|\mathscr{M},y,\theta^*\right) \times \cdots \times\pi\left(\gamma_q^*|\mathscr{M},y,\theta^*,\gamma_1^*,\ \ldots,\gamma_{p-1}^*\right) \tag{23}$$

and then to estimate each of these ordinates from the output of appropriate MCMC runs. To see what is involved, consider the ordinate $\pi\left(\theta_j^*|\mathscr{M},y,\theta_1^*,\ \ldots,\theta_{j-1}^*\right)$ that appears in this decomposition. By definition,

$$\pi\left(\theta_j^*|\mathscr{M},y,\theta_1^*,\ \ldots,\theta_{j-1}^*\right) = \int \pi\left(\theta_j^*|y,\theta_1^*,\ \ldots,\theta_{j-1}^*,\theta_{j+1},\ \ldots,\theta_p,\gamma\right)\ \mathrm{d}\pi\left(\theta_{j+1},\ \ldots,\theta_p,\gamma|y,\theta_1^*,\ \ldots,\theta_{p-1}^*\right)$$

is the full conditional density integrated with respect to the distribution $\pi\left(\theta_{j+1},\ \ldots,\theta_p,\gamma|y,\theta_1^*,\ \ldots,\theta_{p-1}^*\right)$. To calculate this integral by Monte Carlo one can run an MCMC algorithm in which the blocks $\left(\theta_1,\ldots,\theta_{p-1}\right)$ are fixed at their starred values and sampling is over the remaining free blocks, namely $\left(\theta_j,\theta_{j+1},\ldots,\theta_p,\gamma\right)$. This is called a *reduced* MCMC run. Let the sampled draws from this reduced run be denoted by $\left(\theta_{j+1}^{(r)},\ldots,\theta_p^{(r)},\gamma^{(r)}\right), r=1,\ \ldots,M$. Then, provided the full conditional of θ_j is in closed form, we have the estimate

$$\widehat{\pi}\left(\theta_j^*|\mathscr{M},y,\theta_{j-1}^*\right) = M^{-1}\sum_{r=1}^{M}\pi\left(\theta_j^*|y,\theta_1^*,\ldots,\theta_{j-1}^*,\theta_{j+1}^{(r)},\ldots,\theta_p^{(r)},\gamma^{(r)}\right).$$

Each ordinate is estimated in this way from the output of the appropriate reduced runs. Notice that as more blocks are fixed, fewer distributions appear in the reduced runs.

Example 6 *Consider again the hierarchical Bayesian model for clustered data. In this case, we can decompose* $\pi(\theta^*,\gamma^*|\mathscr{M},y)$ *as*

$$\pi\left(D^{-1*},\sigma^{2*},\beta^*|y\right) = \pi\left(D^{-1*}|y\right)\pi\left(\sigma^{2*}|y,D^*\right)\pi\left(\beta^*|y,D^*,\sigma^{2*}\right),$$

so that all computations are marginalized over $\{b_i\}$. *The first term can be estimated by averaging the Wishart density given in Example 5 over draws on* $\{b_i\}$ *from the full MCMC run. To estimate the second ordinate, which is conditioned on* D^*, *we run a reduced MCMC simulation with the full conditional densities*

$$\pi\left(\beta|y,D^*,\sigma^2\right),\pi\left(\sigma^2|y,\beta,D^*,\{b_i\}\right),\ \pi\left(\{b_i\}|y,\beta,D^*,\sigma^2\right),$$

where each conditional utilizes the fixed value of D. The second ordinate is now estimated by averaging the inverse gamma full conditional density of σ^2 *at* $\sigma^2{*}$ *over the draws on* $(\beta\ \{b_i\})$ *from this reduced run. The third ordinate is multivariate normal as given in Example 5 and available directly.*

If the full conditional densities are not in closed form, the marginal likelihood can be computed by the modified Chib method as discussed in Chib and Jeliazkov (2001).

See Also

- ▶ Bayesian Econometrics
- ▶ Bayesian Statistics
- ▶ Econometrics
- ▶ Fixed Effects and Random Effects
- ▶ Longitudinal Data Analysis
- ▶ Markov Chain Monte Carlo Methods
- ▶ Model Selection
- ▶ Simulation-Based Estimation
- ▶ Statistical Inference

Bibliography

Albert, J.H., and S. Chib. 1993. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* 88: 669–679.

Berger, J. 1985. *Statistical decision theory and Bayesian analysis*. New York: Springer.

Chib, S. 1995. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association* 90: 1313–1321.

Chib, S., and B.P. Carlin. 1999. On MCMC sampling in hierarchical longitudinal models. *Statistics and Computing* 9: 17–26.

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

Chib, S., and I. Jeliazkov. 2001. Marginal likelihood from the metropolis-Hastings output. *Journal of the American Statistical Association* 96: 270–281.

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

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

Goel, P.K., and M.H. Degroot. 1981. Information about hyperparameters in hierarchical models. *Journal of the American Statistical Association* 76: 140–147.

Lehmann, E., and G. Casella. 1998. *Theory of point estimation*. New York: Springer.

Lindley, D.V., and A.F.M. Smith. 1972. Bayes estimates for the linear model. *Journal of the Royal Statistical Society B* 34: 1–41.

Tanner, M.A., and W.H. Wong. 1987. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* 82: 528–550.

Tierney, L. 1994. Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics* 21: 1701–1762.

Wakefield, J.C., A.F.M. Smith, A. Racine-Poon, and A.E. Gelfand. 1994. Bayesian analysis of linear and nonlinear population models using the Gibbs sampler. *Applied Statistics* 43: 201–221.

Hierarchy

Luis Garicano

Abstract

Hierarchies lighten the burden of the enormous informational requirements of the price system under uncertainty by acquiring more knowledge and information than any individual can. They are thus are useful for information handling. They can also allow agents to engage in collective actions by decreasing the risk of opportunistic behaviour. But trade-offs are involved because hierarchies impose costs, including communication among agents. This article reviews the literature on this trade-off and its implications for labour markets.

Keywords

Bounded rationality; Earnings distribution; Firm size; Hierarchy; Incentive conflict; Information economics; Monitoring theories; Principal and agent; Queuing theory; Uncertainty

JEL Classifications

L22

Hierarchy deals with individuals' bounded rationality by allowing for more information to be used in decision-making than individual agents could possibly use and by allowing the most skilled agents to leverage their knowledge with the help of others. Hierarchy can also allow agents to engage in collective actions by decreasing the risk of opportunistic behaviour. These benefits of hierarchy do not come without costs. Hierarchies may be slow to react, may introduce noise into the communication process, and generally may require costly communication among agents. This entry reviews the literature on these trade-offs in multiple-layer, multi-agent hierarchies; that is, it leaves aside the simplest, one principal one agent-type model.

Processing Information

Arrow (1974) first observed that the enormous informational requirements of the price system under uncertainty (complete markets require one state-contingent price per commodity per state of the world) place a bound on its performance. A key role of hierarchy is to lighten this burden: hierarchies can acquire more information than any individual can, and thus are useful for information handling.

A large literature has explored this role of hierarchies in enhancing information processing. An example of this class of models is Radner (1993). Decision-making requires observing a linear combination of certain variables, and agents incur some cost of performing additive operations and some cost of communicating their results. All information must be processed. Under these conditions, organizations are asymmetric, to ensure that all agents are occupied. Bolton and Dewatripont (1994) extend this model to the case where cohorts of data arrive all the time; in this case, the optimal structures are balanced trees, and look more like the ones we arguably observe in reality. Radner and Van Zandt (1992), Van Zandt (1999), and Van Zandt and Radner (2001) take a further step by studying the problem of processing in real time, when the information relevant to a given decision is continuously arriving. The key objects of interest are the sign and size of the scale diseconomies resulting from hierarchy. That is, these authors aim to answer the question of the extent to which diseconomies of scale linked to human bounded rationality are the reason we see many firms, rather than one. The answer is not unambiguous. For example, Radner and Van Zandt (1992) find that returns to scale can vary from increasing to sharply decreasing, depending on the correlation of the data and on the cost of incorrect decisions. Vayanos (2003) extends substantially these models beyond associative operations, and considers situations with two realistic characteristics: the decisions of different agents interact; and the aggregation process entails information losses.

A separate branch of the literature, following Crémer (1980), has studied hierarchical resource allocation programmes under limited managerial processing power. Geanakoplos and Milgrom (1991) study a hierarchy in which managers can invest in information collection, but each manager can collect a limited amount of information. By decomposing hierarchically the allocation problem (so that a low-level manager allocates resources among shops, while lower-level managers allocate resources among groups of sources, and so on) the total amount of information used can be increased. Each manager is told by his superior how many resources he gets, and communicates that information to each subordinate. Managers aim to minimize the expected total costs of their units (there are no externalities). Under these conditions, the number of managers used is increasing in value of information and U-shaped in managerial ability (few managers are used if they are unproductive or if they are so productive that a few can achieve all savings). A more uncertain environment increases the number of managers needed and their average skill, and causes a decrease in their span of control.

Organizing Knowledge

In Garicano (2000) a hierarchy, rather than a means to aggregate information, is a means to acquire and conserve experts' knowledge. He considers a set of agents who face a large number of problems. They may or may not invest in learning their solution; they produce only if they do. Some problems are more common than others, and there is an *ex ante* known probability distribution of problem. Agents can ask other agents for help in solving their problems, but, crucially, they do not know who knows what. Garicano shows that an optimal organization has agents specializing in either production or problem solving; that production workers deal with routine problems and problem solvers specialize in the exceptions; and that shape is pyramidal, with fewer agents in each successive layer. The organizing principle is management by exception. The key organizational trade-off is between acquiring knowledge and asking; that is, an extra hierarchical tier increases communication costs but also increases the utilization of expertise and results in lower knowledge acquisition costs. Given this trade-off, an increase in the cost of communication leads agents to learn more and ask less, and managers to learn less and deal with a smaller proportion of problems. Conversely, when the cost of acquiring knowledge rises, the role of the hierarchy increases as managers deal with a larger fraction of problems.

Beggs (2001) also investigates the phenomenon of 'management by exception', although with exogenous knowledge. He uses queuing theory to explore the optimal allocation of workers with exogenously given skills to the different layers of a hierarchy.

This type of organization of work is common in many contexts; for example, in law firms (Garicano and Hubbard 2007) or in medicine. In this professional context, the role of the 'juniors' (associates, residents) is to handle the easier problems to conserve the valuable time of the seniors (attending physicians, partners) for the harder problems. Similarly, in a team engaged in technical support (Orlikowski 1996), experts must answer customer calls, and production is organized so that juniors handle front calls, and transfer the calls they cannot handle to more senior experts.

Hierarchical Allocation of Talent to Positions: The Distribution of Earnings

Another line of research has explored the relation between the distributions of income and the distributions of firm size and hierarchy. This literature has proposed that the reason why the distribution of income is more skewed than the underlying distribution of skills lies in how resources are allocated to individuals. Higher-ability managers raise the productivity of the resources they are assigned more than lower-ability managers. As a result, in equilibrium, more able managers are allocated more resources, and this leads the marginal value of their ability to increase faster than if they were working on their own. Lucas (1978) and Rosen (1982) generate full equilibrium models that yield both an equilibrium firm size and distribution of earnings. In both these papers, the manager increases the productivity under his control, which, depending on the model, may be the number of workers (Lucas 1978), or efficiency units of labour, that is, total units of skill managed (Rosen 1982). In these models, managerial human capital raises the marginal product of the workers or capital they are assigned, but managers' span of control is generally limited implicitly or explicitly by managers' time. Equilibrium assignment patterns involve scale of operations effects, which follow from the complementarity between managerial human capital and productive resources. The main equilibrium result from this class of models is that these production functions involve scale-of-operations effects: more skilled managers are assigned more resources to manage in equilibrium. As a result, the distribution of earnings is more skewed than the distribution of skills. Garicano and Rossi-Hansberg (2005) build on this line of research, but study a model of hierarchy with heterogeneous agents that extends Garicano (2000), and which involves matching between managers and workers – that is, managers do not care only about the efficiency units they manage (which would imply that a top lawyer at a law firm should be indifferent between managing two good associates or a large number of mediocre ones), but instead care about both the quality and quantity of workers. The model generates a continuum of hierarchies and an equilibrium allocation of workers to positions, as well as the income distributions. It allows for the simultaneous exploration of changes in organization and in wage structure, and has been applied to issues such as the formation of cross-country teams (Antràs et al. 2006), or changes in organization and the wage structure as a result of the information technology revolution (Garicano and Rossi-Hansberg 2005).

Monitoring and Authority

An alternative class of theories study managers as agents able to fire underperforming agents or otherwise exercise their authority. Monitoring theories stem from Alchian and Demsetz (1972), who posit that hierarchies are a response to incentive problems associated with team production. In this view, lower-level individuals are directly involved in production, and upper-level individuals are specialized monitors. The view was elaborated formally by Calvo and Weillisz (1978) and Qian (1994). Their basic assumption is that supervision is necessary for ensuring performance. They study an efficiency wage setting like Becker and Stigler's (1974), where agents can work full-time and earn w or shirk and be detected and fired with probability *p,* in which case they earn their reservation utility. Here, the principal can induce work by increasing the monitoring intensity *p* through hierarchical supervision. The hierarchy then trades off the gains due to these lower wages against the cost of the supervisors.

Aghion and Tirole (1997) formally introduce the idea of decision-making agents into the study of hierarchy with incentive conflicts. Delegation by a superior functions as a commitment not to intervene, and as such delegating authority increases incentives for agents to invest. Baker et al. (1999) extend such analysis to a context where delegation is in fact a relational contract in which the centre chooses not to exercise its power. Rajan and Zingales (2001) study how the shape and size of the hierarchy responds to the problem of providing incentives for employees to protect the resources of the entrepreneur and discouraging them from stealing them. Finally, Hart and Moore (2005) consider hierarchies as chains of authority that determine priority in decisions over asset allocation, and derive conditions where optimal hierarchies have generalist coordinators on top. Their theory helps explain why generalists – individuals who know about the interactions between classes of assets – should be senior to specialists.

Overall the models reviewed in this article have the potential to address an important missing link in economic theory: the absence of managers, and of occupations, from both the theory of the firm and the theory of the determination of wages.

See Also

- ▶ Implicit Contracts
- ▶ Incomplete Contracts
- ▶ Information Aggregation and Prices

References

Aghion, P., and J. Tirole. 1997. Formal and real authority in organizations. *Journal of Political Economy* 105: 1–29.

Alchian, A., and H. Demsetz. 1972. Production, information costs, and economic organization. *American Economic Review* 62: 777–795.

Antràs, P., Garicano, L. and Rossi-Hansberg, E. 2006. Offshoring in a knowledge economy. *Quarterly Journal of Economics* 121 (forthcoming).

Arrow, K. 1974. *The limits of organization*. New York: W.W. Norton.

Baker, G., R. Gibbons, and K.J. Murphy. 1999. Informal authority in organizations. *Journal of Law, Economics & Organization* 15: 56–73.

Becker, G., and G. Stigler. 1974. Law enforcement, malfeasance, and compensation of enforcers. *Journal of Legal Studies* 3: 1–18.

Beggs, A. 2001. Queues and hierarchies. *Review of Economic Studies* 68: 297–322.

Bolton, P., and M. Dewatripont. 1994. The firm as a communication network. *Quarterly Journal of Economics* 109: 809–839.

Calvo, G., and S. Wellisz. 1978. Supervision, loss of control, and the optimum size of the firm. *Journal of Political Economy* 86: 943–952.

Crémer, J. 1980. A partial theory of the optimal organization of a bureaucracy. *Bell Journal of Economics* 11: 683–693.

Garicano, L. 2000. Hierarchies and the organization of knowledge in production. *Journal of Political Economy* 108: 874–904.

Garicano, L. and Hubbard, T. 2007. Managerial leverage is limited by the extent of the market: Hierarchies, specialization and the utilization of lawyers' human capital. *Journal of Law and Economics* (forthcoming).

Garicano, L., and E. Rossi-Hansberg. 2005. *Organization and inequality in a knowledge economy. Working Paper No. 11458*. Cambridge, MA: NBER.

Geanakoplos, J., and P. Milgrom. 1991. A theory of hierarchies based on limited managerial attention. *Journal of the Japanese and International Economies* 5: 205–225.

Hart, O., and J. Moore. 2005. On the design of hierarchies: Coordination vs. specialization. *Journal of Political Economy* 113: 675–702.

Lucas, R. 1978. On the size distribution of business firms. *Bell Journal of Economics* 9: 508–523.

Orlikowski, W. 1996. Improvising organizational transformation over time: A situated change perspective. *Information Systems Research* 7: 63–92.

Qian, Y. 1994. Incentives and loss of control in an optimal hierarchy. *Review of Economic Studies* 61: 527–544.

Radner, R. 1993. The organization of decentralized information processes. *Econometrica* 61: 1109–1146.

Radner, R., and T. Van Zandt. 1992. Information processing in teams and returns to scale. *Annales d'Economie et de la Statistique* 25 (26): 265–298.

Rajan, R., and L. Zingales. 2001. The firm as a dedicated hierarchy: A theory of the origins and growth of firms. *Quarterly Journal of Economics* 116: 805–851.

Rosen, S. 1982. Authority, control, and the distribution of earnings. *Bell Journal of Economics* 13: 311–323.

Van Zandt, T. 1999. Real-time decentralized information processing as a model of organizations with boundedly rational agents. *Review of Economic Studies* 66: 633–658.

Van Zandt, T., and R. Radner. 2001. Real-time decentralized information processing and returns to scale. *Economic Theory* 17: 497–544.

Vayanos, D. 2003. The decentralization of information processing in the presence of interactions. *Review of Economic Studies* 70: 667–695.

Higgling

F. Y. Edgeworth

Higgling of the market is described by Adam Smith as a process by which ‘exchangeable value’ is adjusted to its measure ‘quantity of labour’:

> It is often difficult to ascertain the proportion between two different quantities of labour … it is not easy to find any accurate measure either of hardship or ingenuity. In exchanging, indeed, the different productions of different sorts of labour for one another, some allowance is commonly made for both. It is adjusted, however, not by any accurate measure, but by the higgling and bargaining of the market, according to that rough equality which, though not exact, is sufficient for carrying on the business of common life (*Wealth of Nations,* bk. i, ch. v).

Compare Fleeming Jenkin:

> The higgling of the market, ascertaining the result of the relative demand and supply in that market, does not in the long run determine the price of either eggs or tea; it simply finds out the price which had been already determined by quite different means (‘Time-Labour System’, *Papers, Literary, Scientific,* etc., p. 139).

It is possible to accept the writer’s account of the *market process* (ibid. p. 123) without contrasting so strongly the determination of price by demand and supply and by cost of production (cf. Marshall’s *Principles,* Preface to 1st edn, p. xi.). Prof. Marshall at the beginning, when treating of the theory of the equilibrium of demand and supply, gives an excellent type of the action of a market (ibid, 5th edn, bk. v, ch. ii, § 2). The subject can hardly be apprehended without mathematical conceptions. Thus Mill, in his description of the play of demand and supply *(Political Economy,* bk. iii, ch. ii, § 4), in the absence of the idea of a demand-curve or function, may seem to use the phrases ‘demand increases’, ‘demand diminishes’, loosely. A more distinct idea is thus expressed by Fleeming Jenkin in his *Graphic Representations:* ‘If every man were openly to write down beforehand exactly what he would sell or buy at each price, the market price might be computed immediately.’ A similar idea is presented by Prof. Walras *(Éléments d’économie pure,* article 50). In some later passages he has formulated the higgling of the market more elaborately. The present writer, criticizing these passages *(Revue d’économie politique,* January 1891), has maintained that even if the dispositions of all the parties were known beforehand, there could be predicted only the position of equilibrium, not the particular course by which it is reached. Of course special observation may supply the defects of theory. For instance there may be evidence of the incident which Cantillon attributes to the ‘altercation’ of a market, namely the predominant influence of a few buyers or sellers; ‘le prix réglé par quelques uns est ordinairement suivi par les autres’ *(Essai,* part ii, ch. ii. Des prix des marchés). Compare Condillac:

> ‘Aussitôt que quelues uns seront d’accord sur la proportion à suivre dans leurs échanges les autres prendront cette proportion pour règle’ *(Le Commerce et le Gouvernement,* ch. iv: *Des marchés).*

‘Higgling’ is not always qualified as ‘of a market’. The term may be used in much the same sense as the ‘art of bargaining’ is used by Jevons, with reference to a transaction between two individuals, in the absence of competition (*Theory,* p. 124, 3rd edn). Thus Professor Marshall, in an important passage relating to the case in which agents of production are held by two monopolists, says that there is ‘nothing but “higgling and bargaining”’ to settle the proportions in which a certain surplus will be divided between the two (*Principles of Economics,* bk. v, ch. xi). Moses, in the *Vicar of Wakefield,* did not require a fair for the exercise of the skill which is thus attributed to him: ‘He always stands out and higgles and actually tires them till he gets a bargain.’

Bibliography

Cantillon, R. 1755. *Essai sur la nature du commerce en général.* Paris. Ed. H. Higgs, London: Macmillan.

Condillac, E.B.A.M. 1776. Le commerce et le gouvernement considerés relativement l'un à l'autre. In *Oeuvres complètes de Condillac*, vol. 4. Paris: Briére, 1821.

Jenkin, H.C.F. 1887a. Graphic representation of the laws of supply and demand. London. Reprinted in his *Papers, literary, scientific & c.* London: Longmans, Green & Co.

Jenkin, H.C.F. 1887b. Time-labour system. In his *Papers, literary, scientific* & c., ed. S.C. Colvin and J.A. Ewing, London: Longmans, Green & Co.

Marhsall, A. 1890. Preface to his *Principles of Economics.* London: Macmillan. 5th edn., 1907.

Mill, J.S. 1848. *Principles of political economy.* London: J.W. Parker.

Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*. London: W. Strahan & T. Cadell.

Walras, L. 1874–1877. *Eléments d'économie pure.* Lausanne: Corbaz.

Higgs, Henry (1864–1940)

Murray Milgate

The original edition of this *Dictionary* was reprinted (with revisions) by Inglis Palgrave on a number of occasions during his lifetime. The only edition compiled by someone other than Palgrave himself was that which Henry Higgs published between 1923 and 1926. That revised edition, which for the first time incorporated Palgrave's name into the title, important as it was, had to be compiled under the severe restriction of having to use the original plates for the bulk of the text. This permitted Higgs only two avenues for bringing the original up to date. The first was to add, in an appendix to each volume, biographical notices of economists who had died since the compilation of the original and, in a few cases, continuations of articles already to be found in the body of the text. The second was even more inhibiting and involved replacing sections of the text of the original with new material of exactly the same number of words. One senses behind Higgs's bland explanation of this course of action – that to reset the whole would 'have necessitated a prohibitive price for the volume' – more than a note of regret.

Henry Higgs was born on 4 March 1864, the eleventh of thirteen children of a Cornish landowner. At the age of eighteen he entered the civil service as a Lower Division Clerk in the War Office, moving to the Postmaster General's department in 1884. In 1899 he was transferred to the Treasury, and when Sir Henry Campbell-Bannerman took office as Prime Minister in December 1905 Higgs was appointed his Private Secretary. Upon Campbell-Bannerman's death in 1908 he returned to the Treasury. Higgs remained a civil servant until his retirement in 1921. Like his friend James Bonar, Higgs seems to have found a career in the civil service sufficiently flexible to admit of active research into the history of economics and a close involvement with teaching and the professional associations of British economists. In this latter context, particularly to be noted is the instrumental part he played in securing for the British Economic Association (of which he was a founding member in 1890) its Royal Charter in 1902 when it changed its name to the Royal Economic Society. From 1892 until 1905 Higgs served as Secretary of the RES, and from 1896 until 1905 he was the assistant editor (to Edgeworth) of the *Economic Journal*.

In 1884 Higgs began attending lectures on jurisprudence and Roman law at University College, London, finally securing (after having first to matriculate) his LL.B in 1890. His only formal instruction in economics seems to have come from Foxwell, whose lectures at University College Higgs attended in 1885–6 and 1886–7, and it seems likely that it was from this source that his interest in the work of Richard Cantillon derived. (Higgs's book on Physiocracy is dedicated to Foxwell: 'my master and friend'.) Cantillon was probably the perfect subject for Higgs – he had been hailed by Jevons as having established the 'nationality of political economy', but his work and name were scarcely known; he had a mysterious personal history; and the book to which so much credit was being given apparently existed only in a French translation of the missing English original. Research in the Bibliothéque Nationale during his annual vacations provided the material upon which his article on Cantillon for the first volume of the *Economic Journal* in 1891 was based. On the same subject there followed his entry for Palgrave's *Dictionary* in 1894 and, after his retirement, his now standard edition of Cantillon's *Essai* was published under the auspices of the Royal Economic Society (1931).

From these researches undoubtedly sprang Higgs's other great interest in the area: the economics of Physiocracy. In May and June of 1896 Higgs gave a series of six lectures on the Physiocrats at the London School of Economics. These were published in the following year as *The Physiocrats*. In 1894 he had written the entry on the *Economistes* for Palgrave, but that piece ran to only two short paragraphs (one of which was little more than a list of names) and referred the reader to the much longer entry on the Physiocrats which was written by Gustave Schelle. His stamp was thus more permanently impressed on the study of Cantillon than it was upon that of Physiocracy.

Of Higgs's other contributions to the history of economics, only two need to be noted. First, and not surprisingly, Higgs was among the most stalwart of supporters of Palgrave's *Dictionary*, contributing nineteen entries to its original edition (including those on Cantillon, Mirabeau, Turgot, and the *Economistes*), and forty more to his own edition of 1923–6. His entry on 'Débouchés' has been retained in the present work. It is clear that the *Dictionary* is the vehicle which will perpetuate his name. Secondly, in the later years of his life he undertook to edit and produce for the Royal Economic Society bibliographical volumes on the literature of economics. The idea was to capitalize upon, and to record for posterity, the legacy of Foxwell's activities as a scholar and book collector extraordinary. Unfortunately, only one volume appeared (1935) – as Keynes remarked, it may have been that at so late a stage of his life the task no longer suited his gifts (1940, p. 556).

It might also be noted that Higgs's one disservice to the history of economics was his edition of an unfinished manuscript by Jevons which was published in 1905 under the title *Principles of Economics*. So fragmented is this, that it is very difficult to imagine just how Higgs could have been persuaded to print it. Appearing as it did when the climate of opinion about Jevons (largely due to Marshall's efforts) was somewhat less than enthusiastic, it numbers as one of those unfortunate incidents which have combined to diminish the reputation of Jevons in a way that is entirely unwarranted.

In addition to these works, Higgs published two books on the financial system of Britain. The first, *The Financial System of the United Kingdom*, appeared in 1914 was an attempt to provide a connected account of governmental financial procedure. The second, *Financial Reform*, appeared in 1924 and was more an account of the conduct of government policies as Higgs had directly experienced it. He also delivered the Newmarch Lectures at University College in 1892 and 1893 on household budgets, and was the editor of the Centenary Volume of the Political Economy Club.

An exemplary description of Higgs in later life was written by Keynes for the December number of the *Economic Journal* for 1940. So improbable is it that this will ever be bettered, it is reproduced below with the permission of the Society which Higgs helped to found (in the text, Keynes is referring to his recollections of meetings of the council of the Royal Economic Society):

> Becoming, at the last, extremely deaf and quite unable to hear the comments of others present, in which indeed he seemed to take no interest, his argument would continue as an entirely solo performance, frequently on some other item of the agenda than that under discussion; the only Chairman, in my experience, who was able to make him desist until his oration was really finished, being Edwin Cannan, who used to take him almost by the throat, shouting down his ear that we were not discussing that matter, and putting his hand over his mouth until he gave up. Or on other occasions when he had more curiosity as to what was going on, he would push towards whomever was speaking his highly unreliable electrical machine, which would proceed to deliver a thunder-and-lightning storm above which nothing could be heard. I wish I could give some slight indication of Higgs's very individual and oratorical manner of address. It could be a bore and a hindrance if one was in a hurry in this modern age, – or in such circumstances as the above! But if only one could be patient, it had in truth extraordinary finish and a sort of beauty of its own; unquestionably great style in it. These orations to our

> Council, delivered on the wrong items of the agenda, were often delightful in themselves, elaborately prepared beforehand, I sometimes thought, really remarkable in their own way and the best and most characteristic product of his personality (Keynes, 1940, pp. 557–8).

See Also

- ▶ Palgrave's Dictionary of Political Economy

Selected Works

1891. Richard Cantillon. *Economic Journal* 1: 262–91.

1897. *The physiocrats*. London: Macmillan & Co.

1905. W.S. Jevons: *Principles of economics*. Edited with Harriet Jevons. London: Macmillan & Co.

1914. *The financial system of the United Kingdom*. London: Macmillan & Co.

1923–6. (ed.) *Palgrave's dictionary of political economy*. 3 vols. London: Macmillan & Co.

1924. *Financial reform*. London: Macmillan & Co.

1931. (ed.) R. Cantillon, *Essai sur la nature du commerce en général*. London: Macmillan & Co. for the Royal Economic Society.

1935. *A bibliography of economics down to 1700*. Cambridge: Cambridge University Press.

Bibliography

Keynes, J.M. 1940. Obituary: Henry Higgs. *Economic Journal* 50: 555–558.

High-Powered Money and the Monetary Base

Karl Brunner

The concept of high-powered money or a monetary base appears as an important term in any analysis addressing the determinants of a nation's money stock in regimes exhibiting financial intermediation. Two types of money can be distinguished in such institutional contexts. One type only occurs as a 'monetary liability' of financial intermediaries. It characteristically offers a potential claim on another type of money. The contractual situation between customers and intermediaries reveals that this potential claim, to be exercised any time at the option of the owner, forms a crucial condition for the marketability of the intermediaries' monetary liabilities. This second type offers in contrast no such potential claim. While it is exchangeable for other objects, it is a sort of 'ultimate money' without regress to other types of money.

This characterization differs from the widely used classification 'outside-inside' money. 'Inside money' matches in a consolidated balance sheet of 'money producers' a corresponding amount of private debt. Money which cannot be matched in this way forms the outside money. But outside money does not necessarily coincide with the monetary base. The latter magnitude exceeds the volume of outside money by the amount of private debt acquired by the Central Bank in fiat regimes. The two concepts refer, however, to the same magnitude in pure commodity regimes and even in some possible Central Bank regimes with specific arrangements. It follows that the monetary base covers a somewhat wider range than outside money. This difference corresponds to the different analytic purposes of the two concepts. The 'monetary' base is designed for explanations of the behaviour of a nation's money stock, whereas 'outside money' was advanced to express the monetary system's contributions to the economy's net wealth.

The distinction between monetary base and the nation's money stock is hardly informative or relevant for pure commodity money regimes. The distinction becomes important with the emergence of intermediation. Financial intermediation inserts a wedge between the monetary base and the money stock (see article on money supply). But regimes with intermediation cover a wide range of arrangements bearing on the nature of the monetary base. High-powered money may consist of commodity money with or without fiat component or of pure fiat money. These differences are characteristically associated with significant differences in the supply conditions of high-powered money.

The measurement of the monetary base for any country involves, at this stage of monetary evolution, the consolidated balance sheet of the Central Bank system. But the Central Bank is usually not the only producer of 'ultimate money'. The balance sheet of other agencies may also have to be considered. This extension covers in the USA a special Treasury monetary account summarizing the Treasury's money creating activity. In other cases, a balance sheet of the mint or an exchange equalization account may have to be added. But whatever the range of ultimate money producers may be, we need to consolidate their respective balance sheets into a single statement. The monetary 'liabilities' of this consolidated statement, i.e., all items listed on the right-side of the consolidated statement which are money, constitute the monetary base.

The consolidated statement determines that the monetary base can be expressed in two distinct ways. It can be exhibited as the sum of its uses by banks and public. The 'uses statement' thus presents the monetary base as the sum of bank reserves in form of base money and currency held by the public. A 'source statement' complements the uses statement. The sources statement can be immediately read from the balance sheet. The monetary base appears thus as the sum of all assets listed on the left-side of the consolidated statement minus the sum of all non-monetary liabilities. Both statements can be easily derived from the published data in the USA. More difficulties may be encountered for other countries.

The comparatively simple case of the USA may be used to exemplify the sources statement needed for the subsequent discussion. We can write the following expression:

> Monetary Base = Federal Reserve Credit (i.e., earning assets of Central Bank consisting of government securities and advances to banks) + gold stock (including SDR's) minus treasury cash (i.e. free gold) + treasury currency (mostly coin) + a mixture of other assets minus other liabilities (including net worth).

Both uses and sources statement refer to important aspects of the money supply process. The uses statement refers in particular to the allocation of base money, determined by the public's and the bank's behaviour, between bank reserves and currency held by the public. This allocation contributes to shape the link between monetary base and money stock. The sources statement on the other hand directs our attention to an examination of possible (or relevant) supply conditions of base money.

The measurement, but not the definition, of the base clearly depends on prevailing institutions. One particular institution, viz. the imposition of variable reserve requirements on financial intermediaries, suggests a useful extension of the money base. Changes in reserve requirements release or absorb reserves similar to transactions between banks and Central Bank, e.g., an open-market operation. Similar consequences follow with respect to both money stock and 'bank credit'. Thus appeared an extension of the monetary base beyond the 'sources base' (or the volume of high-powered money) defined by the sources statement. The monetary base is understood as the sum of the 'sources base' and a reserve adjustment magnitude (RAM). This magnitude is the cumulated sum of all past releases and absorption of reserves due to changes in reserve requirements. This practice has become the standard procedure in the reports published by the Federal Reserve Bank of St Louis. The extended concept of the base offers the further advantage that the resulting magnitude only reflects actions of the monetary authorities and also reflects all the most important actions proceeding within a given institutional framework.

The sources statement offers a useful starting point for an analysis of the supply conditions of the monetary base. The study of these conditions is motivated by the systematic relation between base and money supply. Changes in the monetary base are a necessary condition for persistently large or substantially accelerated monetary growth in most countries for most of the time. Substantial changes in the monetary base are frequently also a sufficient condition for corresponding changes in the money supply.

The sources statement yields a means to examine the sources of all changes in the base. We can thus investigate which of the sources dominate the trend, the variance of cyclical movements and the variances of middle range or very short-run movements. The patterns shift over time with the monetary regime and vary substantially between countries. Trend and longer-term variance in the USA are dominated, for instance, by the behaviour of the Federal Reserve Credit (i.e., the earning assets of the Central Bank system). We find in contrast for the Swiss case that trend and variance of the base are dominated by the behaviour of the gold stock and foreign exchange holdings. The portfolio of government securities play a comparatively small role. Such examination can also be exploited in order to judge whether movements in the base are essentially temporary or can reasonably be expected to persist with a longer duration.

The stochastic structure of the major and minor source components constitute the supply conditions of the monetary base. These conditions are sensitively associated with a variety of institutional arrangements under the control of legislative bodies or policymakers. The procedures instituted, for instance, by the Federal Reserve system to offer check collection services to banks contribute to the shortest run variance of the monetary base. Reserve requirements imposed on the liabilities of financial institutions offer policy-makers an opportunity to raise the proportion of outstanding government debt held by the Central Bank.

Higher reserve requirements raise the level of the monetary base required to produce a given money supply. Correspondingly a larger volume of government securities can be held by the Central Bank.

The supply conditions may disconnect the behaviour of the base from the economy. This will happen whenever the processes governing the source components operate essentially independently of the economy's movements. In general some dependence may be produced by the prevailing institutions and policies. Such a feedback creates a role for the interaction within asset markets, and also between asset markets and output markets in the determination of the monetary base. The supply conditions of the monetary base acquire thus a central role in our monetary affairs. This is most particularly the case as these conditions emerge from legislative decisions and policy strategies. They fully characterize under the circumstances an important component of a monetary regime. Different monetary regimes are reflected by variations in the supply conditions. The growing dissatisfaction with the discretionary regime, which produced the Great Depression and the inflation of the 1970s, initiated in recent years much public debate about the nature of an adequate monetary regime. A rational examination requires in this case an evaluation of the consequences associated with alternative supply conditions governing the monetary base. This programme still needs some attention by the professions and ultimately (and very hopefully) even by politicians.

See Also

- ▶ Monetary Policy
- ▶ Money Supply

Hildebrand, Bruno (1812–1878)

Hermann Reich

Keywords
Capitalism; Communism; Engels, F.; German Historical School; Hildebrand, B.; List, F.; Müller, A.; Self-interest; Smith, A.; Socialism; Statistics

JEL Classifications
B31

Hildebrand was born in Naumburg (Thuringia), the son of a clerk to the court. He studied in Leipzig and Breslau. In 1841 he was promoted full professor of Staatswissenschaften (of government, which included political economy) at the University of Marburg.

Hildebrand had always been an activist in the liberal and patriotic movement. He faced political persecution before the 1848 revolution, during which he was elected deputy of the Frankfurt National Assembly. In the subsequent period of restoration he was forced to emigrate to Switzerland, where he became not only a professor but also the director of a railway company, and founded the first Swiss statistical office (at Berne). In 1861 he was appointed professor at the University of Jena. He was founder (in 1862) and editor of the *Jahrbücher für Nationalökonomie and Statistik* and contributed to the establishment of the statistical office of the United Thuringian States (in 1864).

Hildebrand is considered as one of the founders of the German Historical School. He was opposed to the deductive method of the classicals and denied the existence of 'natural laws' in economic life (1863). His most important work was *Die Nationalökonomie der Gegenwart und Zukunft* (1848), where he discussed the theories of Friedrich List, Adam Müller, and especially those of Adam Smith. With his sharp criticism of self-interest and egoism as the central determinant of Smith's economic system – and the emphasis on ethical principles and the historically changing patterns of economic development – Hildebrand launched the attacks on Smith and the classical economists that were subsequently continued by many German historical economists.

The largest part of his main work was devoted to a discussion of socialism and communism, which he sharply rejected. Hildebrand focused his attention on the then little known Friedrich Engels and his recently published *Conditions of the Working Class in England* ([1848], 1922, pp. 125–90). He particularly criticized Engel's euphemistic description of pre-industrial conditions and contrasted it with empirical data that showed quite a different picture.

While being aware of current social problems Hildebrand perceived capitalist development most optimistically and envisioned as its last stage of development – the so-called 'credit economy' – a society where an advanced banking system would provide credit to a worker according to his morals and character and where thereby the monopoly of the capitalist class on capital would be broken (Hildebrand [1864], 1922, pp. 354–5). This theory of stages has to be regarded as Hildebrand's capitalist utopia, his liberal answer to socialism and communism.

Hildebrand's importance and his influence on the German Historical School has generally been underestimated; after all Hildebrand was – as Max Weber remarked – the only one really to work with the historical method. He undertook statistical studies – he regarded statistics as an important tool for detailed historical and empirical research (1865) – and wrote historical monographs (1866). He thus anticipated much of the research programme of the 'younger historical school' and the Verein für Socialpolitik, which he joined – as the only economist of the 'older historical school' – as a charter member in 1873.

Hildebrand stood for a kind of progressive liberalism that intended to reshape Germany along the lines of England, which he admired.

See Also

▶ Historical School, German

Selected Works

1848. *Die Nationalökonomie der Gegenwart und Zukunft und andere gesammelte Schriften,* ed. H. Gehrig. Jena: Gustav Fischer, 1922. It contains the articles Die gegenwärtige Aufgabe der Wissenschaft der Nationalökonomie (1863), *Die wissenschaftliche Aufgabe der Statistik* (1865), and *Natural-, Geld- und Kreditwirtschaft* (1864).

1866. Zur Geschichte der deutschen Wollenindustrie. *Jahrbücher für Nationalökonomie und Statistik* 6: S186–S254; 7: S81–S153.

Hilferding, Rudolf (1877–1941)

Roy Green

Keywords

Banking industry; Bernstein, E.; Böhm-Bawerk, E. von; Finance capital; Hilferding, R.; Kautsky, K.; Labour theory of value;

Lenin, V. I.; Luxemburg, R.; Marx, K. H.; Organized capitalism; Social democracy; Socialism

JEL Classifications

B31

Hilferding blended Marxist economics and Social Democratic politics in a career cut tragically short by the rise of fascism in Germany. He studied medicine at the University of Vienna, but soon showed more interest in organizing the student socialist society. After graduating in 1901, he helped Max Adler to found the *Marx- Studien* (1904–23), a series which was to become the theoretical flagship of 'Austro- Marxism'. The first volume contained a vigorous defence of the labour theory of value by Hilferding himself against Böhm-Bawerk's marginalist critique, *Zum Abschluss des Marxschen Systems* (1896). It earned him his intellectual spurs in the German-speaking socialist movement.

At the same time, Hilferding was already contributing to debate within the German Social Democratic Party (SPD) through its journal, *Die Neue Zeit*. There, on the controversial 'mass strike' issue, he steered a course for the party leadership between Eduard Bernstein's 'revisionist' abandonment of the socialist goal and Rosa Luxemburg's revolutionary commitment to it (1903/4, 1904/5). He was rewarded with an appointment in 1906 as economics lecturer at the party school in Berlin, and then as foreign editor of the party newspaper, *Vorwärts*. From 1907, he also wrote regularly for the newly established journal of the Austrian Social Democrats, *Der Kampf*.

Hilferding published his major work, *Das Finanzkapital*, in 1910; it was immediately hailed by such diverse figures as Kautsky (1911), Lenin (1916) and Bukharin (1917), as a path-breaking development of Marxist economic analysis. Essentially, Hilferding argued that the concentration and centralization of capital had led to the domination of industry and commerce by the large banks, which were transformed into 'finance capital' (1910, p. 225). The socialization of production effected by finance capital required a correspondingly increased economic role for the state. Society could therefore plan production by using the state to control the banking system:

> The socializing function of finance capital facilitates enormously the task of overcoming capitalism. Once finance capital has brought the most important branches of production under its control, it is enough for society, through its conscious executive organ – the state conquered by the working class – to seize finance capital in order to gain immediate control of these branches of production . . . . Even today, taking possession of six large Berlin banks would mean taking possession of the most important spheres of large-scale industry . . . (1910, pp. 367–8)

This chain of reasoning, however, tended to exaggerate not only the leverage of the banks over industry, but also the role of the state in the organization of production. While it convinced Hilferding that socialism could be introduced by a determined majority in parliament, it demonstrated to Lenin that socialism would not be possible unless the state was 'overthrown' by a determined minority outside parliament. Their common point of reference was the centrality of the state – rather than society – in the 'latest phase of capitalist development'. It forced socialists to make a choice between parliamentarism and insurrection, the very nature of which contributed to the defeat of the labour movement in Germany and the rise of party dictatorship in Russia (Neumann 1942, pp. 13–38). Although theory cannot be held responsible for the course of history, it may influence political judgements which tip the balance at decisive moments. Hilferding's generation lived through many such moments.

When war broke out in 1914, Hilferding associated himself with the SPD minority which voted against war credits and which later formed the Independent Social Democrats (USPD). He spent most of the war on the Italian front, having been drafted into the Austrian army as a doctor, and returned to Berlin as editor of the USPD journal, *Freiheit*. Hilferding successfully opposed USPD affiliation to the Third International; his speech against Zinoviev at the Halle conference of 1920 – published under the title, 'Revolutionäre Politik oder Machtillusionen?' –

was a decisive turning point. Once the embryonic Communist Party (KPD) forced a split on the issue, however, he saw no alternative to reunification with the remnants of the SPD.

During the 1920s, Hilferding turned his attention almost entirely to the political and economic problems facing the new German republic. He was a leading member of the Reich Economic Council, twice minister of finance and an active participant in the discussions on 'workers' councils' and the government's 'socialization' programme. Hilferding's first stint as minister of finance lasted only seven weeks in the Stresemann government of 1923. Although he had no opportunity to implement his proposals, he devised a plan for currency reform involving the introduction of a *Rentenmark* backed by gold as part of an anti-inflation package. By the time Hilferding returned to the same post in the Müller government of 1928/9, economic conditions had worsened; his predicament was appreciated by Schumpeter who wrote, 'we now have a socialist minister who faces the exceptionally difficult task of curing or improving a situation bequeathed by non-socialist financial policies' (quoted in Gottschlacht 1962, p. 24). A less sympathetic observer, however, portrayed Hilferding at this time as 'the theorist of coalition politics in the period of capitalist stabilisation' (see Gottschlacht 1962, p. 204), blinded by theory to the imminent fascist danger.

Pursuing the logic of *Das Finanzkapital*, Hilferding had developed a theory of 'organized capitalism', a term he first used in 1915 in *Der Kampf*, and then explained more fully in 1924 in *Die Gesellschaft*. He summarized the approach at the SPD's Kiel conference in 1927: 'Organized capitalism means replacing free competition by the social principle of planned production. The task of the present Social Democratic generation is to invoke state aid in translating this economy, organized and directed by the capitalists, into an economy directed by the democratic state' (see Neumann 1942, p. 23). Ironically, this was the very position of an earlier Social Democratic leadership which Marx had singled out for criticism. Commenting on the demand for a 'free state' in the 1875 Gotha programme, Marx wrote:

> It is by no means the goal of workers who have discarded the mentality of humble subjects to make the state 'free'. In the German Reich the 'state' has almost as much 'freedom' as in Russia. Freedom consists in converting the state from an organ superimposed on society into one thoroughly subordinate to it; and even today state forms are more or less free depending on the degree to which they restrict the 'freedom of the state'. (Marx 1891, p. 354.

While Hilferding understood that in capitalist society power lay with capital and was exercised by the representatives of capital in the management structure of the great corporations, he failed to see that democratic control over the productive forces would require a change in the relationship of power *within* the corporation itself.

Organized labour could use the state to accelerate this process of social transformation and to create the centralized institutional machinery necessary for the 'associated producers' to plan directly the whole economy; but the notion that the state itself could perform this task rested upon an illusion. In attempting to replace the domination of capitalist employers with the domination of a 'democratic state', Hilferding and the party leadership achieved only one practical result: 'Unwittingly, they strengthened the monopolistic trends in German industry' (Neumann 1942, p. 21). The state domination which followed was far from democratic.

Hilferding, a Jew, was forced into exile after 1933, first in Switzerland via Denmark and then in France. In an unfinished manuscript, *Das historische Problem*, he set about revising his whole conception of the state. The problem was now said to consist 'in the change in the relation of the state to society, brought about by the *subordination of the economy* to the coercive power of the state .. .' (quoted by Bottomore, Introduction to Hilferding, 1981, p. 16, emphasis in original). Hilferding briefly presented his new approach in the New York *Socialist Courier* in 1940; there, like Marx, he drew a rueful comparison between Germany and Russia. The state had not 'withered away' under Soviet communism:

> History, that 'best of all Marxists', has taught us another lesson. It has taught us that, in spite of Engels' expectations, the 'administration of things'

> may become an unlimited 'domination over men', and thus lead not only to the emancipation of the state from the economy but even to the subjection of the economy by the holders of state power. (1981, p. 376 n.)

It was too late for Hilferding's brave reassessment to influence the course of events. In 1941, he died in the hands of the Gestapo.

Selected Works

Books

1904. *Böhm-Bawerk's criticism of Marx* (Ed. P. Sweezy). London: Merlin Press, 1975.

1910. *Finance capital: A study of the latest phase of capitalist development*. London: Routledge & Kegan Paul, 1981.

Articles

1902/3. Der Funktionswechsel des Schutzzolles. Tendenz der modernen Handelspolitik. *Die Neue Zeit* 21, 2.

1903/4. Zur Frage des Generalstreik. *Die Neue Zeit* 22, 1.

1904/5. Parliamentarismus und Massenstreik. *Die Neue Zeit* 23, 2.

1915a. Historische Notwendigkeit und notwendige Politik. *Der Kampf* 8.

1915b. Arbeitsgemeinschaft der Klassen? *Der Kampf* 8.

1924a. Probleme der Zeit. *Die Gesellschaft* 1, 1

1924b. Realistischer Pazifismus. *Die Gesellschaft* I, 2.

1933. Zwischen den Entscheidungen. *Die Gesellschaft* 10.

1933/4. Revolutionärer Sozialismus. *Zeitschrift für Sozialismus* 1.

1934/5. Macht ohne Diplomatie – Diplomatie ohne Macht. *Zeitschrift fur Sozialismus* 2.

1940. State capitalism or totalitarian state economy. In *Socialist courier*. New York. Reprinted in *Modern Review* 1(1947).

Published Speeches

1919. Zur Sozialisierungsfrage. 10th Congress of the German trade unions, Nuremberg, 30 June–5 July. Berlin.

1920a. Revolutionäre Politik oder Machtillusionen? Speech against Zinoviev at the annual conference of the USPD in Halle, Berlin.

1920b. Die Sozialisierung und die Machtverhältnisse der Klassen. 1st Congress of Works Councils, 5 October. Berlin.

1927. Die Aufgaben der Sozialdemokratie in der Republik. Annual conference of the SPD in Kiel. Berlin.

1931. Gesellschaftsmacht oder Privatmacht über die Wirtschaft. 4th AFA (Allgemeiner freier Angestelltenbund) trade union congress in Leipzig. Berlin.

Bibliography

von Böhm-Bawerk, E. 1896. *Karl Marx and the close of his system* (Ed. P. Sweezy). London: Merlin Press, 1975.

Bukharin, N. 1914. *Imperialism and world economy.* London: Merlin Press, 1972.

Gottschlacht, W. 1962. *Struktur veränderungen der Gesellschaft und politisches Handeln in der Lehre von Rudolf Hilferding*. Berlin: Duncker & Humblot.

Kautsky, K. 1911. Finanzkapital und Krisen. *Die Neue Zeit* 29: 764–772, 797–803, 838–846, 874–883.

Lenin, V.I. 1916. *Imperialism, the highest stage of capitalism*. Moscow: Foreign Languages Publishing House, 1947.

Marx, K. 1891. Critique of the Gotha programme. In *The first international and after*, ed. D. Fernbach. Harmondsworth: Penguin, 1974.

Neumann, F. 1942. *Behemoth*. London: Victor Gollancz.

Hill, Polly (1914–2005)

C. A. Gregory

Keywords

Development economics; Economic anthropology; Hill P.; Inequality; Lewis W. A.

JEL Classifications

B31

Polly Hill was born on 10 June 1914 into a remarkable Cambridge family that includes Nobel Prize winning physiologist A.V. Hill (her father) and J.M. Keynes (her mother's brother) among its many distinguished members. She graduated from Cambridge in 1936 with a degree in economics.

Her first job upon leaving university was with the Royal Economic Society as an editorial assistant, a position she held for two years (1936–8). Her next appointment was a one year (1938–9) research position with the New Fabian Research Bureau (which almost immediately re-amalgamated with the Fabian Society) where she wrote her first book, *The Unemployment Services* (1940). This book was concerned to expose the inefficiency and inhumanity of the system of unemployment relief and to make constructive proposals. Polly Hill's commitment to social justice never waned: economic inequality is the central theme of all her books.

At the outbreak of the war she was obliged, as an unmarried young woman, to become a temporary civil servant. She worked first, briefly, in the Treasury, then for a long time in the Board of Trade and finally in the Colonial Office. She resigned in 1951. After a period of unemployment she became a journalist for the weekly *West Africa.* She married in 1953 and moved to Ghana with her husband where, at the age of 40, she began her academic career. The academic posts she held there involved no teaching and she was able to become, as she put it, 'a pupil of the migrant cocoa farmers of southern Ghana'. She began her fieldwork as an economist and collected data using the questionnaire method, producing her second book, *The Gold Coast Cocoa Farmer: A Preliminary Survey* (1956) with characteristic speed and efficiency. The prevailing orthodoxy had it that sedentary food farmers in southern Ghana had suddenly taken up cocoa farming at the end of the 19th century with such a degree of success that cocoa exports had risen from nil to over 50,000 tons by 1914 – the largest quantity for any country. Polly Hill had uncritically accepted this orthodoxy and her subsequent realization that most farmers appeared to be migrants who had bought their land was to have a profound effect upon her intellectual methods. She abandoned the questionnaire method of data collection in favour of one that sought to develop generalizations on the basis of: (1) detailed fieldwork in one village; (2) fieldwork done by others elsewhere; (3) archival sources. She also began a lifelong struggle with development economists and other purveyors of orthodoxies based on casual empirical observation and 'common sense'. She drifted towards anthropology and history where the qualities of her empirical findings were recognized for what they were: revolutionary. She spent three and a half years collecting detailed evidence to substantiate her claim that the cocoa farmers were migrants and made many fascinating discoveries in the process. For example, she found that the matrilineal farmers adopted an entirely different mode of migration from patrilineal farmers: the former bought family lands with the aid of their kin, and were prepared to grant usufructural rights to their male and female kinsfolk; the latter clubbed together in so-called 'companies', groups of non-kin, the land being divided into strips from a base line, according to the contribution each had made, with subsequent division on inheritance always being longitudinal. Upon hearing of this Professor Meyer Fortes, then Professor of Social Anthropology at Cambridge, encouraged her to apply for a Smuts Visiting Fellowship. This enabled her to write *The Migrant Cocoa-Farmers of Southern Ghana: A Study in Rural Capitalism* (1963) which is now widely regarded as a classic. (She was awarded a Ph.D. in social anthropology from Cambridge under new special regulations in 1966 on the basis of it.) Mainstream writers on development have by and large ignored the book even though it contains telling criticisms of aspects of W.A. Lewis's work.

Following more fieldwork in Ghana, Nigeria and India she produced a further stream of books (1970a, 1970b, 1972, 1977, 1982, 1985, 1986) and many articles of outstanding quality which established her reputation as the world's foremost economic anthropologist. She was appointed a Fellow of Clare Hall in Cambridge in 1965 and subsequently to the prestigious

Smuts Readership in Commonwealth Studies (1973–9). Her publications documented in painstaking detail the complexity of agrarian relations in the tropical regions of the world in which she had worked. The books as a whole constitute an encyclopaedia of knowledge on the socio-economic conditions of poverty and economic inequality and her work ranged in scope from 'agrestic servitude' to 'zamindars'. Her oeuvre was much more than a compilation of facts, though. Her own data and that of others are presented in a theoretical context which broadened as her own field experience widened. She was unrelenting in her empirically based critiques of development economists and her 1986 book *Development Economics on Trial: The Anthropological Case for a Prosecution* was a concerted attempt to make them see the error of their ways.

Selected Works

1940. *The unemployment services.* London: Routledge.

1956. *The gold coast cocoa farmer: A preliminary survey.* Oxford: Oxford University Press.

1963. *The migrant cocoa-farmers of Southern Ghana: A study in rural capitalism.* Cambridge: Cambridge University Press.

1970a. *The occupations of migrants in Ghana.* Ann Arbor: University of Michigan Press.

1970b. *Studies in rural capitalism in West Africa.* Cambridge: Cambridge University Press.

1972. *Rural Hausa: A village and a setting.* Cambridge: Cambridge University Press.

1977. *Population, prosperity and poverty: Rural Kano 1900 and 1970.* Cambridge: Cambridge University Press.

1982. *Dry grain farming families: Hausaland (Nigeria) and Karnataka (India) compared.* Cambridge: Cambridge University Press.

1985. *Indigenous trade and market places in Ghana, 1962–64*, Jos oral history and literature texts. Nigeria: University of Jos.

1986. *Development economics on trial: The anthropological case for a prosecution.* Cambridge: Cambridge University Press.

Hirschman, Albert Otto (born 1915)

M. S. McPherson

Hirschman was born on 7 April 1915 in Berlin. After attending the Sorbonne and the London School of Economics he obtained a doctorate in economic science from the University of Trieste in 1938. His early career was dominated by the struggle against fascism in Europe (Coser 1984). He actively supported the underground opposition to Mussolini while in Italy in the mid-1930s, fought with the Spanish Republican Army in 1936 and later with the French Army until its defeat in June 1940. He stayed on in Marseilles six months more, engaging in clandestine operations to rescue political and intellectual refugees from Nazi-occupied Europe. He avoided arrest by leaving France for the United States in January 1941. There he produced his first book, *National Power and the Structure of Foreign Trade* (1945), which introduced some of the main themes of what is now called 'dependency theory'.

After the war he served as an economist in the Federal Reserve Board until 1952, when he left for Colombia where he stayed four years. Beginning in 1956 he held professorships successively at Yale, Columbia and Harvard, and in 1974 was appointed professor at the Institute for Advanced Study in Princeton.

Hirschman has been a leading figure in economic development since the publication in 1958 of his second book, *The Strategy of Economic Development.* Hirschman's analysis grew out of extensive practical experience in Colombia as an adviser both to its government and to private firms. Characteristically, Hirschman dissented from orthodox views of both right and left, arguing that neither *laissez faire* nor 'rational' economy-wide planning made sense for poor countries. Government needed to encourage 'unbalanced growth', deploying its scarce decisionmaking capacities strategically to set up disequilibria that would stimulate effort and mobilize hidden and underutilized resources. Targeting development efforts on key industries with strong 'linkages' to other parts of the economy could stimulate a favourable dynamic.

Hirschman later provided the label 'possibilism' (1971) for the outlook that shaped much of his thought on development and on which he elaborated in many further books and articles. When social science focuses exclusively on the search for general laws, it obscures the irreducible role of the unique and the unpredictable in human affairs. This causes progress to be viewed either as ensured by the application of general rules or thwarted by the presence of inescapable obstacles. But history reveals that actual social change often follows paths that are *a priori* quite unlikely, turning obstacles into opportunities and confounding rules with unanticipated consequences. From this starting point, Hirschman has cultivated an approach to development problems which embodies respect for complexity and openness to the possibility of genuine novelty – what he once called the discovery of 'an entirely new way of turning a historical corner' (1971, p. 27).

Since 1970, Hirschman has been bringing his possibilist approach to bear on broader problems of social theory. His slim volume, *Exit, Voice, and Loyalty* (1970) revealed the unexpected richness to be found in comparing the implications of dissatisfied clients alternatively *exiting* from an organization or giving *voice* to their complaints. This volume, like Hirschman's more recent work (1982b) on the forces that propel individuals and societies into and out of periods of intense political involvement, explores issues on the borderline between economics and politics. But unlike most economists with an interest in 'public choice', Hirschman shows no inclination to reduce politics to economics. Indeed, both works stress that standard models of economic behaviour fail to make sense of familiar forms of 'public-minded' behaviour such as voicing one's convictions on public matters, participating in demonstrations or working to support candidates for office.

Hirschman's propensity to devise analytical formulations that express rather than conceal the complexities of human motivations and institutions is evident also in his studies of historical views of capitalism (1977, 1982a). Hirschman shows that capitalism has been seen as a powerful civilizing influence and alternatively as a destroyer of the moral and social fabric; still other views have portrayed capitalism, for better or worse, as too feeble to overcome the restraints of preceding social forms. These competing ideological views have evolved, Hirschman notes, in total isolation from one another. A fuller view would recognize that all these contradictory tendencies are present at once, but to recognize this truth would be highly inconvenient, making it 'much more difficult for the social observer, critic, or "scientist" to impress the general public by proclaiming some inevitable outcome of current processes' (1982a).

'But', Hirschman concludes, in a question that captures well his own unique stance in modern social science, 'after so many failed prophecies, is it not in the interest of social science to embrace complexity, be it at some sacrifice of its claim to predictive power?' (1982a, p. 1483).

See Also

- ▶ Exit and Voice
- ▶ Linkages

Selected Works

1945. *National power and the structure of foreign trade.* Berkeley/Los Angeles: Bureau of Business and Economic Research/University of California.

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

1970. *Exit, voice, and loyalty: Responses to decline in firms, organizations, and states.* Cambridge, MA: Harvard University Press.

1971. Introduction: Political economics and possibilism. In *A bias for hope: Essays on development and Latin America,* ed. A.O. Hirschman. New Haven: Yale University Press.

1977. *The passions and the interests: Political arguments for Capitalism before its triumph.* Princeton: Princeton University Press.

1982a. Rival interpretations of market society: Civilizing, destructive, or feeble? *Journal of Economic Literature* 20: 1463–1484.

1982b. *Shifting involvements: Private interest and public action.* Princeton: Princeton University Press.

1986. *Rival views of market society and other essays.* New York: Viking-Penguin International.

Bibliography

Coser, A. 1984. *Refugee scholars in America: Their impact and their experiences*. New Haven: Yale University Press.

Hirshleifer, Jack (1926–2005)

Elizabeth Landaw and David K. Levine

Abstract

Jack Hirshleifer was one of the leaders of the 'information and uncertainty' revolution in economics. His work on the role of time and uncertainty in asset markets and the value of information plays a fundamental role in modern economic thought. Hirshleifer was also a leader in the 'imperial' school of economics, taking the lead in expanding economic thought to areas such as evolution and conflict, which traditionally were studied by other social science disciplines.

Keywords

Altruism; Arms races; Best-shot game; Conflict resolution; Cost–benefit analysis; Evolutionary economics; Free-rider problem; Hirshleifer, J; Indeterminacy of equilibrium; Information economics; Intellectual property; Intertemporal choice; Modigliani–Miller theorem; Patent races; Public investment; Risk; Speculation; Status and economics; Term structure of interest rates; Uncertainty; Value of information

JEL Classifications

B31

Jack Hirshleifer was born on 26 August 1926 in Brooklyn, New York. He graduated at the top of his class of 855 students from Erasmus Hall High School in New York, then enrolled at Harvard in 1942, studying government and other social sciences. He was quickly drawn, however, to economics, which provided him with a 'useful set of tools and methods'. In 1943 Hirshleifer's career as a budding economist went on hold when he enlisted for active service duty in the US Naval Reserve, serving on an aircraft carrier in the Pacific until 1945. This experience inspired in him a long- lasting and deep interest in military arms races. After the war, he resumed his studies at Harvard, receiving a Ph.D. in economics in 1950. Hirshleifer's research career started at the RAND Corporation in Santa Monica. In 1955 he became an assistant professor at the Graduate School of Business at the University of Chicago, and then returned to Los Angeles in 1960 as an Associate Professor of Economics at UCLA, becoming full professor two years later. In 1975 he became what is now called a 'Distinguished University Professor', thus becoming a member of the most elite group of the University of California faculty.

Hirshleifer was an economic theorist with broad-ranging interests. He not only wrote extensively in areas of general economic interest such as capital theory or economics of uncertainty and information, but also wrote and often laid out the foundations for areas outside the traditional scope of economics, including conflict theory and evolutionary modelling.

Some of Hirshleifer's early work focused on the intertemporal theory of interest and investment. Today, this research helps us better understand such topics as intertemporal choice, decisions under uncertainty, the choice of discount rate for public investments, or liquidity and the term structure. His early interest in capital theory led not only to scores of influential articles but also to pioneering and detailed examination of the concepts of interest rate, investment and capital, which are integrated into his book *Investment, Interest and Capital* (1979) and later in the volume of collected articles, *Time, Uncertainty, and Information* (1989). The earlier book and associated articles became a framework for modern finance theory and for understanding investment decisions under uncertainty.

Hirshleifer also made a lasting contribution to the theory of speculation. He showed that differences in taste are not enough to explain speculation; rather, speculation must arise from differences in beliefs. He was the first to analyse speculation in a full general-equilibrium model, with different structures of market completeness carefully considered. Although not generally recognized as such, the 1975 *Quarterly Journal of Economics* paper is also the first paper to point out the indeterminacy of equilibrium when markets are incomplete.

Early in his career, Hirshleifer was instrumental in the information economics revolution and is considered today to be one of its founding fathers. He made the abstract ideas of contingent claims concrete through his examples and applications. In the process, he helped develop fundamental tools, such as the covariance of risks, the analysis of gambling and insurance, the Modigliani–Miller theorem, and the analysis of public investment. Most notably, his 1971 *American Economic Review* paper, 'The Private and Social Value of Information and the Reward to Inventive Activity', became highly influential and one of the most cited papers in the economics of information. The paper demonstrates that competitive markets need not reflect the social value of information. Hirshleifer's example of an inventor who can invest based on the knowledge of the impact of his invention shows that there can be an oversupply of inventive activity. This 'race to be first' has its reflection in the current literature on patent races, starting with Fudenberg et al. (1983) and continuing through such work as Gallini and Scotchmer (2001). It is the key to understanding a fundamental problem in intellectual property law, which the profession is only now coming to grips with. Hirshleifer also identifies what the profession now refers to as the 'Hirshleifer effect': new and more reliable information can have a negative social value if the early information on risks makes these risks uninsurable.

In addition to his founding contributions in information economics, Hirshleifer had a lifelong interest in conflict, beginning with his earliest work on war damages. Late in his career this area was the focus of his contributions, and he was a leader in extending economic methods to problems more traditionally studied in political science. Just as Hirshleifer was first drawn to economics for its methods and tools, he argued that the traditional assumptions of microeconomic theory are too narrow. One such idea, he maintained, the idea of cooperation or 'mutually beneficial exchange via markets', is only one form of many different forms of human interactions. An alternative way would be simply to take what you want away from other parties. This is still economics, since scarcity and competition and optimization and equilibrium are all involved. Conflicts, and indeed all struggles for power and influence, are important economic activities, as important as exchange. He explored an economic approach to conflicts not only in the context of war but also crime, litigation, strikes and political campaigns.

His work on conflict shows how 'Peace is more likely to the extent that the decisiveness of conflict is low, or . . . if the stakes are small or the technology favors the defense. More surprisingly, perhaps, increased productive complementarity between the parties does not systematically favor peace. . .the poorer side is generally motivated to invest more heavily in fighting effort. So conflict can become an income-equalizing process' (1991, p. 133). It is what Hirshleifer calls the 'paradox of power': poor or weaker contestants defeat large ones. Subsequent work shows how a narrow range of possible settlements increases the potential for conflict and how increasing returns followed by diminishing returns explains the monopoly on military force within the state, while also explaining the multiplicity of states. A number of his papers analysing conflict as opposed to cooperation are collected in *Economic Behavior in Adversity* (1987a). Hirshleifer wrote broadly on expanding the domain of economic discourse to include the 'rational' evolutionary analysis of altruism and spite. He believed that the standard economic postulate of fixed preferences is wrong and instead argued that evolution plays a pivotal role in shaping not only people's physical make-up but also tastes. In one of his most influential papers, ‘The Expanding Domain of Economics’ (1985), Hirshleifer reviews how the economic logic of optimization, trade-off and of equilibrium can and should be applied to a wide variety of ‘non-economic’ problems. He writes that economics constitute ‘the universal grammar of social sciences’ (1985, p. 53) but that there is the wide area of ‘noneconomics’ that economists have to become aware of and get over their ‘tunnel vision about the nature of man and social interactions’.

The paper examines different kinds of altruistic preferences, including what would now be called by experimentalists the ‘warm-glow’ effect. As an application, Hirshleifer discusses Becker’s ‘rotten kid’ theorem, showing how a selfish parent can gain from altruism. Still other theories of preferences, including models of status, such as the rat-race are examined. Hirshleifer opened up new areas; by now, much of this ‘non-economic’ economics is widely studied by economists, and models of altruism and status proliferate.

Key to Hirshleifer’s contribution is the underlying point of view of ‘as-if’ rationality – altruism must provide some benefit to the altruist. This was the starting point of much of the modern evolutionary economics literature – for example, the work of Kandori et al. (1993) and Young (1993). From this perspective, Hirshleifer examined models such as the psychological model of ‘anger, gratitude, response’ and argued that this seemingly irrational behaviour does indeed benefit the individual. Yet Hirshleifer’s view of evolution was an eminently practical one: it was firmly grounded in his desire to understand why voluntary exchange arises in some situations, but conflict in others.

Although not primarily an experimentalist, Hirshleifer, together with Glenn W. Harrison, conducted a fundamental experiment on the incentives to free ride (1989). As Hirshleifer surely imagined, increasing incentives to free ride lead to more free riding. The experiment introduced the ‘best-shot’ game, a public goods contribution game in which only the largest contribution to the public good matters. In this type of game it is socially and individually optimal for only one player to contribute, and, unlike many other types of public goods games, this theoretical prediction is exactly what happens in the laboratory.

Hirshleifer’s interest in risk and investment extended to public investment and cost–benefit analysis. Although the fact is not widely known, he co-authored an important study of alternative routes for bringing water from northern to southern California, as well as a follow-up years later after one of the projects was chosen and built. He was fond of saying that much of his scepticism of government arose from the fact that of three routes one was clearly worse than the other two – and that was the one that was actually built.

Jack Hirshleifer’s love of social sciences, particularly economics, was one of his endearing traits. He liked nothing better than contemplating new puzzles and exchanging ideas with his colleagues. Although officially he changed his status to Professor Emeritus in 1991, he never ceased working, writing, reviewing, and lecturing. Colleagues would find him working every day in his office, door open, sitting behind his cluttered desk with an inviting smile. He continued to work until the very end of his life, and was proud that he was able to proofread – he sent back the galleys of the seventh edition of his very popular textbook, *Price Theory and Applications.* He hosted Thursday lunches at the UCLA Faculty Club, which became a gathering place famous for spirited discussions. A kind and approachable man dedicated to his work, Jack’s rule was economics and not gossip. Those who knew him remember him for his personal warmth and sense of humour.

Although the two areas in economics that have especially felt the impact of Hirshleifer’s work are information economics and conflict resolution, Hirshleifer shed light on many other fields including capital theory, finance, bioeconomics and experimental economics. With his insatiable intellectual curiosity, he was never short of good ideas, illustrating them through carefully worked out and accessible examples. He would plant many seeds and often leave to others to develop sophisticated theories. Yet with his standing concern for the value of rigorous scholarship, Hirshleifer was one of the pioneers who transformed economics into the scholarly science that it is today.

See Also

- ▶ Evolutionary Economics
- ▶ Intertemporal Choice
- ▶ Risk

Selected Works

1958. On the theory of optimal investment decision. *Journal of Political Economy* 66: 329–352.

1960. (With J. C. DeHaven.) *Water supply: Economics, technology and policy.* Chicago: University of Chicago Press.

1961. Risk, The discount rate, and investment decisions. *American Economic Review* 51: 112–120.

1965. Investment decision under uncertainty: Choice-theoretic approaches. *Quarterly Journal of Economics* 79: 509–536.

1966. Investment decision under uncertainty: Applications of the state-preference approach. *Quarterly Journal of Economics* 80: 252–277.

1967. (With J. W. Milliman.) Urban water supply: A second look. *American Economic Review* 57: 169–178.

1970. *Investment, interest, and capital.* Englewood Cliffs: Prentice-Hall.

1971. The private and social value of information and the reward to inventive activity. *American Economic Review* 61: 561–574.

1975. Speculation and equilibrium: information, risk, and markets. *Quarterly Journal of Economics* 89: 519–542.

1978. Competition, cooperation, and conflict in economics and biology. *American Economic Review* 68: 238–243.

1979. *Investment, interest and capital.* Saddle River: Prentice Hall.

1985. The expanding domain of economics. *American Economic Review* 75: 53–68.

1987a. *Economic behavior in adversity.* Chicago: University of Chicago Press.

1987b. Conflict and settlement. In *The New Palgrave: A dictionary of economics,* vol.1, ed. J. Eatwell, M. Milgate, and P. Newman. London: Macmillan.

1987c. Investment decision criteria. In *The New Palgrave: A Dictionary of Economics,* vol. 2, ed. J. Eatwell, M. Milgate, and P. Newman. London: Macmillan.

1989a. (With G. Harrison.) An experimental evaluation of weakest link/best shot models of public goods. *Journal of Political Economy* 97: 201–225.

1989b. *Time, uncertainty, and information.* Oxford: Basil Blackwell.

1991. The technology of conflict as an economic activity. *American Economic Review* 81: 130–134.

1992. (With J. Riley.) *Analytics of uncertainty and information.* Cambridge: Cambridge University Press.

2005. (With A. Glazer and D. Hirshleifer.) *Price theory and applications: Decisions, markets, and information,* 7th ed. Cambridge: Cambridge University Press.

Bibliography

Fudenberg, D.R., J. Gilbert, J. Stiglitz, and J. Tirole. 1983. Preemption, leapfrogging and competition in patent races. *European Economic Review* 22: 3–31.

Gallini, N., and S. Scotchmer. 2001. *Intellectual property: When is it the best incentive system?* Berkeley: University of California Press.

Kandori, M., G. Mailath, and R. Rob. 1993. Learning, mutation, and long run equilibria in games. *Econometrica* 61: 29–56.

Young, H.P. 1993. The evolution of conventions. *Econometrica* 61: 57–84.

Historical Cost Accounting

G. Whittington

Historical cost is the cost at which an asset was actually purchased. This is the value traditionally imputed to assets in accounts. Valuation at historical cost was a natural process in the early days of accrual accounting. Historical cost represented money which had been paid out (or a liability created) which was not to be charged against profit because it represented the creation of an asset, rather than an expense. Thus, the logic of double entry suggested that assets should, initially at least, be valued at historical cost.

However, the survival to the present day of historical cost as a valuation basis in accounts is not due merely to its easy assimilation into double entry book-keeping. For accountants, historical cost has at least two attractions relative to current valuation bases, such as current replacement cost or realizable value. Firstly, it is relatively objective, having been established by a verifiable transaction on which two independent accountants would be likely to take the same view, whereas current values involve estimating what would happen if a transaction (replacement or sale) were to occur. Secondly, it is conservative, insofar as it does not recognize gains in value which have taken place since the asset was acquired.

It is alleged (e.g. by Ijiri 1971) that these properties help historical cost accounts to fulfil the stewardship function of providing users of accounts with a relatively objective statement of the financial transactions of those responsible for managing the assets of the business. Accounts which were surrounded by greater uncertainty, due to the subjectivity of the valuation base, might not fulfil this function so well. Furthermore, the conservative practice of not showing any gains in the value of assets due to price rises since the acquisition date is a protection against the manipulation of accounts by unduly optimistic or unscrupulous managers.

On the other hand, the principle of conservatism has been applied so strongly that it has been allowed to modify historical cost in certain cases where current market value is lower than cost. Thus, in the United Kingdom, the valuation rule for current assets such as stocks and work in progress is, in conventional accounts, ‘cost or current market value, whichever is the lower’. For fixed assets of limited life, depreciation is traditionally written off the historical cost of the asset over its lifetime. Written down historical cost does not claim to be a close approximation to current market value, but it is less likely to exceed market value than unadjusted historical cost. The estimation of depreciation reduces the objectivity of historical cost valuation, as does the introduction of lower market values, thus diminishing one of the important advantages claimed for historical cost.

Another common breach of the historical cost system in conventional accounting practice is the periodic revaluation of fixed assets in the balance sheet. This has become accepted practice in the United Kingdom, as a response to the pressure for more relevant information in a period of rising prices. The integrity of historical cost profit is usually preserved by not passing the revaluation through the profit and loss account, that is, the increased value of the assets is regarded as a capital gain, giving rise to an increase in undistributable reserves rather than in profit. On the other hand, the principle of conservatism is applied so that future depreciation charges against profit are based on the revalued amount, so that the charges are higher, and profits lower, than if the revaluation had not taken place. Thus, the effect of the revaluation is to depress the future accounting rate of return by increasing the numerator (profit, after charging depreciation) and increasing the denominator (net assets).

The above description applies to current conventional accounting practice in the United Kingdom. However, historical cost is currently the basis of conventional financial accounts in all major capitalist economies, and in each case there are departures from strict historical cost to meet difficulties which have been encountered in practice; for example, certain Latin American countries which have suffered very high inflation rates have requirements for applying indexation to historical cost. The widespread survival of historical cost accounting can be attributed to two factors. Firstly, the firm transactions base of historical cost accounting gives it a degree of objectivity which, although not as great as might appear at first sight, is not matched by alternative systems. Secondly, vast experience of implementing historical cost has accumulated. Thus, accountants are better equipped to implement it rather than alternative systems, such as current cost accounting.

Accounting practice has evolved as a pragmatic response to practical difficulties, and most accountants think of it in this way rather than as the rational application of theoretical principles. Thus, it seems likely that there will be powerful support from the accounting profession for the continued evolution of generally accepted accounting principles (known in the United States as GAAP), based on historical cost but with an increasing degree of modification, rather than its revolutionary replacement by a different valuation base, such as current cost accounting (as proposed by the Sandilands Report (1975) in the United Kingdom).

See Also

- ▶ Accounting and Economics
- ▶ Inflation Accounting

Bibliography

Ijiri, Y. 1971. A defence for historical cost accounting. In *Asset valuation and income determination, a consideration of the alternatives*, ed. R.R. Sterling. Houston: Scholars Book Co., 1975.

Sandilands Report. 1975. *Inflation accounting: Report of the inflation accounting committee under the chairmanship of F.E.P.* Sandilands, Cmnd. 6225, London: HMSO, September 1975.

Historical Demography

Carl Mosk

Abstract

Historical demography deals with population dynamics prior to and during early phases of industrialization. Using family reconstruction historical methodology, demographers have found partial answers to Malthusian questions revolving around mortality and fertility rates in religious records yielding estimates for marriage, life expectancy and reproduction within marriage. Employing cause of death estimates and Hutterite index measures for the proportion of women married and the level of their reproduction within marriage, historical demographers have developed tentative answers to demographic transition queries. Historical demography has contributed much to our understanding of historical population dynamics.

Keywords

Black Death; Child labour; Climate; Demographic transition; European Fertility Project; Family reconstitution method; Fertility; Fertility transition; Food prices; Henry, L.; Historical demography; Homeostatic equilibrium; Industrialization; Industrial revolution; Iron law of wages; Malthus, T. R.; Malthusian economy; Marriage; Mortality; Population dynamics; Population growth

JEL Classifications

J10; N3

Prior to European industrialization population grew in fits and starts, because the effects of the introduction of new crops like the potato or the reclaiming of uncultivated grasslands and forested slopes for irrigated rice paddy were short-lived, typically ushering in periods of stagnation. Why did pre-industrial populations increase in such a manner, slowly groping upward from one plateau to the next, perhaps even tumbling backward to ever lower plateaus before resuming forward progress? Was it fertility or mortality or an interaction of the two that constrained the growth process?

The Impact of Family Reconstitution

Our understanding of the dynamics of pre-industrial populations has been immeasurably increased by research in historical demography, fuelled by the development of family reconstitution for analysing records of births, deaths and marriages lodged in religious quarters – Catholic, Anglican and Lutheran parishes, Buddhist temples – in clan genealogies and in military records. Developed in the 1950s and 1960s by

French demographers, most notably by Louis Henry, the family reconstitution methodology exploits the fact that individuals are separately listed in vital registers that can be linked together to yield life histories moving from birth to marriage and to death. Henry's ingenuity lay in rigorously defining the period over which a family is under observation for the purposes of deducing its mortality and fertility history.

It should be emphasized that records of religious bodies, of clans and of military organizations are not the only sources that can be tapped by historical demographers. Other sources include censuses (Quebec initiated systematic censustaking in 1665); fiscal documents, for instance taxpayer lists (Japanese population counts for the rice tax paying population of the country are available from the early 17th century); property inventories and wills; archeological remains including preserved garbage dumps; cemetery data, both skeletons and gravestones; and eyewitness accounts recorded in literary documents. Hollingsworth (1969) offers a thorough review of the various methods, pinpointing strengths and deficiencies.

Still, it was the pioneering of a carefully elaborated family reconstitution methodology by French scholars working from records of parishes from the time of Louis XIV and Louis XV that opened the floodgates for systematic analysis of fertility and mortality in pre-industrial Europe and pre-industrial Asia. Particularly important was application of the methodology to England, where several thousand parish registers beginning prior to 1600 exist, and to Japan, where Akira Hayami and others have trained Henry's methodology upon Buddhist religious records (shūmon-aratame-chō) of births, deaths and marriages in analysing the population dynamics of villages during the Tokugawa (1600–1868) period. Hayami (1997) provides a useful history, replete with concrete examples, of the impact that historical demography has had on the understanding of pre-industrial population dynamics in Japan.

What is clear from the analysis of Buddhist registers for Japanese villages is that fertility within marriage was kept quite low in many parts of the country from the early 18th century onward, the intervals between births being drawn out through a combination of infanticide and taboos against having too many small offspring in the household at any one time. Whether Japanese peasants were concerned about excess competition for the family headship (only one child could take over the headship from the patriarch of the household), responding to a falling off in the demand for child labour on densely populated paddy rice fields, or whether they were attempting to maximize survivorship rates for each child allowed to live remains a matter for scholarly debate. What historical demography has shown is that the debate must be about why fertility was fairly low, not why mortality was fairly high.

Low- and High-Pressure Homeostatic Equilibriums

Systematic analysis of the English parish data has yielded one of the crowning achievements of post-Second World War historical social science: the securing of over 3.5 million totals for baptisms, burials and marriages drawn from 404 carefully selected Anglican parish records by a research team at Cambridge University headed up by E. A. Wrigley and R. Schofield. Developing a novel technique for projecting back population totals from the census of 1871 and from national level estimates of births and deaths generated from the 404 parish figures, Wrigley and Schofield (1981) were able to estimate population totals, and fertility (including the gross reproduction rate that gives the number of female births a woman is expected to have across her reproductive life) and mortality rates (including life expectancy at age zero) for England between 1550 and 1871. The Wrigley–Schofield 1981 volume was path-breaking not only in offering a remarkable data-set and a remarkable set of estimates for pre-industrial fertility and pre-industrial mortality. It was also pathbreaking in contesting the standard Malthusian interpretations of pre-Industrial Revolution British population dynamics.

In the standard argument the force explaining fluctuations in population size and growth rates was mortality. The Black Death reduced the ranks of the populace in the 14th century. More generally, plagues occurring between 1350 and 1660 acted as negative exogenous shocks absorbed by the British population, peasants and aristocrats alike being decimated by these waves of disease. In the Malthusian model this mechanism for regulating numbers is the positive mortality check, and populations so regulated are described as operating in a high-pressure homeostatic equilibrium, feedback running from population increase to increased food prices to enhanced mortality, thereby reducing population.

Wrigley and Schofield (1981) suggested that pre-industrial England operated as a low-pressure rather than a high-pressure equilibrium system, fluctuations in fertility driving fluctuations in population size over the long run. Indeed, the authors went so far as to suggest that there was a 50-year lag at work, surges in real wages generating surges in marriage and in births over a 50-year period. In offering a theory based upon the idea that the real wage drives population growth through its impact upon births, Wrigley and Schofield (1981) put forward a novel interpretation of the iron law of wages. This proposition states that increases in real wages due to accumulation of capital or technological improvements are ultimately choked off by population increase initiated by the improvement in wages.

The low-pressure homeostatic story accounting for the iron law of wages was not satisfactory to R. Lee, who devoted much effort to analysing the response of real wages to exogenous fluctuations in population size. For instance, Lee (1980) estimated an elasticity of minus one-and-a-half for the impact of population increase on real wages, a ten per cent increase in human numbers diminishing real earnings by 15 per cent. In Lee (1987), he pointed out that the 50-year lag is only one story that is consistent with the long-run movements in fertility and real wages advanced by Wrigley and Schofield.

In any event, the 50-year lag of Wrigley and Schofield and Lee's estimates for the impact of population increase on real wages are both based upon long-run movements in population, fertility and mortality. Equally interesting are the short-run dynamics for pre-industrial populations, fluctuations in climate –when the spring thaw permitting planting of new crops in the fields takes place, when the onset of cold fall temperatures dictates harvesting –driving movements in food prices, resulting in fluctuations in marriages, pregnancies, births and deaths. Analysing a large number of historical cases, Lee (1987) concluded that the vital rates do respond to upward and downward movements in food prices, pre-industrial societies being regulated in a homeostatic fashion that was responsive to exogenous changes in climate.

To examine more systematically the impact of fluctuations in food prices upon demographic behaviour in pre-industrial Europe and Asia, the Eurasian Project in Population and Family History has pioneered the use of longitudinal databases of household and individual records, eschewing the computation and analysis of aggregate demographic statistics generated from massive family reconstitution exercises like that carried out by Wrigley and Schofield (1981). A good illustration of the type of analysis stemming from this approach is Bengtsson et al. (2004). Generating results for Scania in southern Sweden, for eastern Belgium, for three villages in northern Italy, for a village in northern Japan, and for Liaodong in north-eastern China, the Eurasian Project suggests that demographic responses to short-run stress (that is, spikes in food prices) were fundamentally different in the West and in the East. In the East power, especially gender-based power, played a crucial role in shaping household demographic behaviour in the face of food scarcity, females getting less access to nutrition than males in the typical scenario. In the West, socio-economic status, especially ownership of land, mattered a great deal. When climatic variation forced up the price of foodstuffs, the landless suffered in Europe. In Asia it was young females who bore the brunt of the crisis.

Onset of the Demographic Transition

In addition to shedding light on Malthusian questions – on the relative importance of the positive mortality check and the preventive fertility check – historical demography has shed light on the question of when the fertility and mortality transitions began. To what extent did the onset of industrialization influence fertility and mortality? Is there evidence of fertility decline in early industrializing – or even completely pre-industrial – settings? The general overlap of industrialization and the demographic transition is evident. Heavily industrialized countries enjoy low fertility and low mortality. What is not evident is that there is a direct relationship between the onset of industrialization and the onset of mortality and fertility declines.

Nor is the short-run relationship between mortality and industrialization obvious. In the 19th century before the germ theory of disease had led to advances in sanitation (for example, chlorination of water) and the treatment of food and drink (for example, pasteurization of milk), densely populated cities were unhealthy places. Germs spread as waves of immigrants flocked into metropolitan centres rife with a diverse menu of infections, the immigrants coming from rural isolates too tiny to support the host of infectious diseases with which they were now assailed.

Only in the late 19th century and after did cities become healthy as knowledge of water purification, the importance of proper sewer systems, and flush toilets spread in the West. With the 20th century development of sulpha drugs followed by the chance discovery of penicillin and the mass manufacture of antibiotic drugs, the scale economies in distribution enjoyed by cities came to the fore. Preventing infection through public health and treating infectious cases came at a lower unit cost in dense, congested, jurisdictions that had once been mortality sink holes.

In the remainder of this article our focus will be on the onset of the fertility transition and its connections with industrialization.

The most important project that laid out the empirical groundwork for analysing questions about the onset of the fertility transition is the European Fertility Project that carried out at the Office of Population Research at Princeton University during the 1960s and 1970s under the direction of A. Coale. Coale and his colleagues wanted to construct measures of fertility and its components –reproduction within marriage, proportion married, the incidence of reproduction outside of marriage (illegitimate fertility) – that could be generated from a relatively small amount of data, data that they could secure for every province throughout 19th-century western Europe and Europe.

The European Fertility Project hit upon the ingenious procedure of comparing the actual fertility experiences of the populations they were studying with the fertility experience of the Hutterites who thereby entered the historical demography literature as a much utilized standard. Why use Hutterite reproduction as a standard? Hutterite women in the period between the world wars married at very young ages and had as many children as possible. The Hutterite sect took very seriously the Biblical injunction to 'be fruitful and multiply'. Moreover, the Hutterites who settled in the great plains of the United States and the prairies of Canada lived on large farms and had a strong demand for child labour. A typical Hutterite woman had a total fertility rate (the sum of the age specific birth rates, an approximation to the total number of children she would give birth to over her reproductive life) of more than 12. Using the Hutterite standard allows us to estimate the degree to which a population falls short of its maximal reproductive potential.

The Hutterite indices generated by the European Fertility Project measure the relative level of marital fertility, illegitimate fertility, proportion married and overall fertility for any jurisdiction that has counts of births classified by legitimacy status and counts of population classified by gender and marital status in the five-year age groups. The idea is to use figures on women and married women in the five-year age groups in a given population of interest to the researcher to compute the level of fertility and marital fertility that would occur if these women reproduced at the rate of Hutterite women in the cohorts of the 1920s and 1930s. The age specific rates (for five-year age groups) at which Hutterite wives reproduced are known and these are used in conjunction with the actual data on population and births to compute the Hutterite indices.

In assessing why populations fall below maximal reproductive potential it is important to separate out the impact of low proportions married from the impact of sharply diminished reproduction within marriage. The Hutterite index for marital fertility (I_g) for a given population is the ratio of the legitimate births occurring in that population to the number that would occur if the women reproduced at the rate of the Hutterites. The Hutterite index for proportion married (I_m) is the ratio of married women weighted by the Hutterite fertility schedule – take the number of married women in each age group and multiply this number by the corresponding level of Hutterite fertility for the age group, thereby giving heaviest weight to the most reproductive ages – divided by the total number of women weighted by the Hutterite schedule. The Hutterite index for illegitimate fertility (I_h) is the ratio of the number of illegitimate births to those that would occur had the unmarried women reproduced as the Hutterite women had reproduced. The overall Hutterite index of fertility (I_f) is the ratio of total births occurring in a population to those that would have occurred had the women been as fruitful as the Hutterite women. The last measure offers an overall summary for fertility.

Not surprisingly, the Hutterite indices for illegitimate fertility – in 19th-century Europe and Asia – tend to be low, typically falling below a value of 0.10. By contrast the Hutterite index for marital fertility in most 19th-century western European provinces tended to be fairly high, around 0.80 in many cases.

One of the convenient properties of the Hutterite indices is their multiplicative property. If the index of illegitimate fertility is zero (typically it is close to zero), then the Hutterite index for overall fertility is the product of the Hutterite indices for marital fertility and proportion married, namely $I_f = I_g * I_m$.

To see why constructing these indices yields useful information about the nature of pre-transition fertility and the dating of the fertility transition, consider the following. In 19th-century western Europe prior to the sustained decline in marital fertility (the European Fertility Project defines the onset of the fertility transition as a drop in I_g of ten per cent initiating irreversible decline, no subsequent return to the pre-decline level occurring), a typical value for the Hutterite index for proportion married was around 0.5, the corresponding Hutterite index for marital fertility being between 0.8 and 0.9. Multiplying the two gives a value of between 0.4 and 0.45, meaning that in western Europe women reproduced far less than did the Hutterites, not because of what they did within marriage, but rather because they were not marrying very young or, in some cases, at all. By contrast in pre-decline Japan, China and Korea, the levels of I_m were usually between 0.8 and 0.9, women marrying very early and almost universally. However, levels of reproduction within marriage I_g were quite low in pre-transition Asia, around 0.5 in many cases. Again, taking the product, we get a range for Ig between 0.4 and 0.45.

So, in both pre-transition Asia and pre-transition western Europe, overall levels of reproduction were modest, but for different reasons in the two regions. In Europe the key was late marriage and low proportions marrying. This was something Malthus approved of, believing that the path of demographic virtue lay in late marriage and abstinence outside of marriage. In Asia the key was relatively low levels of reproduction within marriage, something Malthus was less enthusiastic about. Indeed, he probably would have labelled it vice.

To return to the question of what the European Fertility Project's findings tell us about the relationship between industrialization and the fertility transition, some of the most striking findings of the project need stating. First, France was the region in western Europe enjoying the earliest decline in marital fertility, its irreversible fall beginning in the early 19th century, occurring prior to sustained industrialization there. Second, the irreversible decline in English marital fertility did not occur until the 1870s, a full century after the Industrial Revolution began there. Third, language and culture seem to have been important in shaping the spread of marital fertility decline. For instance in Belgium, language difference separates early-decline provinces from late-decline provinces. For these reasons the European

Fertility Project concluded that stopping behaviour within marriage –having a specific number of offspring, then ceasing having more children altogether – was an innovation. As Coale and Watkins (1986) demonstrate, the consensus opinion in the European Fertility Project was that the innovation of regulating reproduction diffused through contact between individual households, this diffusion channelled through and within distinctive cultural groups.

In short, there is no simple story for western Europe involving the short-run relationship between the onset of industrialization and the onset of marital fertility decline. It is apparent that both are important to modernization. But the interaction of the two is certainly complex.

When we turn to Asia, the complexity of the relationship is even more evident. For instance, in Japan, China, Korea and Asiatic Russia marital fertility appears to have risen before it began its irreversible decline in the 20th century. Mosk (1983) offers one hypothesis about the rise in fertility in Japan that is consistent with the idea that there is a long-run linkage between industrialization and low marital fertility. His explanation rests on the idea that in the short run a rising standard of living may actually induce a rise in marital fertility provided marital fertility has been suppressed through infanticide and sexual taboos aimed at lengthening the intervals between live births. In particular, he argues that rural areas that were experiencing land reclamation due to the diffusion of rice seed varieties from the southwest to the north-east spawned new family managed farms, increasing the demand for child labour and easing pressure on parents concerned with finding marriage and/or farming opportunities for their offspring. Additionally, improved food consumption affected the length of intervals between live births considered optimal, promoting a rise in I_g between the 1880s and the 1920s. Better-fed households felt less constrained to space their births far apart lest they fall short of the nutritional resources required to guarantee survival for all of their youngsters. To these arguments one can add the fact that the opening of the country to international trade in the late 19th century created a strong export market for silk, which was produced by family labour especially in the north-east and the Japanese Alps.

In sum, the literature dealing with the overlap of industrialization and the onset of the demographic transition suggests that the interaction of the two secular transformations crucial to defining modernity is complex and intriguing. As with the issues involving the Malthusian economy, much is known. The general contours of the issues involved are clear enough. But, as with so many other things, the devil is in the details. At the detailed level, it is clear what we do not know is as important as what we know. In this sense historical demography has opened up as many questions for future research as it has provided answers to questions thrown up by previous generations of scholars.

See Also

- ▶ Demographic Transition
- ▶ Industrial Revolution
- ▶ Malthus, Thomas Robert (1766–1834)
- ▶ Malthusian Economy

Bibliography

Bengtsson, T., C. Campbell, J. Lee, et al. 2004. *Life under pressure: Mortality and living standards in Europe and Asia, 1700–1900*. Cambridge, MA: MIT Press.

Coale, A., and S. Watkins, eds. 1986. *The decline of fertility in Europe: The revised proceedings of a conference on the European fertility project*. Princeton: Princeton University Press.

Hayami, A. 1997. *The historical demography of premodern Japan*. Tokyo: University of Tokyo Press.

Hollingsworth, T. 1969. *Historical demography*. Ithaca: Cornell University Press.

Lee, R. 1980. An historical perspective on economic aspects of the population explosion: The case of preindustrial England. In *Population and economic change in developing countries*, ed. R. Easterlin. Chicago: University of Chicago Press.

Lee, R. 1985. Population homeostasis and English demographic history. *Journal of Interdisciplinary History* 25: 635–660.

Lee, R. 1987. Population dynamics of humans and other animals. *Demography* 24: 443–465.

Lee, J., and C. Campbell. 1997. *Fate and fortune in rural China: Social organization and population behavior in Liaoning 1774–1873*. Cambridge, MA: Harvard University Press.

Mosk, C. 1983. *Patriarchy and fertility: Japan and Sweden, 1880–1960*. New York: Academic Press.

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

Historical Economics, British

J. Maloney

Keywords

Ashley, W.; British historical economics; Cliffe Leslie, T. E.; Cunningham, W.; Economic man; Evolutionary economics; Foxwell, H. S.; German Historical School; Hewins, W. A. S.; Historicism; Hobson, J. A.; Holism; Individualism versus holism; Ingram, J. K.; Jones, R.; Marginalism; Marshall, A.; Protection; Rogers, J. E. T.; Toynbee, A.

JEL Classifications

B1

A group of economists whose heyday was from 1875 to 1890 and whose major figures were John Kells Ingram (1823–1907), James E. Thorold Rogers (1822–1890), T.E. Cliffe Leslie (1827–1882), William Cunningham (1849–1919), Arnold Toynbee (1852–1883), William Ashley (1860–1927) and W.A.S. Hewins (1865–1931). H.S. Foxwell (1849–1936) was sympathetic to their approach but outside the group's mainstream. All were united by an inductive approach to economics, a determination to stress that no economic theory or policy could be appropriate to all times and places, and a conviction that classical and neoclassical economics alike were already too abstract to give state or citizen much practical help, and were getting worse.

The movement's most important forerunner was Richard Jones (1790–1855), whose criticisms of Ricardian economics – both for its hyper-deductive character and its pretensions to universality – enjoyed intelligent public attention without much persuasive power. Jones offered neither a historically relative political economy to put in Ricardianism's place nor even any substantial contribution to economic history. But, in any case, the time was not right for Jones's ideas to take hold. By the 1870s a number of factors had combined to prepare the ground for a far more influential historical critique of orthodox economics. There was the influence of John Stuart Mill, who in his later years both practised and lent his philosophical authority to a more inductive approach to political economy. Yet when Mill's influence was removed by his death in 1873, silencing the most authoritative voice in economics, the collapse of classical orthodoxy was further accelerated. And of its two main potential heirs, marginalism and historicism, it was the historicists who were more in tune with the general intellectual climate of the time.

As Darwinian ideas were absorbed into social science, the call went up for an evolutionary (and hence relativistic) science of political economy. (No one was to call for it more loudly than Marshall.) The Comtean critique of overspecialization within social science was still near its zenith, and applied with especial force to the increasingly narrow world of neoclassical economics. 'Straight' history was increasingly emphasizing its economic aspects in the work of F.W. Maitland, F. Seebohm and P. Vinogradoff. And, for those who were prepared to listen, Karl Marx was reiterating the potential scope and grandeur of economic dynamics.

The representatives of the English historical school drew on such influences with varying degrees of emphasis. Ingram used his presidency of Section F of the British Association (the social science section) to mount an explicitly Comtean attack on political economy's 'narrowness' in 1878. Ashley painstakingly catalogued the aspects of Marxism with which he was and was not in agreement. The one conditioning factor which, oddly enough, was of limited influence was the work of the German Historical School of economists. English historicists might invoke the authority of their German contemporaries; Ashley and Hewins had important contacts with the later

German Historical School; but it is hard to point to any German historicist as a major formative influence on any English counterpart.

What, then, was the detailed message of the Historical School? (In answering this question we shall be able to throw light on how far it should be regarded as a distinct 'school' at all.) First, as has already been mentioned, they were reacting against the narrow scope of orthodox economics. Thus, Ingram's address of 1878, while accepting the arguments in favour of doing 'one thing at a time', warned that the social sciences were still branches of one subject 'and the relations of the branches may be precisely the most important thing to be kept in view respecting them'. Ingram saw the narrow intellectual vision of orthodox economists as both cause and consequence of their neglect of moral issues, and further argued that once it was accepted that 'the idea of forming a true theory of the economic frame and working of society apart from its other sides is illusory' it necessarily followed that 'the economic structure of society and its mode of development cannot be deductively foreseen but must be ascertained by direct historical investigation' (Ingram 1878).

But should one's methodological stance in fact depend on one's assessment of the appropriate intellectual boundaries of economics? J.A. Hobson was later to argue that the two issues had nothing whatever to do with one another. However, historicists to a man – albeit with different degrees of emphasis – followed Ingram's lead in using their calls for a broader-based discipline to buttress their onslaught on unbalanced deductivism. The link was 'economic man', seen by historicists as an unreal psychological stereotype wholly unable to support the pyramids of deductive logic burdened upon him by Ricardians and Jevonians alike. Whether it was wealth or utility that he was supposed to maximize, he turned out very much the same, 'an abstraction confounding a great variety of different and heterogeneous motives which have been mistaken for a single homogeneous force' (Cliffe Leslie 1879). Other Ricardian propositions which, in Leslie's view, contradicted actual experience included the quantity theory of money and the contention that competition operated so as to equalize rates of profit across the economy.

Leslie's suggestion that the whole edifice of Ricardian economics be levelled to the ground, prior to economists making a fresh and cautious start, marked the high point of historicist iconoclasm. There were a number of different stopping-places (most of them inhabited by Ashley at one time or another) along the road from orthodoxy to this extreme point. Yet the historicists hang together as a school because of their common emphasis on factual and statistical thoroughness, on the relativity of economic doctrines, and on entering unfamiliar territory with an open mind and doing painstaking research before allowing the first tentative inductive generalizations to filter through. The most orthodox of the school, Thorold Rogers, made the most impressive statistical contribution with his *History of Agriculture and Prices in England* (1866) which, among other objectives, sought to marshal the figures needed to refute Ricardian rent and wage theory. Ashley's verdict, however, that Rogers' practice of merely illustrating his preconceived opinions with historical material was alien to a genuine historical method has been endorsed by modern commentators.

It would be wrong to conclude from the above that the Historical School was hostile to deduction as such. 'Deduction', said Ingram, 'is a legitimate process when it sets out not from *a priori* assumptions, but from proved generalisations'. The historicist position, in effect, was that one had to ascertain by factual investigation exactly how amenable to deductive analysis different economic phenomena actually were. That the calculating maximizing spirit (where it existed) was amenable to Ricardian treatment was conceded on all sides. This point had been heavily stressed by Walter Bagehot (in his centenary essay on *The Wealth of Nations)* in the hope of rendering orthodox economics more plausible by demarcating its boundaries as those of the modern commercial world. Ashley's inaugural lecture at Harvard in 1893 endorsed this point; Cunningham's *Modern Civilisation in Some of its Economic Aspects* (1896) asserted that deductive analysis was coming into its own because 'business of a modern type is being extended over a larger and larger area'. That this last tendency was – on balance – welcomed by Ashley and regretted by Cunningham may help explain the difference in their attitudes to Marshallian economics. Ashley (who was to become professor of commerce at Birmingham in 1901) shared Marshall's enthusiasm for most of what the modern businessman represented. In Cunningham, by contrast, distaste for the modern world and nostalgia for the Middle Ages predominated. But personal temperament counted for just as much in explaining the contrast between Ashley's relatively placatory attitude to Marshall and Cunningham's violently hostile one.

Marshall's inaugural lecture at Cambridge in 1885 had met, head-on, the historicist assertion that the forces of custom and habit in economic life were strong enough to make orthodox economics, with its basic postulate of maximization, widely redundant. Marshall predicted that 'economic science' would soon be even more successful than it was already in 'break[ing] up and explain[ing] economic customs'; asserted that statements that this or that economic arrangement was due to custom were little more than confessions of ignorance of true causes; and entrusted economic analysis with the illumination of such ignorance – the demonstration, for example, that 'rents seldom diverge much for a long time from their Ricardian level in the East' (Marshall 1885). Cunningham, while regarding the whole lecture as a personal and public affront, fastened especially onto this last point, telling the British Association (1889) that 'Professor Marshall, instead of accepting the description of mediaeval or Indian economic forms as they actually occur, sets himself to show that the accounts of them can be so arranged and stated as to afford illustrations of Ricardo's law of rent.' Marshall's *Principles of Economics,* published the following year, opened with a long historical introduction which Ashley saw as a conciliatory gesture and Cunningham as a further provocation. (Today it reads as neither.) In 'The Perversion of Economic History' *(Economic Journal,* September 1892), Cunningham joyously rebuked what he saw as Marshall's hasty and amateurish style of historiography. It would all have read more convincingly if Cunningham had refrained from grotesquely out-of-context quotation, even at one point inserting a rogue word into Marshall's text to make it sound marginally more implausible.

Marshall's reply to Cunningham's criticisms (it took Cunningham three years and seven polemics to induce it) was seen in most quarters as the final statement in the dispute (if only because the *Economic Journal* refused Cunningham the space for a counter-riposte.) Ashley, in his Harvard inaugural the following year, praised the historical chapters in the *Principles* and claimed that 'to most of us the recent exchange of hostilities between two distinguished English economists has seemed almost an anachronism'.

The methodological debate, then, subsided after the early 1890s. But the protectionist controversy which began when Joseph Chamberlain disavowed free trade in 1903 saw survivors of the old historicists grouping reconstituted for a new battle. The episode is best approached via a general look at historicist attitudes to policy questions.

It is no coincidence that the entire Historical School, regardless of whether as individuals they were of the 'left' or the 'right', favoured an acceleration of the existing trend towards increased state intervention in the economy. Irish social reform, the recognition and legal protection of the trades unions, and the conditions of industrial and agricultural workers were all seen as urgent areas of responsibility for the state. The general view was well summarized by Foxwell (1885):

> We have been suffering for a century from an acute outbreak of individualism unchecked by the old restraints and invested with almost a religious sanction by a certain soul-less school of writers. The narrowest selfishness has been recommended as public virtue.

Ingram praised the German Historical School for upholding the power of the state as 'the organ of the nation for all ends which cannot be adequately effected by voluntary individual effort'. Cunningham's *Politics and Economics* (1885) introduced his readers to 'National Husbandry', Cunningham's scheme for an economic policy holistic in its inspiration and nationalistic in its objectives: 'the duty we owe to posterity [is] to make the future of our nation as great and noble as lies within our power'.

The link between holism (refusal to isolate the individual as a unit of analysis) and historical relativism was an irreproachably logical one: only if an individual can be isolated from his social context can a theory involving him be isolated from time and place. And Cunningham for one kept his readers' eyes firmly on the fact that policy recommendations were as historically relative as economic principles, even suggesting at one point that the fact that a measure had worked well in very different circumstances was a consideration *against* proposing it here and now. Such pragmatism characterized much of the protectionist campaign. If free-trading economists were to be charged with inflexible dogmatism, intellectual arrogance and subservience to abstractions, it was essential that no such taint could be thought to cling to the protectionist cause. Ashley, indeed, never went beyond recommending temporary and selective tariffs for purposes of retaliation, and stressed that 'with England as she has been for some centuries the notion that imports are paid for by money which might otherwise be spent at home is the crudest of popular fallacies' . Cunningham – eventually – did arrive at a more thoroughly protectionist stance than this, but it took him until 1910 to do so. And by 1910 the steam was running out of the protectionist campaign anyway, at least as far as the Historical School was concerned. Ashley's administrative responsibilities at Birmingham and Hewins's parliamentary ones virtually terminated their contributions to serious economic debate; Cunningham turned his attention to the relations between Christianity, political practice and social science. The Historical School's achievements were complete by 1914.

How significant were they? Today their part in the foundation of economic history as a subject in its own right is more obvious than their contribution to economics. Their lack of facility with marginal analysis – no historicist tried to master the neoclassical 'paradigm' and it must be doubted whether most of them would have been able to handle it even if they had tried – relegated them to outsiders' roles once the dominance of neoclassicism was secured. Could they have prevented this dominance? The answer depends on whether one thinks that the inductive, historically based economics which they demanded but ostentatiously failed to supply could ever have been a feasible project. As it was, their lack of solid achievement inevitably weakened their position even as critics. Yet they forced both Marshall and his disciples to change both their thoughts and their presentation of these thoughts in a number of ways. Economic concepts were more carefully defined, and the bounds of their applicability more precisely demarcated. Policy recommendation became more cautious and less likely to be accompanied by exaggerated statements of the contributions of pure theory. The modern economist, said L.L. Price (1906),

> evinces a readiness to recognise without reserve those qualifications of subtle delicate theory which a comparison with rough, unyielding facts must necessarily require. This reasonable attitude is largely due to the abiding influence of the vigorous controversy in which Cliffe Leslie bore a leading part.

See Also

► Historical School, German

Bibliography

Ashley, W.J. 1903. *The tariff problem*. London: P.S. King & Son.

Coats, A.W. 1954. The historicist reaction in English political economy, 1870–90. *Economica* 21 (May): 143–153.

Cliffe Leslie, T.E. 1879. *Essays in political and moral philosophy*. Dublin: Hodges, Figgis & Co..

Cunningham, W. 1885. *Politics and economics*. London: Kegan Paul, Trench & Co..

Cunningham, W. 1896. *Modern civilisation in some of its economic aspects*. London: Methuen.

Foxwell, H.S. 1885. What is political economy? *The Eagle* No. 79.

Ingram, J.K. 1878. *The present position and prospects of political economy.* Dublin/London: Longmans. Reprinted in *Essays in economic method,* ed. R.L. Smith. London: Duckworth, 1962.

Koot, G. 1980. English historical economics and the emergence of economic history in England. *History of Political Economy* 12: 174–205.

Marshall, A. 1885. *The present position of economics*. London: Macmillan.
Rogers, J.E.T. 1866–1902. *A history of agriculture and prices in England*. Oxford: Clarendon Press.
Semmel, B. 1960. *Imperialism and social reform*. London: George Allen & Unwin.

Historical School, German

Heath Pearson

Abstract

The German Historical School was an influential heterodoxy in 19th-century political economy. It diverged from the classical school crucially in its scepticism that universal laws of social behaviour could be established. Its members were also more interventionist, tending to favour protection, regulation colonization and the welfare state, though by no means unanimously on every point. In line with their relativity, they accepted that their policy recommendations, too, were historically contingent. Their influence among economists was greater in developing countries than in western Europe, but it has everywhere had a lasting impact on allied branches of social science such as sociology.

Keywords

American Academy of Political and Social Science; American Economic Association; Brentano, L.; Bücher, K.; Chayanov, A.; Commons, J.; Durkheim, E.; German Historical School; Germany, economics in (20th century); Gide, C.; Hildebrand, B.; India, economics in; Italy, economics in; Japan, economics in; Jones, R.; Keynes, J.M.; Knapp, G.; Knies, K.; Lamprecht, K.; List, F.; Lubbock, J.; Maine, H.; Malinowski, B.; Marshall, A.; Methodenstreit; Mitchell, W.; Polanyi, K.; Relativity; Roscher, W.; Schmoller, G.; Schumpeter, J.; Sombart, W.; Spiethoff, A.; Veblen, T.; Verein für Socialpolitik (Association for Social Policy); von Inama-Sternegg, K.; Wagner, A.; Weber, M

JEL Classifications

B1

The German Historical School has a fair claim to be the most thoroughgoing and influential heterodoxy in 19th-century political economy. Scholars conventionally date its origins to 1843, with the publication of Wilhelm Roscher's *Outline of Lectures on Political Economy, according to the Historical Method.* Again by convention, the school is divided chronologically into three generations. The 'older' generation included Roscher (1817–94), Bruno Hildebrand (1812–78), and Karl Knies (1821–98). It was succeeded by a 'younger' generation, led by imperial Germany's most prominent economist, Gustav Schmoller (1838–1917), and including Lujo Brentano (1844–1931), G.F. Knapp (1842–1926), K.T. von Inama- Sternegg (1843–1908), and Karl Bücher (1847–1930). A 'youngest' generation included Werner Sombart (1863–1941) and Arthur Spiethoff (1873–1957). All were professors of political economy or of *Staatswissenschaft* ('state science'), and were widely known outside academic circles. A full account of the German Historical School would include many lesser-known figures, as well as several famous scholars who have often been associated with its agenda: Friedrich List (1789–1846), Adolph Wagner (1835–1917), Karl Lamprecht (1856–1915), and Max Weber (1864–1920), among others. There is no consensus date for the school's demise, but most would agree that by 1918 it was losing momentum, and that by 1945 it was a spent force.

In so far as they are remembered for belonging to the school, history has assigned to these economists the role of dramatic foil vis-à-vis classical political economy. Where the classicals were cosmopolitan children of the Enlightenment, the historical economists are remembered as romantics, idealists, nationalists; where the former were motivated to understand the nature and prospects of the commercial society taking shape around them, the latter were oriented to the economic past and its evolution towards the present; where the former were Newtonian in their aspirations for a master theory of the market order, the latter were satisfied to explore the peculiarities of specific situations; where the former offered a robust defence of private enterprise, the latter were just as robust in their vindication of state intervention; and where the classicals proved endlessly adaptable as circumstances varied, the German Historical School was sterile, a creature of one time and place. There is something to be said for each of these contrasts, but each of them can be – and has been – overdrawn, and the trend of recent scholarship has been to mitigate them.

Method

The Historical School partook of the ethic of professional historiography, seeking not to ransack the past but rather to understand it on its own terms. As history was an integral part of the *Staatswissenschaft* curriculum through which most German economists passed, it could hardly have been otherwise. But it is also fair to say that in this curriculum history was yoked, not to say subordinated, to the established discipline of *Statistik,* which had long meant the comparative study of social phenomena for purposes of effective statecraft. This more instrumentalist approach to historical inquiry is clearly in evidence in the works of the Historical School; it accords too with the participation of many members (notably Knies, Hildebrand, Inama-Sternegg, Bücher, Knapp, Brentano, Wagner, and Spiethoff) in the development and use of statistics in its more strictly modern sense, and the interest of many others (notably Roscher, Schmoller, Inama-Sternegg, and especially Bücher) in contemporary ethnography. In this sense they bear more than a passing resemblance to an Enlightenment polymath named Adam Smith, whom Roscher (1843, p. 150) named among the forefathers of historical economics.

But how would those copious data be used? A widespread view, born especially from the famous *Methodenstreit* between Schmoller and Carl Menger in the early 1880s, is that the historical economists took their scientific brief to include description, collation, and not much else; valid theoretical knowledge would emerge, if at all, only in the fullness of time and of its own accord. It is indeed true that in the heat of the dispute Schmoller made some ill-advised statements to this effect, and it is true also that he and his colleagues consistently denounced what they saw as the deductive excesses of Ricardian theory. However, in general they were far from denying the validity of deductive inference in principle (see Schmoller 1901–04, pp. 108–11); in practice they did engage in generalization and in theoretical speculation, sometimes so sweepingly as to make a classical economist blush. Prominent among the grander visions was their penchant for evolutionary ‘stage theories’ of economic development. It is these stage theories which have attracted the taint of holism, teleology, and crypto-Hegelian idealism. Once again, these charges are less than outrageous (see Weber 1902–05) but significantly overstated. Turgot, Smith and Marx made similar efforts to reduce such broad historical processes to patterns of individual behaviour, which in turn are explicable in terms of the overall context. Even as it evolves, Roscher wrote in the introduction to his influential textbook, the economy remains ‘a natural product of the faculties and drives which make the human being human’ (1854, s. 14).

Perhaps the best single term to distinguish their brand of science is ‘relativity’. Unlike Newton and his admirers, who envisioned law-like relationships that were invariant as to time or place, the historical economists thought this a hopeless task for the human sciences. The theories they aspired to had fewer constants, more variables (psyche, environment, institutions, and so on) and in some versions a large error term – but they were still recognizable as theories. According to Roscher’s formulation, the classicals had postulated rules, to which recent critics had pointed out myriad exceptions. ‘Now it would be above all necessary’, he went on, ‘to broaden the rules themselves to the point where those exceptions are incorporated’ (quoted in Eisermann 1956, p. 150). Or, as Schmoller himself put it near the end of his career, German economists of his persuasion had achieved progress in economic theory the way Smith had, ‘by placing man and society at its center; but they did not thereby exclude the methods of natural science, or general concepts, or regularities. They did not claim that all the phenomena of economic life are individual and unique' (1911, p. 434).

Policy

As regards policy recommendations, it is clear that the modal opinions of German historical economists were distinctly more interventionist than the Anglo-French norm. The renowned Verein für Socialpolitik (Association for Social Policy, arguably the world's first economic think tank) was founded in 1872 primarily by members of the school, with a mission that stood as a plain rebuke to the principle of laissez-faire. It is for this reason that the epithet *Kathedersozialisten* ('socialists of the lectern') was rather indiscriminately applied to them in their own day, and for this reason, too, that historians have come to view government activism as just as intrinsic to historical economics as any methodological precept. Paradigmatic in this view is Schmoller's belief in a 'social monarchy' and its capacity to reconcile the goals of private property, national development, and distributive justice through a programme of protection, regulation, and colonization. Once again, however, this generalization must be handled gingerly. It understates the diversity of political opinion within historical economics; specifically, it ill serves those economists who called for participatory government (Brentano, Bücher), who tended towards state socialism (Wagner, young Sombart), who opposed Bismarck's tariffs (Brentano, Bücher, Weber), and who doubted the capacity of regulation to improve upon market outcomes in general (Roscher, Hildebrand, Weber). It also ignores their essentially relativistic outlook: like List before them, who had promoted protection as the policy for *his* time but not for *all* time, the German historical economists – not excluding Schmoller himself – recognized the historical contingency of their specific recommendations.

Influence

It is also the case that the German Historical School was less parochial than has been suggested. It is true that the influence of German economists was conspicuously weak in francophone Europe through most of the 19th century, despite an early translation of Roscher's *Principles,* and despite the efforts of the Belgian economist Emile de Laveleye. This situation began to improve after about 1880, however, with the creation of the first chairs of political economy in the law faculties of the French universities. Since the new professors were perforce trained in law, and since French jurisprudence had already begun to fall under the sway of German historicism, they were better disposed to the historical economists than their predecessors (Gide 1908). Charles Gide's critical appreciation of the German economists was characteristic of this younger generation, as was their reception by Paul Cauwès, François Simiand and Emile Levasseur. Historical economics fared better in the United Kingdom, thanks largely to the indigenous examples of Richard Jones, Henry Maine, John Lubbock, and others. T.E. Cliffe Leslie praised their endeavours, as did W.J. Ashley, William Cunningham, J.S. Nicholson, and W.A.S. Hewins. Alfred Marshall, whose name is not typically associated with historical economics, in fact affirmed that the school's work 'has thrown light on economic theory, has broadened it, has verified, and has corrected it' (1890, p. 74).

Despite these successes, in western Europe the German historical economists remained exotic specimens of a minor genus. Elsewhere they fared better, with their influence waxing in rough proportion with the developmental ambitions of the society in question. The historical school's rise to prominence in Gilded Age America (c.1876–1914), due largely to Germany's pre-eminence as a site for higher education in economics, has been well documented (Dorfman 1955; Herbst 1965; Rodgers 1998). The American Economic Association (AEA) and the American Academy of Political and Social Science were both originally modelled on the Verein für Socialpolitik; all told, 20 of the first 26 presidents of the AEA had studied in Germany. While the leading American 'institutionalists' of the early 20th century did not have first-hand experience in Germany, they can be seen as carrying on the school's agenda: Thorstein Veblen its methodological dissent, W.C. Mitchell its statistical inquiries, J.R. Commons its social reformism. The Italian case offers a fairly close parallel. Young Italian economists were drawn to advanced study in Germany, and while the guardians of orthodoxy could inveigh against the trend of *germanismo economico,* they could not staunch the school's influence among economists such as Luigi Cossa, Vito Cusumano, Giacomo Luzzatti, and Achille Loria (Schiera 1989). Elsewhere the German historical school achieved something close to intellectual hegemony by the turn of the 20th century, for example in Russia (Balabkins 1988; Kingston-Mann 1999; Barnett 2004) and in Finland (Heinonen 2002). Interestingly, its greatest influence relative to other schools was in Asia. In British India, the German dissent from orthodoxy was praised in the work of M.G. Ranade, G.K. Gokhale, R.C. Dutt, and G.S. Iyer, and left its imprint in the field of 'Indian Economics' that they founded. In Meiji Japan, meanwhile, the school established a decisive beachhead at the Imperial University in Tokyo in the early 1880s, thanks to direct German influence and especially to that of German-inspired American professors (Pyle 1974; Sugiyama and Muzuta 1988). A Society for Social Policy was founded there in 1897 on the model of the Verein für Socialpolitik, membership in which soon became an essential qualification for professional economists in that country.

Finally we turn to the question of the German Historical School's influence, or lack thereof, into the later 20th century and beyond. The school had a lasting impact on allied branches of social science. In economic sociology, Emile Durkheim's early works were fairly deeply engaged with historical economics (Steiner 2003); Joseph Schumpeter held Schmoller up as a pioneer in the field (Schumpeter 1926); and Max Weber was so deeply rooted in the school that he himself has occasionally been called a member. In the field of economic anthropology, Bronislaw Malinowski, Karl Polanyi and A.V. Chayanov had all been exposed to this literature in their youth (Kahn 1990). It is only within economics itself that the German Historical School's star waned quickly after 1930. The reasons are no doubt complex and entwined with the drama of German political history; but surely the 'formalist revolution' in economic theory at large – where clarity and elegance gained great popularity, occasionally at the expense of verisimilitude and relevance – played a significant role. The matter is crystallized in J.M. Keynes's obituary for Marshall in 1924, where he characterized the school's work as 'learned but half-muddled'. One pictures Marshall nodding in reluctant agreement, and then asking aloud what more could be asked of true social science. For Keynes's successors, however, that indictment could hardly have been more damning.

See Also

- ▶ American Economic Association
- ▶ Bücher, Karl Wilhelm (1847–1930)
- ▶ Economic Anthropology
- ▶ Economic Sociology
- ▶ Hildebrand, Bruno (1812–1878)
- ▶ Historical Economics, British
- ▶ Institutionalism, Old
- ▶ Knies, Karl Gustav Adolf (1821–1898)
- ▶ List, Friedrich (1789–1846)
- ▶ Methodenstreit
- ▶ Roscher, Wilhelm Georg Friedrich (1817–1894)
- ▶ Schmoller, Gustav von (1838–1917)
- ▶ Sombart, Werner (1863–1941)
- ▶ Spiethoff, Arthur August Kaspar (1873–1957)
- ▶ Weber, Max (1864–1920)

Bibliography

Balabkins, N. 1988. Schmoller in Tsarist Russia. *Journal of Institutional and Theoretical Economics* 144: 581–590.

Barnett, V. 2004. Historical political economy in Russia, 1870–1913. *European Journal of the History of Economic Thought* 11: 231–253.

Dorfman, J. 1955. The role of the German Historical School in American economic thought. *American Economic Review* 45: 17–28.

Eisermann, G. 1956. *Die Grundlagen des Historismus in der deutschen Nationalökonomie*. Tübingen: Enke.

Gide, C. 1908. L'Ecole économique française dans ses rapports avec l'Ecole anglaise et l'Ecole allemande. In *Die Entwicklung der deutschen Volkswirtschaftslehre im neunzehnten Jahrhundert*, ed. S. Altmann et al., vol. 1. Leipzig: Duncker and Humblot.
Heinonen, V. 2002. The influence of the German Historical School in Finnish economic thought around the turn of the century. In *Economic thought and policy in less developed Europe: The nineteenth century*, ed. M. Psalidopoulos and M. Mata. London: Routledge.
Herbst, J. 1965. *The German Historical School in American scholarship: A study in the transfer of culture*. Ithaca: Cornell University Press.
Kahn, J. 1990. Towards a history of the critique of economism: The nineteenth- century German origins of the ethnographer's dilemma. *Man* 25: 230–249.
Kingston-Mann, E. 1999. *In search of the true west: Culture, economics, and problems of Russian development*. Princeton: Princeton University Press.
Marshall, A. 1890. *Principles of economics*. London: Macmillan.
Pyle, K. 1974. Advantages of followership: German economics and Japanese bureaucrats, 1890–1925. *Journal of Japanese Studies* 1: 127–164.
Rodgers, D. 1998. *Atlantic crossings: Social politics in a progressive age*. Belknap: Cambridge, MA.
Roscher, W. 1843. *Grundriß zu Vorlesungen uber die Staatswirthschaft, nach geschichtlicher Methode.* [Outline of Lectures on Political Economy, according to the Historical Method.] Göttingen: Dieter.
Roscher, W. 1854. Grundlagen der Nationalökonomie. In *System der Volkswirtschaft*, 24th ed., vol. 1. Stuttgart: Cotta.
Schiera, P., eds. 1989. *Gustav Schmoller e il suo tempo: la nascita delle scienze sociali in Germania e in Italia*. Milan: Il Mulino.
Schmoller, G. 1901–04. *Grundriß der allgemeinen Volkswirtschaftslehre.* 2 vols. Leipzig: Duncker and Humblot.
Schmoller, G. 1911. Volkswirtschaft, Volkswirtschaftslehre und -methode. In *Handwörterbuch der Staatswissenschaften*, 3rd ed., ed. J. Conrad et al., vol. 8. Jena: Fischer.
Schumpeter, J. 1926. Gustav v. Schmoller und die Probleme von heute. *Schmollers Jahrbuch für Gesetzgebung, Verwaltung und Volkswirtschaft* 50: 337–388.
Steiner, P. 2003. Durkheim 's sociology, Simiand's positive political economy and the German Historical School. *European Journal of the History of Economic Thought* 10: 249–278.
Sugiyama, C., and H. Muzuta, eds. 1988. *Enlightenment and beyond: Political economy comes to Japan*. Tokyo: University of Tokyo Press.
Weber, M. 1902–05 [1975]. *Roscher and Knies: The logical problems of historical economics*. Trans. G. Oaks. New York: Free Press.

History and Comparative Development

Louis Putterman

Abstract

What role do historical factors play in explaining the large differences in level of economic development among different countries and world regions today? Our answer begins with the observation that cross-country differences in level of economic development, as well as in rate of economic growth since the end of the European colonial era, are strongly correlated with the average levels of technological and political advance in the places the current population's ancestors lived on the eve of that era and earlier. We explore recent literature in economics that addresses the evidence for and the causes of this correlation. The interplay between geography and human capital (broadly understood as including culture, norms, and institutional capability) has a central part in our discussion.

Keywords

Agricultural revolution; Ancestry; Colonial era; Colonies; Industrial revolution; Less developed countries; Linguistic distance; Migration; neolithic revolution

JEL Classifications

N5; O10; Z1

Economists have increasingly sought to understand what accounts for long-standing differences in the levels of economic development among world regions, and why some poor countries have achieved more economic growth than others in recent decades, by reference to relatively exogenous and persistent factors including geography, institutions, culture and early history.

Frequently considered geographic factors have included the advantages of a coastal location and the disadvantage of being landlocked (from the standpoint of participating in low-cost international trade) and the disadvantages of a tropical climate (from the standpoint of pests, disease and soil structure; e.g. Gallup et al. 1999). A different set of geographic factors capturing considerable attention in recent work are those that helped to determine the locations of and degrees of interaction between early agrarian civilisations. The implications of these initially geography-based differences for differences in comparative development, especially those in the non-European world from the 15th century onwards, receive considerable attention.

In the early 2000s, Acemoglu et al. (2001, 2002) argued for an approach emphasising formal institutions and property rights as an alternative to the geography-centred approaches. A number of economists – for example, Guiso et al. (2006) and Nunn (2014) – have offered evidence that culture, too, is a fundamental long-run determinant of economic outcomes.

This article discusses the roles of history, geography, culture and institutions in shaping differences in economic development around the globe. It adopts a long-term historical approach in which differences in the starting times and diffusion of major technological traditions are attributed to geographic factors, including the suitability of a large swath of Eurasia for a common set of domesticates, the absence of such suitability in sub-Saharan Africa, and the cutting off of Oceania and the Americas from the techniques and ideas shared by Eurasian societies due to limited navigation during the centuries leading up to 1500 CE. The long-term persistence of technological advantages, albeit with greater dynamism being exhibited at times by well-connected peripheries, will be argued to have been a rule of economic history. Culture will make its appearance in the approaches that will be focused on here in the form of broad human capabilities and orientations, and as a feature contributing to differences in likelihood of trade relationships and diffusion of ideas. Institutions will be identified as one of culture's important manifestations.

The Agricultural and Industrial Revolutions

The long-run view considered here emphasises that the transitions from foraging to agriculture and from there to industrial societies are both recent relative to the more than 100,000 years of anatomically modern human existence and the 60,000 or so years of modern humans' spread beyond Africa. It also emphasises the fact that both the agricultural and the industrial revolutions diffused from their points of origin in non-instantaneous and geographically uneven fashions.

It took about five millennia for the population growth and improvements in farming, animal husbandry, construction and metallurgical techniques that followed the first agricultural revolution in the Near East to give rise to sizeable cities, states, and record-keeping bureaucracies in the third millennium BCE. The agricultural techniques first developed in the region also took millennia to reach all of Europe and southwest Asia. Meanwhile, another agrarian tradition sprang up at least partly from indigenous roots in what is now north central China, gradually spreading to the south and to what are today Korea and Japan.

Critically, the time lapse between the onsets of Near Eastern and East Asian agricultures was in the neighbourhood of a single millennium, and geographic contiguity (Europe, the Near East and East Asia sharing a common land mass) permitted the technological developments associated with these core areas to ultimately cross-fertilise one another and form a shared body of technical knowledge. In contrast, the agricultural traditions that independently emerged in Mesoamerica and the Andes began significantly later and knew no meaningful contact with outside civilisations until the 15th century. When the American civilisations were (at best) in their infancies, Assyria was already a regional superpower on what (given the available communication technologies) may as well have been a different planet. The American cradles of agriculture would themselves give rise to populous civilisations and states, and their domesticates would achieve immense global importance after the opening of global contact

(Nunn and Qian 2010), but they failed to become technological matches for the successors of Assyria before that contact was made. In the early 16th century, conquistadors from an Old World civilisation, by then millennia their senior, were able to vanquish the two New World civilisations, weakened by diseases to which they lacked resistance, with mere hundreds of men, war horses and small arms.

Compared to the spread of agricultural ways of life, the spread of industry has been rapid. Whereas agriculture took ten millennia to travel from the Near East to Australia (with a regional agriculture even failing to reach Australia from neighbouring New Guinea), at least some industrial factories, office buildings, automobiles and electrical power had reached every country in the world within three *centuries* of the onset of the Industrial Revolution. Technologies of recent years, like the digital camera and the mobile phone, have achieved their initial geographic spread still more rapidly.

Yet the spread of the industrial way of life was not instantaneous. Even today, the capacity to make and use sophisticated equipment, as opposed to simple possession of consumer gadgets, fails to spread to all parts of the world overnight. If we can explain the unevenness in the spread of technologies associated with the two epoch-making revolutions, we may go a long way towards explaining the economic differences between the world's societies today.

What accounts for where and when these revolutions occurred and why they spread to or were taken up more rapidly by some societies than by others? A plausible answer can be built largely out of geographic facts, which from the standpoint of our subject matter are clearly exogenous. These facts include the shapes, terrains and broad climatic profiles of the continents, and the fact that humans evolved in Africa and that the first 50,000 years of their dispersal depended on overland journeys and on sea voyages of relatively short spans only.

Exactly why agriculture arose when it did is a question to which economists (for example Dow et al. 2009) have recently made significant contributions. With respect to sites, Hibbs and Olsson (2004; see also Olsson and Hibbs 2005) find support for the factors emphasised by Diamond (1997), such as the wild habitats of large-seeded grasses and of the precursors of domesticated animals. Regardless of the exact reasons for the timing and locations of the earliest domestications, what is central for our purposes is that the first agricultural revolution led to larger and ultimately more specialised populations and that this in turn facilitated innovation in other areas, including metallurgy, writing and mathematics. Equally important is that the suite of domesticated grains, legumes, ruminants and fowl and the myriad of other technologies following after them faced few geographic barriers to diffusing – however gradually by today's standards – throughout temperate and sub-tropical Europe and Asia and also into those parts of Africa lying along the Mediterranean and the Nile river valley. Coined money, the use of horses and later camels in transport, the plough and later technologies including printing, gunpowder, paper and the compass, diffused from west to east or east to west across a swath of civilisations that included Europe, the Ottoman Empire, Persia, Mughal India, China and neighbouring kingdoms like Vietnam, Korea and Japan.

In contrast, before the 15th century European expansion began, the technologies just mentioned were unknown to the descendants of those who had reached the Americas millennia earlier, or those reaching much of Oceania more recently. The non-grain-based agricultural societies of highland New Guinea, whose agricultural tradition may long predate those of the Americas, was equally isolated from developments in Eurasia, and failed to give rise to a civilisation even on the scale of those in the Americas. Home-grown agricultural traditions also developed in West Africa and spread to much of that continent over roughly the same period that the American agrarian societies were developing, but while productive enough to displace foragers from many of their remaining African habitats, African farming was mainly horticultural and gave rise to only comparatively small states that never incorporated the majority of the continent.

The earlier start, broader expanse and greater contiguity of Eurasian lands of similar climate appear largely sufficient to explain why its civilisations were far more technologically advanced than others in the 15th century. And this in turn, or at least with the help of a view of human nature that accommodates the presence of selfish alongside benevolent motives, appears adequate to explain why, when Eurasian shipbuilding and navigation techniques finally made regular transport across wider expanses of ocean more feasible, people from a Eurasian core society came to exercise power over natives of the Americas, sub-Saharan Africa and Oceania. Who was to conquer whom in the resulting encounters was rendered that much less in doubt by the fact that Eurasians had been building up resistance to a range of diseases capable of felling non-Eurasians in large numbers. In addition, the defences of most non-Eurasian peoples were weakened by their lack not only of iron, steel and gunpowder, but also of the horse, with its considerable military importance until the early 20th century.

Not all are satisfied with this explanatory framework, its application to Africa being especially debated. Acemoglu and Robinson (2012) point out that there was considerable contact between coastal east Africa and ocean-going traders from Arabia long before 1500, and that North African traders plied trade-routes across the Sahara in large caravans. Contacts were sufficiently regular from the 7th and 8th centuries CE that the Islamic world, which had been a transmitter of much thenadvanced knowledge to early modern Europe, had offshoots well south of the Sahara. These contacts not only helped give rise to Sudanic empires, including Mali and Ghana (not to be confused with the modern nations by the same names), but transmitted at least some technological and cultural ideas further south in West Africa, where states such as the Oyo and Benin empires arose by the 15th century. Regular contacts between Egypt and Nubia and between Ethiopia and both Egypt and the Arabian Peninsula existed in earlier millennia. Even if all but the Mediterranean and the southern tip of Africa were climatically (and with respect to disease environment) unsuitable for most core Eurasian crops and domesticated animals, one still needs to explain why other technologies were not picked up more rapidly south of the Sahara.

Useful clues might be found by looking at other cases in which occasional contacts existed, but technologies shared by the core Eurasian civilisations made little impact. Here, it seems relevant that Eurasia itself remained technologically and socially heterogeneous in the 15th century. Parts of Southeast Asia were still sparsely populated by people practicing slash-and-burn agriculture. Manchuria, Mongolia and much of Central Asia were relative technological backwaters peopled by pastoralists whose periodic clashes with both China and Europe are a recurring feature in the history of civilisations up to the early modern era. The deserts of North Africa and Arabia were still home to peoples such as the Bedouin, who maintained ways of life that had changed little over the centuries. In the northernmost parts of Eurasia, reindeer herders had honed a way of life magnificently suited to an extreme climate, but borrowed little from the technology, culture and political forms of the ‘advanced’ Eurasian civilisations. The beginnings of book learning and modern science, and in some cases participation in inter-regional trade, thus tended to leave untouched these parts of the Old World in which the relevant forms of agriculture and animal husbandry were absent, and Africa is not exceptional when viewed in this light.

This suggests that it was not simply contact with other civilisations that led to the spread of the technologies and organisational forms that distinguished civilisations like those of China, India, Persia and Europe from the rest of the world. There also needed to be sufficient similarity in way of life and some already shared elements of mind set for ideas to diffuse. Major differences in way of life, including population density and degree of sedentarism, tended to correlate with differences in political and social structures and, arguably, of mind set, to make transmission of ideas less likely. These way-of-life differences were in turn a function of differences in ecology, climate and soil. In the case of Europe, Anatolia and parts of West, South and

East Asia, ecological conditions were sufficiently similar that similar agrarian ways of life emerged, resulting in both motive and ease of borrowing from one another's mathematics, cultivation and animal breeding techniques, building methods, and transportation technologies. In contrast, Eurasia's arctic and sub-arctic fringe, the Central Asian steppe and most of sub-Saharan Africa were ecologically unable to support similar ways of life. Such lack of overlap in economic adaptation, as much as or more than the physical obstruction posed by the Sahara, may account for lower transmission rates to the sub-Saharan region, as well as to parts of Indonesia and the Philippines that had ongoing contacts with Asia but remained technologically far from the major Eurasian civilisations in the 15th century.

Colonisation and Exceptions to Persistence

As evidence for the argument above that millennia-old differences in early starts and subsequent barriers to diffusion of ideas play a large part in explaining who was colonised and when, beginning in the 15th century, Ertan et al. (2012) estimate regression models with data on 111 countries, accounting for 95% of the world population outside of Europe. They find that each of three measures of development in 1500 CE are individually significant predictors of their outcome variables, with higher pre-modern development making colonisation both less likely to occur and substantially later in time if it does occur. Together, the three indicators of early development explain about a quarter of the observed variation in both colonisation's occurrence and its timing. By inspection, countries colonised by Europeans between the 1467 settlement of Cape Verde by the Portuguese and the 1842 takeover of Hong Kong by the British had far lower levels of experience with agriculture and associated technologies, and of state-level polities, on average, than were common in core Eurasian civilisations in 1500. Late-colonised countries, and even more so those never colonised, look more like the colonising countries, on average, in terms of these indicators. The never-colonised include several countries that had been rough technological peers of Europe in medieval times – e.g. Turkey, Persia, China and Korea.

While the origins of the Eurasian advantage may be largely due to geography, once colonisation was under way differences in command of coercive technologies played at least as large a part as did the pre-existing technological gap itself in explaining why the opening (or expansion) of contacts between Old World civilisations on the one hand and indigenous American, Oceanic and sub-Saharan African societies on the other did not lead to a quick 'upgrading' of technological capacity in the latter. Rather than narrowing, the developmental gaps between Europe and the indigenous people of its colonies effectively widened because the colonial polities restricted the latter's ability to fully learn or put to their own use the new technologies now potentially at their disposal, while the technological capabilities of the colonisers were growing more rapidly than ever.

Our historical narrative thus far has emphasised persistence: Eurasian civilisations got a head start in developing a suite of technologies, which made it likely that one of them, and not a sub-Saharan, Amerindian or Oceanic society, would emerge as dominant once progress in navigation led to global interactions. The persistence of technological advantages from the agricultural revolutions to today has been brought out by studies including Hibbs and Olsson (2004), Olsson and Hibbs (2005), Putterman (2008), Comin et al. (2010) and Putterman and Weil (2010). The first three papers find a link between early agriculture and recent income, while the fourth finds such a link for broader indices of technology adoption in 1000 BCE, 1 CE and 1500 CE, and the last uses both kinds of measure as well as indices of depth of large-scale state experience.

The colonial era is marked by two major disruptions of the overall pattern of persistence, however. One of these is that while the European actors that initiated the era had only a small technological lead over a half dozen other Eurasian civilisations when it began (Maddison 2001; Ertan et al. 2012), they were to gain a much larger lead – a phenomenon dubbed 'the great divergence' (Pomeranz 2000) – by the time it ended. Presumably, the combination of access to cheap resources, coerced labour, expanded markets for their products and vast territories to which surplus population could be exported helped European countries to begin pulling ahead, while the Ottoman Empire, Persia, India, China and Japan remained relatively stagnant. Although debate continues as to how important European expansion was to the location of the first Industrial Revolution in northwest Europe, it is difficult to dismiss it as a contributing factor.

The other major departure from persistence is that some non-European territories which had been relatively advanced within their regions when the colonial era began – the Aztec and Inca Empires being the earliest cases in point – were falling behind by the era's end, whereas some that had been relatively backward – e.g. what are now the USA, Canada, Australia and New Zealand – became relatively advanced. These examples form part of the pattern that Acemoglu et al. (2002) call the 'reversal of fortune'. They show that in general, colonised countries that had been more developed in 1500 as measured by urban proportion of population or population density, were less developed in 1995.

One way to see the examples of this reversal is that they in fact reflect how technological gaps created by the unevenness of *agricultural civilisation*'s spread allowed colonisers, including those soon to be leading industrial powers, to exert control over the regions they now dominated. In previously underdeveloped areas, with low population densities rendered lower still by Eurasian diseases, the colonisers helped themselves to territories they wished to settle in (such as the USA, Canada, Argentina, Australia and New Zealand) and brought slave or other labourers to work plantations in those they preferred to exploit using a relatively small number of overseers, only (Cape Verde, Cuba, Jamaica, Haiti, Dominican Republic). In more populous colonies, with land suitable for export crop production or mining, the pressing of locals into service occurred (for example, in Sri Lanka, Indonesia and Bolivia), with less dramatic impact on the composition of the population, but with important and often negative impacts on the local social structure.

Acemoglu et al. recognise that preconditions in the countries colonised – a disease environment deadly to Europeans (emphasised in Acemoglu et al. 2001) or population density (emphasised in Acemoglu et al. 2002) – helped determine which subsequently prospered and which did not. But the critical intervening factor, in their view, is whether the colonisers imposed *extractive* institutions, which favour the forced exploitation of labour and extraction of natural resource wealth, or *inclusive* institutions, especially property rights that encourage investment in skills and physical capital.

Glaeser et al. (2004) question whether it was European institutions or the human capital that Europeans brought with them to colonised countries that accounts for their development. Easterly and Levine (2012) find that the share of Europeans in a colonised country's population at roughly the mid-point of its years as a colony is a strong positive predictor of the post-colonial country's recent income, even for countries with relatively small numbers of European settlers.

Putterman and Weil (2010) point out that the changes in who lived not only in countries of the Americas, Australia and New Zealand, but also countries including Cape Verde, South Africa and Singapore, need to be properly accounted for in order to assess claims of the persistence of early developmental advantages into recent times. They construct a matrix of geographic origins of current populations' year 1500 ancestors classified by the country borders of today. Using it, they recalculate early indicators for contemporary countries based on the places their people's ancestors lived in, rather than the places the descendants inhabit today. They find that the ability of indicators of early and year 1500 technological development, including time of transition to agriculture, depth of experience with large-scale home-based states, and the technology indicator for year 1500 used by Comin et al. (2010), have considerably better ability to predict levels of development in 2000 when countries are assigned the weighted average early development values of their populations' ancestors rather than the unadjusted values for the history of the present territory.

Chanda et al. (2014) use Putterman and Weil's ancestry accounting methodology to revisit the reversal of fortune in Acemoglu et al.'s sample of colonies as well as in a corresponding sample that excludes migration outliers such as the USA, and in a broader sample of non-European countries. They replicate the reversal of fortune for countries as territories, but find no sign of a reversal of fortune for populations and their descendants. This suggests, for example, that the fact that what is now the USA developmentally overtook what is now Mexico despite the latter's lead around 1500, is explicable by the changes in population ancestry that occurred since that time. Most ancestors of contemporary Mexicans are Amerindian, whereas most ancestors of US residents are European, and Europeans had a technological edge over Amerindians even in 1500, so the result of apparent reversal adheres to a pattern of persistence after all.

Why Europe?

Why Western and Northern Europe rather than Europe's previously more advanced south or some other part of the Eurasian civilisational core, including China, was the coloniser of the other continents, and thereafter the first to industrialise, is a question that has loomed large for economic historians and growth theorists. The story of the 15th century Chinese fleet commanded by Admiral Zheng He, which eclipsed the expeditions of Columbus, da Gama and Magellan in ship number and size, and which visited ports throughout the Indian Ocean simply to display China's might, has become well known thanks in part to its role in discussions of this issue. Contributors including Diamond (1997) and Landes (1998) suggest plausibly the role of the difference between the rule over the vast area of China by a single emperor and the competition among smaller states in Europe, which increased the likelihood that differing strategies could find adequately resourced sponsors. They speculate that the difference in political unification may result from one of geography: China's core river valleys are joined by common adjacent plains, and it had a largely inland-centred civilisation with few major peninsulas and islands, whereas Europe is in a sense a central peninsula jutting westward at Iberia, sprouting several large appended peninsulas (e.g. Italy, Scandinavia) and neighbouring islands (Britain, Ireland, Sicily). Morris (2010) also attributes the European rather than Chinese 'discovery' of the Americas to the much smaller distance and more favourable currents and winds in the passages from Europe to those continents compared with any potential Chinese route across the Pacific. Voigtlander and Voth (2013) see high Black Death mortality playing a key role by raising Western European wages, which encouraged greater urbanisation that (in their view) raised mortality rates and wages still further. Acemoglu and Robinson (2012) focus on the emergence of favourable political institutions.

Part of the answer to the question of 'why Europe?', and why northwest Europe in particular, may lie in a phenomenon that fails to align with an overly strict notion of persistence of advantages, but that might be embraced by students of persistence as an important *caveat* to their framework. It is that once they have incorporated a sufficient share of the useful technological and cultural ideas of an established core, other areas may achieve more rapid progress than that core itself due to resource advantages, political flexibility or simply greater openness to creativity and innovation. On the resource front, a technology might be developed in one place – in the case of agriculture and animal domestication, for instance, due to location of the wild progenitor – but may actually prove more hardy in another environment (some European soils proved superior to those of the Near East). Core area resources may also be depleted by overexploitation. On the political front, being peripheral to a battle for influence between empires in a civilisational core might prove an advantage, since the core suffers repeated stresses that the periphery avoids. Olsson and Paik (2013) document a reversal of fortune among the Old World lands that inherited the agricultural and other technologies initially spawned in the Near East. Something similar appears to have happened on a more condensed time scale with respect to the comparative development of Japan, Korea and China between 1800 and the late 20th century.

Northwest European dynamism also invites a '(reverse) resource curse'-type explanation. Goods from Asia, including spices, tea and porcelain, enjoyed considerable demand in Europe, whereas only the precious metals Europe could muster (gold from Africa, and later silver from the New World) were much valued in the east. Venice and Genoa had been hubs for the Asian trade via the Near East, but much of the profit had long been lost to middle-men who bore the goods through overland routes on the Asian mainland. The Atlantic-facing monarchies were motivated by the large profits to be made by monopolising new sea-based trade routes that cut out those middle-men. Chinese and other Asian traders, in contrast, saw little incentive to sail to Europe, since it had little to offer them and had an overall economy and population that were small relative to Asia's.

Finally, as technological advances like the compass and the lateen sail made ocean-going trade more feasible, those parts of Eurasia with better ports and ice-free sea access were advantaged. Although the Chinese civilisation had developed maritime capabilities, they were concentrated on its south-eastern coast at a considerable distance from the country's more northerly and central core. That core consisted of inland river valleys whose rulers had for centuries seen their key defensive challenge as being protection from nomadic invaders from the inner Asian steppe. The decision of the later Ming emperors to mothball Admiral Zheng He's fleet, which they considered an expensive extravagance, has a certain internal logic given that their predecessor (and Zheng He's benefactor), the Yong-le Emperor, had died while engaging in a military campaign against the Mongols. Indeed, the Ming would later be overthrown by the semi-nomadic Jurchen (Manchu) aided by Mongol forces, whereas the Portuguese and Dutch colonisers active in the Indian Ocean gave China little to worry about in the early 1600s. On the other hand, a navy might have appeared more affordable had the Ming not engaged in massive construction programs centred on their own tombs and palaces. So the grandiose and self-centred mind set of emperors who saw their realm as the only centre of true civilisation cannot entirely be left out of the equation.

Diffusion of the Industrial Revolution

Despite the considerable difference in pace, the spread of the Industrial Revolution and its correlates has some qualitative similarities to that of the agricultural one(s). Much as agriculture spread from the Near East to Europe, from West to central and southern Africa, from Mesoamerica northward, and so forth, so industrial technologies spread more quickly to neighbours, giving rise to inequalities of development resembling those of the earlier agricultural revolutions. (Thanks to its Malthusian character, however, the earlier revolution's inequalities were more ones of technology than of income; see Ashraf and Galor 2011.) One important qualitative difference is that improved transport allowed pre-existing cultural and linguistic affinities to play a larger relative role, compared with geographic proximity, in diffusing the Industrial Revolution. The shrinking of the globe by improved navigation also made it unlikely that any society would begin a separate industrial revolution in its own corner of the world before learning of the one already in progress.

When comparing pairs of countries such as the USA or Australia on the one hand to China or Korea on the other, 'the last becoming first and the first last' during the later centuries of the 2nd millennium may be explained not only by migration (as in the USA/Mexico comparison above) but also by the relative technological lethargy exhibited by the East Asian countries that had earlier been close to, or even occupying, the world technological frontier. Whatever the reason for its relative lethargy going into the 20th century, however, bringing Asia into the discussion can help us transition to another important observation: that the colonial era reversal of fortune began to be reversed itself by a post-colonial era 'catch up' process that also shows signs of differentiation by pre-modern development levels. Within the non-European world, that is, countries populated by descendants of societies that had enjoyed a head start before 1500 (for example China, Taiwan, Singapore and to some extent India and Turkey) have been growing faster than their counterparts (for instance Guatemala, Haiti, Malawi and New Guinea). Chanda and Putterman (2007) note that among non-European countries that manifested a large income and technology gap with Europe and its off-shoots circa 1950 CE, countries that had had relatively advanced agrarian societies circa 1500 were more likely to experience rapid growth in the late 20th century than ones mainly characterised by pastoralism or horticulture, such as the countries of Central Asia, North Africa and sub-Saharan Africa. Thus, the early post-colonial catch-up phase for non-European countries exhibited a 'persistence-like' tendency for formerly advanced societies to catch up more quickly. This is all the more strongly confirmed when the migration-accounting methodology discussed above is brought to bear (Chanda et al. 2014).

Comin et al. (2008) provide detailed support for the interpretation of differences in levels of economic development as reflecting lags in the adoption of more advanced methods of production. They study late 20th century lags in catching up with the world leader's intensity of use of individual technologies, including electricity, personal computers, the Internet, cell phones and landline phones, air shipment, passenger aviation, commercial and passenger vehicles, and tractors. They find strong correlations between given countries' adoption rates and their per capita incomes, and they conclude that differences in the intensity of usage of such technologies might account for a large part of cross-country differentials in factor productivity. The importance of such adoption lags is consistent with persistence of development levels if given countries tend to sustain similar relative adoption lags in relation to technological leaders over long periods of time.

Spolaore and Wacziarg (2013a, b) address the tendency for innovation to spread more rapidly between genetically, linguistically or geographically closer populations, suggesting that it may reflect essentially the same phenomenon as the persistence of development levels. Populations living at similar geographic distances from a common technological leader are likely to have adopted given innovations at similar times, thus scoring similarly on measures of early technological adoption or sophistication. Given their geographic proximity, they may also be relatively closely related in terms of ancestry and language. To study the impact of relatedness between populations, Spolaore and Wacziarg (2009) use Cavali-Sforza et al. (1994)'s estimates of genetic distance. These estimates are based on differences in genes believed to be non-trait-encoding, which should be subject to random drift at common rates and thus able to serve as measures of populations' historical separation times. Using these estimates, the authors predict differences in average income between 6,800 pairs of countries in 1990. They find genetic distance strongly predictive of differences in income in 1990 when controlling for geographic distance. Genetic distance is also predictive of estimated income differences for the years 1500 and 1700, as well as differences of institutional quality measured by expropriation risk, and differences of schooling, of investment in physical capital, and of other factors often treated as proximate explanations for development.

Spolaore and Wacziarg (2012, 2013b) take an additional step towards demonstrating a link of the results just mentioned to the question of diffusion rates. They find that the relative genetic distance of each pair of countries from a technological leader, such as Italy or the UK in 1500, or the UK or USA today, is a stronger predictor of the difference in incomes between those countries than is the absolute genetic distance between the pair. This implies that genetic distance from the technological leader is what explains level of development, with this kind of distance acting over and above geographic distance as a barrier to diffusion of technology. The authors find similar results for differences in the adoption of specific components of the Comin et al. year 1500 technology index and for differences in use of a number of specific industrial technologies studied by Comin and Hobijn (2010). Because linguistic distance (number of language branchings since two languages shared a hypothesised common ancestral language) is highly correlated with genetic distance, it is possible that the measured effects of genetic distance are to some extent proxying for differences in language or culture.

Culture, 'Broad Human Capital' or 'Social Capability' as Key

Putterman (2000) and Chanda and Putterman (2004) advance the view that a key determinant of the differential rates with which different low-income economies have been catching up to world leaders since the mid-20th century is the set of outlooks and capabilities collectively held by a population, which Putterman dubs 'broad human capital' and which coincides at least partly with what economic historian Moses Abramovitz (1986, 1995) called 'social capability'. 'Broad human capital' is to be distinguished from a conception of human capital focusing mainly on formal learning. It includes tacit knowledge (Nelson and Winter 1982) as well as what Heckman and others (see, for example, Heckman et al. 2006) call non-cognitive skills. The word 'culture', in the anthropological sense of the shared mental software of a society, has essentially the same meaning, and is avoided here primarily due to the tendency of non-anthropologists to associate the term exclusively with distinctive tastes and practices. 'Social capability' as used by Abramovitz includes an outlook compatible with empirical science, with an effective incentive structure and with effective political institutions, as well as the spread of education, experience with administering large-scale organisations, and with the functioning of capital markets. The term also spills over into the actual presence of effective political and financial institutions, adding similarity to what Hall and Jones (1999) call 'social infrastructure'. While differing in exact coverage and degree of prescriptiveness, all of these terms treat intangible collectively-held attitudes and capabilities as a key factor influencing the relative performance of societies in the 'catch up' or 'late development' process.

Several of the above and other economists have offered evidence of the importance of the factors in question. Abramovitz (1986) found a tendency towards convergence in productivity within a sample of OECD countries but not in a broader sample including middle and low income countries during 1950–1980. He concluded that 'a country's potential for rapid growth is strong not when it is backward without qualification, but rather when it is technologically backward but socially advanced'. Hall and Jones (1999) constructed a social infrastructure index based on an index of country risk and an index of openness to international trade. When they instrumented for 'social infrastructure' by the extent to which French, German, Spanish and English are spoken as first languages and distance from the equator, they found that it could account for a substantial share of the variation in output per worker among countries in their global sample. Temple and Johnson (1998) measured 'social capability' by an index of social modernisation indicators compiled in the 1960s. While controlling for human capital and investment rates, they found the index to be a good predictor of the rates of economic growth of 60 developing countries during 1960–85. They offered this as evidence that social capability is a potentially measurable concept that affects economic growth through channels other than investment and education.

Putterman (2000) linked his broad human capital concept to social evolutionary schema. In the latter, modes of social organisation and economic/ecological adaptation to environments, which range from foraging to horticulture, pastoralism, intensive agriculture and industrial society, form a loose continuum with respect to criteria including time of initial appearance, population density, social complexity and stratification, and scale of largest effective political unit (Service 1971; Johnson and Earle 2000; Richerson et al. 2001). Putterman argues that habituation to unbroken periods of intensive labour and frequency of interactions in larger organisations, in which family and personal relationships are subordinated to hierarchical structures, tend to increase along the progression from foraging to modern industry. Because of this, he suggests that societies such as China, India, Iran, Korea, Japan and Turkey, which had operated as large-scale agrarian states for centuries prior to the Industrial Revolution, have been able more easily to adopt the organisational forms and technologies of an industrial society than have ones that had depended mainly on foraging or horticulture. In the latter societies, such as most in central and east Africa and Papua New Guinea, large-scale states were absent or a relative novelty. Where societies of the latter type were not reconstituted by large numbers of immigrants (as in the Australian or US cases), catching up with industrialised societies could be expected to take longer and to proceed initially at a slower pace. Focusing on rates of growth during 1960–1990, rather than income level, Burkett et al. (1999) found support for this conjecture when using population density and cultivated land per capita as proxies for level of pre-modern development. Bockstette et al. (2002) and Chanda and Putterman (2007) found parallel support for the proposition that higher levels of pre-modern development are associated with faster economic growth in the late 20th century, using depth of historical experience with large-scale states as the indicator of early development, and using the growth rates of 1960–1995 and 1960–1998 as dependent variables.

The social evolutionary framework just discussed arguably parallels both the persistence approach of Comin et al. (2010) and the barriers to diffusion approach of Spolaore and Wacziarg (2012, 2013b). Over a sufficiently short time period, at least, it is to be expected that societies recently characterised mainly by horticulture and/or pastoralism will have lower quality of governance in their more recently formed large-scale states, and will have lower adoption rates of advanced technologies, than will societies that were at an advanced stage of intensive agrarianism in their last pre-industrial period. In so far as people having more recent genetic and linguistic commonalities are more likely to have been at similar stages of social evolutionary change, the greater ease with which people at more ‘advanced’ positions on the evolutionary continuum could adopt new industrial organisational forms and technologies could contribute, along with other cultural affinities including that of language, to the speedier diffusion of technology. For example, European speakers of Indo-European languages of the Celtic, Romance, Germanic and Slavic groups tended to be in various stages of transition from agrarian to industrial society by the late 19th century, whereas sub-Saharan African speakers of Bantu and Nilotic languages tended to still be living in horticultural and pastoral societies. To the smaller linguistic and geographic distances between less developed European countries and Britain than between African countries and Britain, one might therefore add the smaller distance of the European countries with respect to social evolutionary stage as a reason why the former countries were able to follow Britain into industrialisation more quickly.

Proximity with respect to early development may also help to explain the principal outlier of 19th century industrial diffusion: industrialisation’s leap from northwest Europe to Japan. Whereas this instance of diffusion involves exceptionally large linguistic and geographic distances, it is less of an outlier with respect to the differences between Japan’s levels of agricultural and urban development and those in Europe. Subsequent transmissions from Japan to its neighbours accord with expectations about both geographic and linguistic distance on the one hand and difference in level of development on the other.

Conclusion

Economists studying very long-run determinants of economic growth and standard of living have focused on geography, institutions, culture and history. This article has discussed institutions only peripherally and mainly with respect to the emergence of large-scale states. At least some institutional quality measures, for instance prevalence or absence of corruption and general efficiency or quality of bureaucratic performance, are closely linked to culture in the sense of the last section's 'broad human capital'. The strong performance of institutional measures in studies seeking to explain differential rates of growth or levels of development may accordingly also be linked to the themes discussed here, and there need be no fundamental tension between these approaches (see, for example, Ang 2013).

Our discussion considered history, geography and culture within a long-term historical rubric in which differences in the origination and spread of technological traditions are attributed to geographic factors which include the potential for relying on a similar agricultural way of life in civilisations spread widely over the Eurasian landmass, the absence of such potential in most of sub-Saharan Africa, the relative ease of contacts across Eurasia, and the cutting off of Oceania and the Americas given known methods of communication and travel during the centuries leading up to 1500. The long-term persistence of technological advantages, coupled with dynamism on well-connected peripheries, has been a rule of economic history. And large differences in culture, language and stage of development have until now served as barriers to the diffusion of technology.

With large technological gaps still dividing rich and poor countries today, finding ways to reduce the barriers to diffusion could be one of the most promising ways in which to seek remedies to underdevelopment and poverty in the global 'South'. Programs that accelerate the building of capacity in technological, institutional and other dimensions should accordingly be emphasised. Even without such programs, however, the rapid urbanisation of the developing world and the diffusion of a global culture via television, the Internet and other media, may work powerfully to close existing gaps, thus helping to consign some differences in outlook and capability to history.

See Also

- ► Agriculture and Economic Development
- ► Development Economics
- ► Regional Development, Geography of

Bibliography

Abramovitz, M. 1986. Catching up, forging ahead, and falling behind. *Journal of Economic History* 46: 385–406.

Abramovitz, M. 1995. The elements of social capability. In *Social capability and long-term economic growth*, ed. D.H. Perkins and B.H. Koo, 19–47. New York: St Martin's Press.

Acemoglu, D., and J. Robinson. 2012. *Why nations fail: The origins of power, prosperity, and poverty*. New York: Crown Publishing.

Acemoglu, D., S. Johnson, and J. Robinson. 2001. The colonial origins of comparative development: An empirical investigation. *American Economic Review* 91(5): 1369–1401.

Acemoglu, D., S. Johnson, and J. Robinson. 2002. Reversal of fortune: Geography and institutions in the making of the modern world income distribution. *Quarterly Journal of Economics* 117(4): 1231–1294.

Ang, J. 2013. Institutions and the long-run impact of early development. *Journal of Development Economics* 108: 1–18.

Ashraf, Q., and O. Galor. 2011. Dynamics and stagnation in the Malthusian epoch. *American Economic Review* 101(5): 2003–2041.

Bockstette, V., A. Chanda, and L. Putterman. 2002. States and markets: The advantage of an early start. *Journal of Economic Growth* 7: 347–369.

Burkett, J., C. Humblet, and L. Putterman. 1999. Pre-industrial and post-war economic development: Is there a link? *Economic Development and Cultural Change* 47(3): 471–495.

Cavali-Sforza, L., P. Menozzi, and A. Piazza. 1994. *The history and geography of human genes*. Princeton: Princeton University Press.

Chanda, A., and L. Putterman. 2004. The quest for development: What role does history play? *World Economics* 5(2): 1–31.

Chanda, A., and L. Putterman. 2007. Early starts, reversals and catch-up in the process of economic development. *Scandinavian Journal of Economics* 109(2): 387–413.

Chanda, A., C. J. Cook, and L. Putterman. 2014. Persistence of fortune: Accounting for population movements, there was no post-Columbian reversal. *American Economic Journal – Macroeconomics* [in press, to appear in August issue].

Comin, D., and B. Hobijn. 2010. An exploration of technology diffusion. *American Economic Review* 100(5): 2031–2059.

Comin, D., B. Hobijn, and E. Rovito. 2008. World technology usage lags. *Journal of Economic Growth* 13(4): 237–256.

Comin, D., W. Easterly, and E. Gong. 2010. Was the wealth of nations determined in 1000 B.C.? *American Economic Journal: Macroeconomics* 2(3): 65–97.

Diamond, J. 1997. *Guns, germs and steel: The fate of human societies*. New York: Norton.

Dow, G., C. Reed, and N. Olewiler. 2009. Climate reversals and the transition to agriculture. *Journal of Economic Growth* 14(1): 27–53.

Easterly, W., and R. Levine. 2012. *The European origins of economic development*. NBER working paper no. 18162.

Ertan, A., L. Putterman, and M. Fiszbein. 2012. *Determinants and Economic consequences of colonization: A global analysis*. Brown University Department of Economics working paper 2012–5.

Gallup, J.L., J. Sachs, and A. Mellinger. 1999. Geography and economic growth. In *Annual World Bank conference on development economics 1998*, ed. B. Pleskovic and J. Stiglitz, 127–178. Washington: The World Bank.

Glaeser, E., R. La Porta, F. Lopez-de-Silanes, and A. Shleifer. 2004. Do institutions cause growth? *Journal of Economic Growth* 9(3): 271–303.

Guiso, L., P. Sapienza, and L. Zingales. 2006. Does culture affect economic outcomes? *Journal of Economic Perspectives* 20(2): 23–48.

Hall, R., and C. Jones. 1999. Why do some countries produce so much more output per worker than others? *The Quarterly Journal of Economics* 114: 83–116.

Heckman, J., J. Stixrud, and S. Urzua. 2006. The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics* 24(3): 411–482.

Hibbs, D., and O. Olsson. 2004. Geography, biogeography, and why some countries are rich and others are poor. *Proceedings of the National Academy of Sciences* 101(10): 3715–3720.

Johnson, A., and T. Earle. 2000. *The evolution of human societies: From foraging groups to Agrarian State*, 2nd ed. Stanford: Stanford University Press.

Landes, D. 1998. *The wealth and poverty of nations*. New York: Norton.

Maddison, A. 2001. *The world economy: A millennial perspective*. Paris: Development Centre of the OECD.

Morris, I. 2010. *Why the West rules – For now: The patterns of history, and what they reveal about the future*. New York: Picador/Farrar/Straus and Giroux.

Nelson, R., and S. Winter. 1982. *An evolutionary theory of economic change*. Cambridge, MA: Harvard University Press.

Nunn, N. 2014. Historical development. In *Handbook of economic growth*, vol. 2, ed. P. Aghion and S. Durlauf. Amsterdam: Elsevier.

Nunn, N., and N. Qian. 2010. The Columbian exchange: A history of disease, food, and ideas. *Journal of Economic Perspectives* 24(2): 163–188.

Olsson, O., and D. Hibbs. 2005. Biogeography and long-run economic development. *European Economic Review* 49(4): 909–938.

Olsson, O., and C. Paik. 2013. A western reversal since the Neolithic? The long-run impact of early agriculture. Working papers in economics 552. University of Gothenburg.

Pomeranz, K. 2000. *The great divergence: China, Europe, and the making of the modern world economy.* Princeton: Princeton University Press.

Putterman, L. 2000. Can an evolutionary approach to development predict post-war economic growth? *Journal of Development Studies* 36(3): 1–30.

Putterman, L. 2008. Agriculture, diffusion, and development: Ripple effects of the Neolithic revolution. *Economica* 75: 729–748.

Putterman, L., and D. Weil. 2010. Post-1500 population flows and the long run determinants of economic growth and inequality. *Quarterly Journal of Economics* 125(4): 1627–1682.

Richerson, P., B. Vila, and M. B. Mulder. 2001. *Principles of human ecology.* Unpublished. Davis: University of California. Available at http://www.des.ucdavis.edu/faculty/Richerson/BooksOnline/101text.htm. Accessed 12 March 2014.

Service, E. 1971. *Cultural evolutionism: Theory in practice*. New York: Holt/Rinehart and Winston.

Spolaore, E., and R. Wacziarg. 2009. The diffusion of development. *Quarterly Journal of Economics* 124(2): 469–529.

Spolaore, E., and R. Wacziarg. 2012. Long-term barriers to the international diffusion of innovations. In *NBER international seminar on macroeconomics 2011*, ed. J. Frankel and C. Pissarides, 11–46. Chicago: University of Chicago Press.

Spolaore, E., and R. Wacziarg. 2013a. How deep are the roots of economic development? *Journal of Economic Literature* 51(2): 325–369.

Spolaore, E., and R. Wacziarg. 2013b. Long-term barriers to economic development. NBER working paper #19361.

Temple, J., and P. Johnson. 1998. Social capability and economic growth. *Quarterly Journal of Economics* 113(3): 965–990.

Voigtlander, N., and H.J. Voth. 2013. The three horsemen of riches: Plague, war and urbanization in early modern Europe. *Review of Economic Studies* 80(2): 774–811.

History of Economic Thought

Craufurd D. Goodwin

Abstract

Attention was paid to the history of economic thought (HET) by pioneers of economics such as Dupont de Nemours and Adam Smith. Classical economists like J.R. McCulloch in the 19th century used HET to establish a canon of economic literature, and their successor marginalists such as William Stanley Jevons to demonstrate progress in the subject. From the First World War until the 1960s, leading economists, from Jacob Viner to Wesley Mitchell, employed HET to cast light on current research. In the 1970s HET became a separate sub-discipline with its own periodicals and meetings. The number of scholars who worked in HET did not decline, even though the major research and postgraduate training centres lost interest.

Keywords

American Economic Association; Austrians economics; Blanqui, J.-A.; Böhm-Bawerk, E. von; Cannan, E.; Classical economics; Du Pont de Nemours, P. S.; Economic history; Historical School; History of economic thought; History of Economics Society; Ingram, J. K.; James, E. J.; Jevons, W. S.; Knight, F. H.; Marginal revolution; Marshall, A.; McCulloch, J. R.; Mitchell, W. C.; Physiocrats; Schumpeter, J. A.; Smith, A.; Twiss, T.; Viner, J.

JEL Classifications

B2

The history of economic thought (hereafter HET) is explored today for the most part within a sub-discipline of economics. (The literature on this topic is rather limited. Blaug 1991 is an anthology of relevant articles. Two useful bibliographical works are Howey, 1982 and Stark 1994. The history of economic thought in Britain is examined in Backhouse 2004. Selected histories of economic thought are reprinted in Backhouse 2000.) It shares a category in EconLit, the indexing service of the American Economic Association, with methodology, where it is called 'Schools of Economic Thought'. Scholars in the sub-discipline conduct various kinds of studies: interpretive biographies, narrative accounts of the growth of ideas and their impact on society, rational reconstructions of the emergence of theory, the behaviour of scientific and intellectual communities, and more. Some 476 members of the American Economic Association declared 'methodology and the history of economic thought' as a field of interest in 2006. There are more than 1,000 scholars seriously interested in HET worldwide. The three main journals in the field (*History of Political Economy, Journal of the History of Economic Thought* and *European Journal of the History of Economic Thought*) have a combined circulation of about 2,000. Approximately 200 scholars attend each of the annual meetings of the continental societies for the study of HET, and the Japanese society has over 800 members.

The location and style of HET today are in contrast to those of the histories of most other scientific disciplines, which are found usually not within the discipline under study but within one of the sub-disciplines of history known as 'history of science' or 'intellectual history'. Only the more humanistic disciplines like literature and art history and, within the social sciences, political science tend still to study their history within their known communities. Unlike those studying most other scientific disciplines, historians of economics have generally been trained as economists rather than as historians; this training gives them the perspective on their subject of insiders, but also, sometimes, the historical skills of amateurs. Scholars of HET are likely to teach in economics, not in history.

From approximately the First World War until the 1960s HET was lodged comfortably in the 'core' of economics. One or two courses were required of students at both the undergraduate and graduate levels, taught alongside micro and macro theory and statistics. Economics faculty began their courses on almost any subject with an introduction to the evolution of relevant theory. Indeed, HET was thought of as simply an historical extension of theory, and practitioners as simply a special kind of theorist with a long time horizon. Scholars of HET met other economists at conferences of the national and international economics societies. They did not think of themselves as a separate sect within the discipline, and saw no reason to have their own meetings or associations. They published in the mainstream economics journals and in the publications of several friendly adjacent disciplines such as history, philosophy, sociology and political science.

However, in the 1950s and 1960s this landscape changed. HET was banished from the core of economics to the margins of the discipline, ostensibly to make room for more technical economic theory and burgeoning econometrics. From being a requirement in the curriculum, HET became an option for graduate and undergraduate students – if there was someone to teach it, and increasingly there was not. The mainline professional societies and journals showed less and less hospitality to HET. Even the sister sub-discipline, economic history, then in the grip of the cliometric revolution and under scrutiny itself for relevance, seemed more and more uneasy about close relations with a subject that was 'literary'. More and more of the major postgraduate training programmes abandoned HET formally when those who taught the subject retired and were not replaced.

The response to this crisis among those in HET in the 1960s was to regroup and create a new infrastructure in which to operate, and a sub-discipline of HET effectively came into existence. The first journal dedicated exclusively to the field, *History of Political Economy*, began in 1969, and the History of Economics Society (HES) for specialists in the subject was established in 1974. Both of these new institutions, although based in the United States, were intended to serve a worldwide community. Joint sessions of the HES with the American Economic Association and other bodies of economists continued, but the HES annual meetings became the most popular gatherings where specialists might gather and interact. A paradoxical situation, then, exists in HET in the first decade of the 21st century. While the memberships in societies and numbers of books and articles published annually is at least stable, coverage of the subject in the premier graduate training and research centres and in the mainstream periodicals of economics has steadily decreased almost to nothing. In the United States those few graduate students who specialize in the field do so usually through a jerry-built tutorial programme with a faculty mentor, and a dissertation dictated by the job market made up of only one essay in HET and two in more saleable fields. External funding for HET, unless it is camouflaged as policy studies or theory, is almost non-existent. So what then explains the impressive place gained by HET in the economics discipline at the middle of the 20th century and its precipitate fall, in prestige and respect, within the larger discipline at least, by the end of the century? The answer lies in the subject's own history, beginning in the 18th century. Five distinct historical periods can be discerned.

Period I. The Enlightenment: HET as Rhetoric

HET began at about the same time as the discipline that it studies. The 18th-century Physiocrats clearly held in low regard many of the early thinkers on questions with which they were engaged; for example, they often denigrated the thinking of Colbert, the French Minister of Finance. But they used HET less as a weapon against those with whom they disagreed than to proclaim their own remarkable accomplishments. The Physiocrats assigned Pierre Samuel Du Pont de Nemours the task of historian. His short monograph, *De L'Origine et des progrès d'une science nouvelle* (1768), may be considered the earliest treatise in HET. Dupont claimed that Quesnay and his colleagues had for the first time discovered a body of doctrine that 'following the nature of man, exposed the laws *necessary* for a government to make for man in all climates and in all countries' (1768, p. 35). His book was mainly a celebration of this achievement.

Adam Smith was not as cautious in his criticism as were the Physiocrats. He was exceptionally well read, knew the economic literature of his day intimately, and was not shy about offering judgements. He cited some writers on economic topics in support of his views, from Aristotle onwards, and condemned others. But he did not in any sense produce a serious and balanced history of economic thought. He had favourites, such as his friend David Hume, and pointed out some whose ideas were intriguing, like Matthew Decker and Bernard Mandeville. But he did not present the work of his predecessors as constituting a unified body of thought or leading inexorably to his own. Smith praised the work of the Physiocrats, and especially that of 'the very ingenious and profound author ... Mr. Quesnai'. But he also condemned out of hand earlier thinkers who held fundamentally different views. About as charitable as Smith could be towards those who had expressed policy conclusions at variance with his own was that their 'arguments were partly solid and partly sophistical' (Smith 1776, p. 433). Neither the Physiocrats' self-congratulation nor Smith's imaginary debates with his predecessors were important contributions to a history of economic thought.

Period II. Classical Political Economy: HET for Cartography and Doctrinal Cleansing

Thomas Robert Malthus and David Ricardo were neither very interested in nor respectful of their intellectual ancestors; they made occasional references to earlier work (Smith's Wealth of Nations was particularly important to them) but they made no systematic attempt to frame it as a whole. Not so their immediate successors, the second generation of what came to be known as the classical economists: James Mill, Nassau Senior, Robert Torrens, James Ramsay McCulloch and others. These later classicals came increasingly to believe that, despite the continuing disputes over important points of theory, something approaching ultimate truth had been achieved in the work of the founders. Senior suggested in the 1820s even that the core of political economy could be expressed in a few simple propositions derived from the founding fathers' work. From these propositions could be inferred both principles of high policy by which governments should abide, and principles to guide individual human action (Senior 1827, pp. 35–6). Yet among the loose community of businessmen, journalists, public servants and others who pursued classical political economy during the first three-quarters of the 19th century there was relatively little agreement about what should be included in the canon. There were various elementary primers for those entering the field but no definitive textbooks (with the possible exception of J.S. Mill's Principles in 1848), professorial oracles, or dominant professional periodicals to which one might turn for definitive judgements. Indeed, virtually anyone could make a claim for inclusion of his ideas in classical political economy simply by publishing in one of the many generalist reviews.

It was to correct this condition of seeming doctrinal anarchy and inconsistency, and to impose some discipline upon an unruly conversation, that the classical economists turned to HET. Historical investigation could, perhaps, help map the new discipline and discern who and what were respectable contributions to political economy and who and what were not. Each of the doctrinal cartographers and cleansers had his own ideas of what orthodoxy should be imposed (Villeneuve-Bargemont 1841, even named consistency with Christian theology as a criterion for inclusion). Some were Smithians, some Ricardians and some paid allegiance to an amalgam of doctrines. But their common purpose in going to the past was to sort out just what should guide the present. An example of a work to this end is the book *View of the Progress of Political Economy in Europe since the Sixteenth Century* (1847), which contained a course of lectures delivered by Travers Twiss, Professor of Political Economy in the University of Oxford. Twiss aimed to demonstrate that genuine works of political economy, as the subject had evolved since Adam Smith, employed the scientific method, which he described as testing theory by history so as to produce results that could benefit society: 'leading doctrines are the conclusions of an enlarged experience, and are not, as many persons suppose, mere deductions from arbitrary premises skillfully assumed' (1847, p. v). Twiss described the ill effects that could follow from the 'unsound theory' of such writers as Colbert and John Law. Twiss explained clearly how he proposed to use HET as a device to purge political economy of any false doctrines by which it had become corrupted. 'I have attempted in the course of the above inquiry to assign to the chief writers their due shares respectively in furthering the progress of sound opinions, but I have purposely omitted the names of many authors of eminence, who have struggled to retard that progress, although they may have indirectly furthered it by the controversy which they have provoked' (1847, p. viii).

On several occasions John Ramsay McCulloch, like Twiss, gave an account of the progress of political economy as a morality tale. He pictured truth ultimately conquering error despite the strong forces massed against it. In his pioneering textbook *Principles of Political Economy* (first published in 1825) McCulloch included a chapter on 'the rise and progress of the science'. He explained that dissension amongst early economists had tended to discredit the subject among scientists generally, and political economists needed to present a united front: 'The differences which have subsisted among the most eminent of its professors have proved exceedingly unfavourable to its progress, and have generated a disposition to distrust its best-established conclusions' (1825, p. 14). One of McCulloch's primary objectives was to sort out truth from falsehood, so that political economy could gain the reputation and influence that it deserved: 'the errors with which this science was formerly infected are now fast disappearing; and a very few observations will suffice to shew, that it really admits of as much certainty in its conclusions as any science founded on *fact and experiment* can possibly do' (1825, p. 15).

McCulloch's view was that there had to be broad agreement in any subject for it to be considered a science, and therefore the history must be presented as leading towards consensus.

Histories of economic thought in the classical period often took on a distinctly nationalist tone. The cartographic function was perceived not only as filling in the map of the new discipline but also as making sure that some of the territory at least bore the home country's colours. Not all the map should be British red. The publication of these histories seems almost like the intellectual equivalent of the scramble for colonies that was in progress among the European nations at this time. Adolphe Blanqui, in what was as much an economic history of Europe as a history of economic thought, gave two chapters to Smith and Malthus wedged in between segments on the Physiocrats, Rousseau, the French Revolution and J.B. Say. He was relieved that the doctrines of the British 'industrial school' were no longer accepted without question thanks to the work of Sismondi and other French critics (Blanqui 1837, p. 262). A similar work in Italian was by Luigi Cossa (1876).

Period III. Neoclassical and Historical Economics: HET as Literature Review

Beginning with the marginal revolution of the 1870s HET took on a new role derived from what had become fashionable in the physical sciences and mathematics: the literature review. If economists were to be seen as true scientists, insisted economists such as William Stanley Jevons, they must walk and talk like them. They must not use the history of their subject to demonstrate a stable orthodoxy, as McCulloch and others had sought to do. The past of a science contained not an accumulation of what was true but of what had been found to be false and had been displaced by current doctrine. In praising the accomplishments of the Austrian marginalists, Böhm-Bawerk used an evolutionary metaphor to describe HET as the study of illness in scientific infancy and childhood. The Austrians, he wrote, 'are of the opinion that the errors of the classical economists were only, so to speak the ordinary diseases of the childhood of science ... Their greatest fault was they were forerunners; our greatest advantage is that we came after' (Böhm-Bawerk 1973, p. 362). The essence of science was progress and change. The purpose of the literature review to be included with any major work in science should be twofold: to pay due respects to worthy ancestors, and more particularly to use the past to demonstrate how certain prior works led inexorably to the present, superior, one. The literature review in a work of theory while acknowledging worthy predecessors also established claims to priority in the novel ideas set forth. HET had come into the service of Whig history. Above all, the emphasis had to be on change rather than on stability. William Stanley Jevons insisted that attention to the past should be seen as liberating and not as stifling deference to orthodoxy. He observed how ‘in the other sciences the weight of authority has not been allowed to restrict the free examination of new opinions and theories; and it has often been ultimately proved that authority is on the wrong side’ (1871, pp. v–vi). In the books of the marginal revolutionaries the literature review was placed usually in the preface or in an appendix. Jevons used both. The marginalists were as ready as Twiss or McCulloch to dismiss some predecessors out of hand; but their dismissal was focused especially upon those who differed in particular methods or results from the work currently being presented. All predecessors had necessarily been supplanted. Those who walked the right road, though too slowly, deserved to be remembered. Those who took the wrong road deserved to be condemned. Here is what Jevons wrote of McCulloch’s heroes: ‘When at length a true system of Economics comes to be established, it will be seen that that able but wrong-headed man, David Ricardo, shunted the car of Economic Science on to a wrong line – a line, however, on which it was further urged towards confusion by his equally able and wrong-headed admirer, John Stuart Mill’ (Jevons 1871, pp. li–lii). Jevons could congratulate Von Thunen, Dupuit, and Cournot; but for others, like John Stuart Mill, who were not on the right road to the marginal revolution, he had only contempt.

Each of the pioneer marginalists had his own way of incorporating a review of the literature into his text. In Menger the historical commentaries were long footnotes that so annoyed the translators of his *Principles of Economics* (1871) into English in 1950 that they appear there as a series of appendices. Marshall began with an introductory historical section on ‘the growth of economic science’ in the first edition of his *Principles of Economics* (1890) but shifted this material in the fifth edition (1907) to an appendix. Irving Fisher, lacking a single broad-based treatise of his own to which he could append an historical review of the literature, attached one to the translation of Augustin Cournot’s *Researches into the Mathematical Principles of the Theory of Wealth* (1897). These reviews of the literature by the marginalists often have a strikingly unsystematic and personalized appearance with offhand comments that seem out of place in a carefully reasoned text. For example, the following comment by Marshall in a generally laudatory mention of Ricardo and his work seems to reflect more his own casual prejudices than a serious study of history. Marshall wrote:

> his [Ricardos’s] aversion to inductions and his delight in abstract reasonings are due, not to his English education, but, as Bagehot points out, to his Semitic origin. Nearly every branch of the Semitic race has had some special genius for dealing with abstractions, and several of them have had a bias towards the abstract calculations connected with the trade of money dealing, and its modern developments; and Ricardo’s power of threading his way without slip through intricate paths to new and unexpected results has never been surpassed. But it is difficult even for an Englishman to follow his track; and his foreign critics have, as a rule, failed to detect the real drift and purpose of his work’. (Marshall 1920, p. 629 n.)

Edwin Cannan’s *A History of the Theories of Production and Distribution in English Political Economy 1776–1848* (1917) was a generalized and highly critical literature review of what had become settled doctrine a half century before the ‘new’ economics of Alfred Marshall. It set out to demonstrate that only with marginal tools had economics become a science. Cannan’s book was not like those of Twiss and McCulloch, which had sought to sift the wheat from the chaff in the confident belief that a pile of genuine truth would thereby be revealed. Cannan’s message was that everything before marginal economics was hardly worth a glance because none of it was science.

The marginalists were not the only ones in the late 19th century to use HET to bolster the legitimacy of their approach. The Historical School also concluded that a literature review demonstrated the strength of their position. The essence of their claim was that the usefulness of economic theory was relative to the circumstances in which the theory was applied. Different circumstances required different theory, and the history and appraisal of past theory had to keep in mind the tasks for which the earlier theory had been designed. The American historical economist E.J. James suggested that:

> the axioms and theorems which apply to one form of society may have little or no applications in another form, and any attempt to make such application may result in the most absurd conclusions ... Nor will a theory which is adequate to the demands of an industrial state like England or America suit such a country [*sic*] as India or Africa. (Ingram 1888, p. vii)

The historians wrote specifically in opposition to 'The assertion of J.B. Say' a doctrinal cleanser 'that the history of Political Economy is of little value, being for the most part a record of absurd and justly exploded opinions' (Ingram 1888, p. 2). This they found to be an unjustified dismissal of early economic thought.

The correct way to view the history of ideas, they were convinced, was as the record of how theory was useful at particular times and places and not either as a gradual but final movement towards some kind of ultimate truth, or as a steady accretion of scientific understanding. At the same time it must be conceded that the consequence of this posture by the historians was not very different from that of the marginalists; the details of HET, they implied, were largely of antiquarian interest. The difference between them was that the historical economists looked with more sympathy upon their predecessors, even those with whom they disagreed in their modern application.

The most detailed history in English taking the historical approach was by the Irish economic historian, John Kells Ingram. Ingram's findings were in part similar to and in part a contrast to those of Cannan. He agreed with Cannan on the failings of the classical economists, and he insisted on the need to discover new theory. But his road map was different from that of Cannan. He found that the marginal successors were far too much like the classical economists they followed. He wanted a turn to modern science, but a different kind of science: an empirical science unconstrained by a body of high theory. He said: 'the science must be cleared of all the theologico-metaphysical elements or tendencies which still encumber and deform it. Teleology and optimism on the one hand, and the jargon of "natural liberty" and "indefeasible rights" on the other, must be finally abandoned' (1888, p. 241). Instead, economics must become an experimental science 'forming only one department of the larger science of Sociology' (1988, p. 242). Only in this way could economists change 'the attitude of true men of science towards this branch of study, which they regard with ill-disguised contempt, and to whose professors they either refuse or very reluctantly concede a place in their brotherhood' (1988, p. 240).

Other contemporary interpretations of HET in the same tradition as Ingram, that economic ideas were necessarily embedded in economic history, were posited by Price (1891) and Ashley (1894).

Period IV. The Golden Age: HET as Heuristic Device

Beginning around the First World War and continuing for almost half a century, HET went through a remarkable transformation. After serving in the 19th century as little more than a minor weapon in the arsenals of combatants in one professional conflict or another, and appropriately consigned to prefaces and appendices by major figures and taken up extensively by no more than a few minor ones, HET came now to be pursued with energy and great seriousness by many of the leading figures in economics. Many of these converts produced significant book-length studies; others wrote articles. Some who did not devote years or an entire career to the subject still engaged in it soberly for the production of one or two studies before moving on. This new approach was not the 'throwaway HET' that had come before. Nor was it simply hagiography by members of a proud new community of professional economists. The authors in the golden age were committed to understanding problems through use of HET as an analytical device. They saw HET as heuristically significant. The golden agers did not think of HET as a separate new sub-discipline, as ultimately it was to become, but as an overlay of all economics, a distinct approach to all economic problems that should be explored as fully as other theoretical and empirical approaches. Moreover, the new interest was not confined to those holding any one ideological, methodological or doctrinal position. The following is an incomplete but illustrative list of some of those prominent economists who engaged in HET during this golden age apparently in search of answers to pressing questions: among the Austrian marginalists, J.A. Schumpeter, Gottfried Haberler, Karl Pribram, Erich Schneider, and Fritz Machlup; among English and American marginalists, John Hicks, Lionel Robbins, Frank Knight, George Stigler, and Jacob Viner; among the American Institutionalists, Wesley Mitchell, John R. Commons, Clarence Ayres, and John Kenneth Galbraith; among those intrigued with Marx, Eric Roll, Martin Bronfenbrenner, John Elliott, and Maurice Dobb; and among the new macroeconomists Piero Sraffa, G.L.S. Shackle, Gunnar Myrdal, and John Maynard Keynes himself. It was during this time that serious interpretive HET, rather than simply obituary notices, literature surveys, and review articles, entered the main publications of the profession, in writings by major figures such as those listed above, and lesser lights. HET was not only welcomed by the 'top' journals during the golden age, it became routinely the subject of presidential addresses and other ceremonial pronouncements. Most of the senior economists who took up HET also gave graduate courses in the field, and they encouraged some of their best graduate students to write dissertations in the area and to contemplate specializing in the field professionally.

Why this sudden turnabout? Why this unexpected fascination with history at the highest levels in the discipline? The most likely explanation lies in the circumstances of the time, which were certainly very different from those of the century before. Above all, a loss of confidence struck economics after the First World War. Before the war, economists of the mainstream such as Alfred Marshall, John Bates Clark, Léon Walras and Carl Menger concluded that they worked in an advancing science of a conventional sort and that they had the answers to most observable problems. The First World War, and the depression that followed, shattered all illusions that economic problems were that simple. No longer was it clear that relatively unconstrained rational men living in democracies and free market economies could count on enjoying peace and prosperity. The evidence seemed to prove the contrary and to suggest that all social constructs perfected during the Victorian age, including the global economy based on European empires, had to be re-examined from bottom up. Economics could not yet think of itself, as Keynes suggested it might be able to do some day, as analogous to dentistry seeking progress through technical improvements in familiar procedures. Where there had once been certainty now there was mainly doubt. And all of a sudden it seemed for many economists that HET might point the way toward undiscovered answers to some at least of the challenges newly arisen. HET was recognized as a vital tool in research. It could help economists find their bearings at several levels as they sought to be useful.

Another factor behind the new interest in HET may have been the kind of scholar attracted to the economics discipline at this time. The questions that were coming to the fore were not of a type that could be addressed effectively by narrow technicians, and the questions attracted persons who insisted on supplementing conventional economic analysis with philosophical, sociological, psychological and historical enquiry. So what were these questions that prominent economists came to believe might be tractable through HET? They were methodological, including, how to reconcile and integrate the approaches of the different national traditions of marginalist economics, for example British and American partial equilibrium with the general equilibrium of the Walrasians? Were mathematical economics and econometrics essential to progress within the discipline, and how should they be used? More generally, was it possible to retain under one disciplinary tent economists who were so different in their approaches and objectives as the varieties of neoclassical marginalists, Institutionalists, economic historians, Marxists, Keynesians and others? Was such heterogeneity virtue or vice? The questions were also theoretical; might early and forgotten theory cast light on such topics of sudden new concern as imperfect markets or business cycles? And some questions were directly policy oriented. What was the proper place for economics, and economists, in the policy process? Should the economists, rejecting the advice of most marginalist pioneers, sally forth from their ivory towers and connect directly with policymakers, perhaps even entering government as the German historians had done? If so, how? Should there be a ministry of economic affairs? Advisory councils to political leaders? Think tanks entirely outside of government? What about central planning? The Russian Revolution of 1917 raised this question for urgent public reconsideration even though it seemed to be settled for most professional economists by that date.

On all these questions, in contrast to the sense of self-confidence that characterized the first decade of the 20th century, when the most serious issues of economic policy were how to perfect the fine-tuning of the welfare economics of A.C. Pigou, the post-war mood demanded creative and fresh thinking. A notorious manifestation of this thinking across the disciplines was the hugely successful set of short biographies, *Eminent Victorians* (1918), by Lytton Strachey in which four prominent 19th-century institutions were held up for re-examination and reform: the military, the Church, the public schools and Victorian woman. Might this kind of historical enquiry reveal where the economy and economics had gone wrong, and show how they might be put back on the right track? Certainly Keynes believed so when he wrote *The Economic Consequences of the Peace* (1919), patterned substantially after *Eminent Victorians*.

Large structural questions without and within the economics discipline also were raised by the First World War and its aftermath. Was economics truly a science? This question became critical again during and after the Second World War, when public support for science, more than for other forms of enquiry, was contemplated and then implemented. These were years when the sub-disciplines of economics were just getting organized, and questions of boundaries and inclusions or exclusions had to be addressed. To some prominent scholars HET seemed a promising place to seek guidance. Jacob Viner's *Studies in the Theory of International Trade* (1937), Joseph Spengler's *French Predecessors of Malthus* (1942), Arthur Marget's *The Theory of Prices* (1938–42), Gottfried Haberler's *Prosperity and Depression* (1937), George Stigler's *Production and Distribution Theories* (1941) and Arthur Cole's *The Historical Development of Economic and Business Literature* (1957) were all milestones in HET and in the formation of the sub-disciplines of, respectively, international economics, economic demography, macroeconomics, industrial organization and management science. Not all of those who pursued HET in the golden age were the Renaissance men of the discipline. A few specialists did focus on single figures from the past, for example, Werner Stark on Bentham, Piero Sraffa on Ricardo, William Jaffé on Walras, and Joseph Dorfman on Veblen.

Not many of the giants of the golden age explained in detail the reasons for their new commitment to HET. Often the most we have to go on is an offhand remark or two. Jacob Viner said that his objectives were 'to resurrect forgotten or overlooked material worthy of resurrection, to trace the origin and development of the doctrines which were later to become familiar, and to examine the claims to acceptance of familiar doctrine' (Viner 1937, p. xiii). For Joseph Schumpeter the study of HET was an integral part of discovering a vision of economic evolution, which contained the key to understanding the economy (Schumpeter 1954). Frank Knight remarked that 'A major lesson to be learned from the history of ideas is to realize the 'glacial' tardiness of men, including the best minds, in seeing what it later seems should have been obvious at the first look' (1973, p. 46). Wesley Mitchell explored the question at some length at the start of his classes in HET at Columbia University and his reflections are revealing. In the transcription of his lectures, edited by Joseph Dorfman, Mitchell says that HET is necessary not so much to understand modern economics as to advance the subject through graduate education and research:

> All that I contend for is that so long as the social sciences continue to make progress each generation of economists will find problems in the history of their science which earlier generations have not thought out, and that these problems will probably attract workers who feel their fascination; that is, I think there is a difference between the social

> sciences and the natural sciences, which makes the past history of their subjects more interesting and more pertinent to the workers in the social field than to workers in the natural-science field.
>
> Our interest in the history of economics changes with the development of economics itself. The history of economics needs to be re-written by every generation of economists for the same reason that history at large needs to be re-written. (Mitchell 1967, p. 2)

Mitchell's point was an important one. He suggested that HET was valuable especially for graduate students and young scholars who had the responsibility ultimately to move economic science forward. Without historical sensibility graduate students would be at a serious disadvantage on the research frontier. Mitchell said that the HET he was teaching was fundamentally different from that which had come before:

> Working in this spirit we find ourselves concerned more with the larger aspects of economic history than our predecessors. What we can get light upon and what we therefore think most about is not the letter of the laws laid down, the traces of a man's thinking to be found in his predecessors, the logical inconsistencies which minute criticism may develop among his formulations – it is not these things which interest us so much as the type of problems the man attacks, his way of formulating them, what materials he had to work with, the general method he employed, the things he took for granted without inquiry, the grounds for the confidence he felt in his results, what use he put these results to, their acceptance or rejection by his contemporaries and the reaction of his scientific work upon social processes. (1967, pp. 6–7)

Mitchell suggested that HET should help economists gain 'knowledge of ourselves and free us from over-narrow specialization' (1967, p. 7). It would also 'give us clearer insight into the conditions which promote or retard the progress of knowledge in the social sciences. Perhaps some at least among these conditions will prove to be amenable to control' (1967, p. 7). HET might also give students the background with which to select among rival theoretical claims. 'Some of them become neo-Marshallians, some neo-Marxists, some neo-Austrians, some mathematical theorists, some institutionalists. If anyone is going to make any such choice he ought to make it with open eyes; i.e. he ought to understand what other types of theory are; what they offer. If he knows, perhaps he wont become an ardent follower of any school' (1967, p. 10). Finally, Mitchell noted that the sheer joy of historical inquiry should attract students to it. 'The fascination of the work itself, the possibility of gaining keener insights and more certainty as we follow up our leads, may have more to do with the future progress of such work than the indirect gains it promises for economic theory' (1967, p. 8).

This golden age of HET came to an end in the 1950s and 1960s. The cause of its death is as much a puzzle as its birth. One explanation could be that most of the leading figures retired or left the field. But that is a description of what happened more than an explanation. Why did these leaders not have successors? Why was not the next generation of leaders in economics fascinated in the same way by the history of their subject? The best explanation seems to be that by the 1960s economics had once again regained its self-confidence and there was a reversion to the set of attitudes that prevailed before the First World War. Most of the issues that appeared after the war (depression, doctrinal conflict, war itself) seemed either to be answered or to have gone away by the 1950s. There was no longer a need to look backwards, it seemed, only ahead. One of the most powerful forces leading to a high level of self-confidence in economics was its own performance during the Second World War compared with that during the First. Macroeconomic understanding proved helpful in maintaining full employment with price stability, while optimizing models taken directly from applied microeconomics and sometimes including the new tool of game theory were found to be useful in processes as different as aiming a machine gun and planning air raids.

Period V. Building a New Sub-discipline of HET

Most close observers of HET in the 1950s and 1960s might have predicted that its life within the economics discipline was over and that it was on its way to join the histories of other academic subjects in the deep recesses of history departments. At best it might leave a few champions within the larger discipline, such as Edwin Cannan proclaiming the faults of the old and the promise of the new. But this did not happen. HET lived on in economics, albeit without the powerful leaders of the golden age, without a place in most of the prominent research departments, and indeed without many opportunities for graduate training. So, without these assets how did the field survive? Several factors seem to have been in play.

The most important factor may have been the momentum carried over from the golden age. While most of those who had turned to history as a heuristic tool were gone by the 1960s, a few remained, and during this decade they joined in preparing a response to the new charge of irrelevancy. In the lead were George Stigler, Lionel Robbins, Terence Hutchison, Joseph Spengler, Joseph Dorfman and Martin Bronfenbrenner. Also sympathetic but less directly involved were Kenneth Arrow, Kenneth Boulding, James Buchanan, John Chipman, Earl J. Hamilton, Paul Samuelson and James Tobin. But more important than this rearguard action by the last golden agers was a cadre of young and middle-aged scholars trained in HET and committed now to retaining it within the economics discipline. These children of the golden age were well placed in teaching jobs and their careers had often been encouraged by their mentors. From being an overlay of the economics discipline during the golden age, HET moved during the 1960s and 1970s to become an independent sub-discipline, led by, among others: in Britain, R.D.C. Black, Mark Blaug, Tony Brewer, A.W. Coats, David Collard, Ronald Meek, Denis O'Brien, Andrew Skinner, and Donald Winch; in the USA and Canada, William Allen, William Barber, Hans Brems, Robert Ekelund, Frank Fetter, William Grampp, Samuel Hollander, Todd Lowry, Larry Moss, Mark Perlman, Warren Samuels, Robert Smith, Vincent Tarascio, Carl Uhr, Anthony Waterman and Donald Walker; in Israel, Haim Barkai and Ephraim Kleiman; John Pullen, Michael White, and Peter Groenewegen in Australia. Outside the English-speaking world leadership was taken by, among others, Pier Luigi Porta, Maria Cristina Marcuzzo, and Pierangelo Garegnani in Italy; Erich Streissler in Austria, Heinz Kurz, Harald Hagemann, and Bertram Schefold in Germany; Arnold Heertje in the Netherlands; Yuichi Shionoya and Takashi Negishi in Japan; and Lars Jonung and Bo Sandelin in Sweden. In addition to building and supporting the infrastructure of specialized periodicals and societies, such as *HOPE*, *JHET*, *EJHET*, and others, these scholars helped to mobilize and sustain a variety of other resources that have strengthened the field: translations and republications of canonical writings, collected works and letters of major authors, variorum editions, and ephemera, as in the Kress-Goldsmith micro-film project of works published before 1800. Collections of manuscripts of prominent economists, saved sometimes at the last minute from the garbage dump, made possible for the first time the close study of the interactions among economists and how they constructed their articles and books. The most substantial of these is the Economists' Papers Project at Duke University in the United States. In the United Kingdom the guide to archives prepared by Paul Sturgess documented where materials were located in that country. Access to manuscripts made possible meticulously documented biographies of great economists, for example, of Marshall by Peter Gronewegen (1995), of Hayek by Bruce Caldwell (2004) and of Keynes by Donald Moggridge (1992). Increasingly HET was defined as ending as recently as yesterday, and so oral history too became an essential tool of the historian.

An important movement that began in the 1960s was to explore ways in which HET could be incorporated more successfully into the curriculum of graduate students, economics majors, and even non-specialist liberal arts undergraduates. The teaching of HET in the golden age had been confined very largely to graduate and honours students using original sources and a few commentaries from the secondary literature. The textbooks that were available were by then very old – for example, those by Gray (1931), Gide and Rist (1909), and Haney (1911) – and not very appealing. The first rigorous new-style textbook, mainly for graduate students, was Mark Blaug's *Economic Theory in Retrospect* (1962). It concentrated on expressing old ideas in modern guise. Other similar texts that joined it over the years were by Hans Brems (1986) and Jurg Niehans (1990). A plethora of textbooks for undergraduate courses were published with styles, degrees of rigour, and ideologies for most tastes (for example, those by Landreth 1976; Ekelund and Hebert 1975; Rima 1967; and Spiegel 1971). One of the pioneering works in this genre was William Barber's *History of Economic Thought* (1967). An important publication landmark was Robert Heilbroner's *The Worldly Philosophers* (1953) which, with sales reputedly above a million copies, attracted generations of undergraduates to a more extended investigation of HET. Although leaders of the economics discipline in the years after the golden age expelled HET from the graduate curriculum (not even Blaug's new textbook could stem that tide), it was important for the employment prospects of those trained in HET that the appeal of the subject as an elective course for undergraduates remained.

Progress in research in HET since the 1970s has helped to sustain the positive response to the challenge of the 1950s and 1960s. The creation of a new sub-discipline was strengthened by the flush of interest in the philosophy of science in the 1970s. There were stimulating attempts to use new interpretive tools derived from the writings of Thomas Kuhn (1962), Imre Lakatos (1970), and others to understand the history of economics. And Deidre McCloskey's examination of the rhetoric of economics (1998) reverberates still in HET. Other substantial research projects that were a stimulus to the new sub-discipline of HET, both as inspiration and as source of consternation, include Samuel Hollander's reconsideration and reinterpretation of classical economics (1973, 1979, 1985, 1996), Philip Mirowski's exploration of the linkages between the history of economics and progress in other disciplines (1989), Roy Weintraub's account of the mathematization of economics (2002b), and studies of developments in modern economics by Mary Morgan (1990), Esther-Mirjam Sent (1998), Judy Klein (1997), and others.

The emergence of a new generation of leaders of HET in the decades after the golden age, leaders who were able to gain secure positions in colleges and universities, has been a reassuring development. These include Jurgen Backhaus, Roger Backhouse, Bradley Bateman, Peter Boettke, Mauro Boianovsky, Bruce Caldwell, Jose Luis Cardoso, Avi Cohen, David Colander, William Coleman, John Davis, Robert Dimand, Neil De Marchi, Ross Emmett, Jerry Evensky, Evelyn Forget, Dan Hammond, Wade Hands, Robert Hebert, Kevin Hoover, Sue Howson, John King, Judy Klein, Robert Leonard, John Lodewijx, Harro Maas, Steven Medema, Perry Mehrling, Don Moggridge, Mary Morgan, Malcolm Rutherford, Margaret Schabas, Neil Skaggs, Karen Vaughn and Jim Wible. Often these scholars have combined their interest in HET with commitment to another sub-field of economics, sometimes by keeping their interests in HET quiet until they achieved tenure. These grandchildren of the golden age, as it were, have kept the momentum for the perpetuation of the new sub-discipline alive into the 21st century.

Certain developments outside HET as well as within helped to strengthen the field in the latter decades of the 20th century. A number of distinguished economists moved to history rather late in their careers. Usually they addressed questions still alive in their original sub-disciplines, but they have employed the historian's tools and perspectives. Examples of these mid-career migrants to HET include Walter Eltis, Geoff Harcourt, Don Patinkin, David Laidler and John Whitaker.

A second kind of migrant has been more problematic for HET. When the homogenization of economics reached a crescendo in the 1980s and 1990s, some of those who felt alienated or squeezed out of the discipline for methodological or ideological reasons found comfort and welcome in HET. Some who resisted the increasing technical complexity of the new theory also sought refuge in this cross-over. These refugees, while providing welcome additions to the ranks of HET and offering different perspectives on a variety of issues, have tended to mark the entire sub-discipline as made up of malcontents.

A third kind of migrant to HET came from specialized communities within economics that had become too small or marginalized to continue on their own. They sought and received hospitality within HET whether their interests were primarily historical or not. They include some Marxists, neo-Austrians, Post Keynesians, Institutionalists, Sraffians, and others.

HET has been enriched in recent years by visits, short or long, from members of other disciplines who came not as refugees but attracted by specific research questions. They came from social, intellectual, political and economic history, as well as from sociology, philosophy and political science. Prominent visitors have included Peter Clark, Robert Skidelsky, Heath Pearson, S.M. Amadae and Yuval P. Yonay.

The Prospect Ahead

The future of HET is uncertain (Weintraub 2002b). On the one hand, the strong infrastructure of societies and publications is encouraging, as are the numbers of scholars who identify with the field. It is gratifying, moreover, that the field has demonstrated persuasively its capacity to survive adversity and to face challenges constructively. But though these are reasons for optimism for the future there are reasons also for unease. And this leads to a final question. What uses will be found for HET in the future, and can any of these be discerned from study of the past? The original use for HET in the rhetoric of policy debates persists, but mainly on the surface. Libertarians wear Adam Smith ties and opponents of an active government in the economy dismiss their opponents collectively as Keynesians, but in both cases the combatants understand little beyond the labels. HET as doctrinal cleansing is still performed, but mainly in review articles and chapters, such as those in the *Journal of Economic Literature* and the various Handbook series, prepared not by specialists in HET but by high priests of the various sub-disciplines. The more focused and celebratory literature reviews, such as those that gained popularity after the marginal revolution, can be found still in Nobel Prize acceptance speeches and presidential addresses, but neither serious history nor professional historians of economic thought are much involved. In this spirit are the innumerable biographical and hagiographic dictionaries of ‘great economists’ categorized in various ways, as women, dissenters, or something else. The use for HET which was its greatest strength during the golden age, in the training of graduate students and in the search for answers to large questions on the research frontier, has largely disappeared, and there seems no immediate prospect of it being resurrected. Among the more recent uses for HET as a home for refugees of various kinds and as a component in the undergraduate curriculum to relieve the tedium of increasingly technical abstraction, only the latter seems secure and likely to grow in strength.

The overriding question remains: can a sub-discipline survive for long when it is little valued by the discipline of which it is a part and where there is no graduate training available through which to sustain and renew the leadership? One bright spot may be the liberal arts college, where breadth as well as depth is still rewarded and which is likely to express forcefully in the labour market its preferences for kinds of faculty training. Or it may take another loss of confidence within the economics discipline overall, such as that experienced early in the 20th century, to cause economists to find once again something of relevance in their past!

Bibliography

Ashley, W.J. 1894. *An introduction to english economic history and theory.* 3rd edn. London: Longmans Green.

Backhouse, R.E. 2000. *Early histories of economic thought 1824–1914*. London: Routledge.

Backhouse, R.E. 2004. History of economics, economics and economic history in Britain, 1824–2000. *European Journal of the History of Economic Thought* 11: 107–124.

Barber, W. 1967. *A history of economic thought*. New York: Penguin.

Blanqui, A. 1837. *Histoire de L’Economie Politique*. Paris: Guillamin.

Blaug, M. 1962. *Economic theory in retrospect*. Homewood: Irwin.

Blaug, M. 1991. *The historiography of economics*. Aldershot: Edward Elgar.

Böhm-Bawerk, E.V. 1973. The Austrian economists. In *Dictionary of the history of ideas*, ed. P. Wiener. New York: Charles Scribner's Sons.

Brems, H. 1986. *Pioneering economic theory, 1630–1980*. Baltimore: Johns Hopkins University Press.

Caldwell, B. 2004. *Hayek's challenge*. Chicago: University of Chicago Press.

Cannan, E. 1917. *A history of the theories of production and distribution in english political economy 1776–1848*, 1994. Bristol: Thoemmes.

Cole, A.H. 1957. *The historical development of economic and business literature*. Boston: Harvard Graduate School of Business Administration.

Cossa, L. 1876. *Guida allo studio dell'economica politica*. Milano: Ulrico Hoepli. (Eng. trans. 1880.)

Cournot, A. 1897. *Researches into the mathematical principles of the theory of wealth*, 1963. Homewood: Irwin.

Du Pont De Nemours. 1768. *De L'Origine et des progrès d'une science nouvelle*, 1910. Paris: Geuthner.

Eckland, R.B., and R.F. Hebert. 1975. *A history of economic theory and method*. New York: McGraw-Hill.

Gide, P.H.C., and C. Rist. 1909. *Histoire des doctrines économiques*. Paris: L. Larose and L. Tenin.

Gray, A. 1931. *The development of economic doctrine*. London: Longmans Green.

Gronewegen, P.D. 1995. *A soaring eagle*. London: Edward Elgar.

Haberler, G. 1937. *Prosperity and depression*. Geneva: League of Nations.

Haney, L. 1911. *History of economic thought*. New York: Macmillan.

Heilbroner, R. 1953. *The worldly philosophers*. New York: Simon and Schuster.

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

Hollander, S. 1973. *The economics of Adam Smith*. Toronto: University of Toronto Press.

Hollander, S. 1979. *The economics of David Ricardo*. Toronto: University of Toronto Press.

Hollander, S. 1985. *The economics of John Stuart Mill*. Toronto: University of Toronto Press.

Hollander, S. 1996. *The economics of Thomas Robert Malthus*. Toronto: University of Toronto Press.

Ingram, J.K. 1888. *A history of political economy, with a preface by E.J. James*. New York: Macmillan.

Jevons, W.S. 1871. *The theory of political economy*, 1965. New York: Augustus M. Kelley.

Keynes, J.M. 1919. *Economic consequences of the peace*, 1920. New York: Harcourt Brace.

Klein, J. 1997. *Statistical visions in time*. Cambridge: Cambridge University Press.

Knight, F. 1956. *On the history and method of economics*. Chicago: University of Chicago Press.

Knight, F. 1973. Economic history. In *Dictionary of the history of ideas*, ed. P. Wiener. New York: Charles Scribner's Sons.

Kuhn, T.S. 1962. *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Lakatos, I., and A. Musgrave. 1970. *Criticism and the growth of knowledge*. Cambridge: Cambridge University Press.

Landreth, H. 1976. *History of economic thought*. Boston: Houghton Mifflin.

Marget, A.W. 1938–1942. *The theory of prices*. New York: Prentice-Hall.

Marshall, A. 1920. *Principles of economics*. 8th edn, 1964. London: Macmillan.

McCloskey, D. 1998. *The rhetoric of economics*. 2nd edn. Madison: University of Wisconsin Press.

McCulloch, J.R. 1825. *The principles of political economy*, 1830. London: William.

McCulloch, J.R. 1845. *The literature of political economy*, 1938. London: London School of Economics and Political Science.

Menger, C. 1950. *Principles of economics*. Trans. and ed. J. Dingwall and B.F. Hoselitz. Glencoe: Free Press.

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

Mirowski, P. 1989. *More heat than light*. Cambridge: Cambridge University Press.

Mitchell, W.C. 1967. *Types of economic theory*. New York: Augustus M. Kelley.

Moggridge, D. 1992. *Maynard Keynes: An economist's biography*. New York: Routledge.

Morgan, M. 1990. *The history of econometric ideas*. Cambridge: Cambridge University Press.

Niehans, J. 1990. *A history of economic theory*. Baltimore: Johns Hopkins University Press.

Price, L.L. 1891. *A short history of political economy in England, from Adam Smith to Arnold Toynbee*. London: Methuen.

Rima, I. 1967. *Development of economic analysis*. Homewood: R.D. Irwin.

Samuelson, P.A. 1947. *Foundations of economic analysis*. Cambridge, MA: Harvard University Press.

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

Senior, N. 1827. *An introductory lecture on political economy*. London: J. Mawman.

Sent, E.-M. 1998. *The evolving rationality of rational expectations*. Cambridge: Cambridge University Press.

Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*, 1976. Oxford: Clarendon.

Spengler, J.J. 1942. *French predecessors of Malthus*. Durham, NC: Duke University Press.

Spiegel, H. 1971. *The growth of economic thought*. Durham, NC: Duke University Press.

Stark, W. 1994. *History and historians of political economy*. New Brunswick: Transaction.

Stigler, G.J. 1941. *Production and distribution theories, 1870 to 1895*. New York: Macmillan.

Strachey, L. 1918. *Eminent victorians*. London: Chatto and Windus.

Twiss, T. 1847. *View of the progress of political economy in Europe since the sixteenth century*. London: Longman.

Villeneuve-Bargemont, J.-P.-A. 1841. *Histoire de l'économie politique*. Paris: Guillaumin.

Viner, J. 1937. *Studies in the theory of international trade*, 1955. London: George Allen and Unwin.
Weintraub, E.R., ed. 2002a. *The future of the history of economics*. Durham, NC: Duke University Press.
Weintraub, E.R. 2002b. *How economics became a mathematical science*. Durham, NC: Duke University Press.

History of Forward Contracts (Historical Evidence for Forward Contracts)

Kim Oosterlinck

Abstract

A forward contract is an agreement between two parties who specify today the terms (price, underlying asset, quantity, etc.) of an exchange that is to take place at a known future date. Forward contracts can be traced to Greek and Roman times, and may have occurred earlier still; they have been widely traded in Europe (and subsequently elsewhere) since the Middle Ages.

Keywords

Forward contracts; Futures

JEL Classifications

F30; G10; N20; N23

A forward contract is an agreement between two parties who specify today the terms (price, underlying asset, quantity, etc.) of an exchange that is to take place at a known future date. The creation of forward contracts logically followed the geographical extension of trade. Indeed, as the distance to be covered increased, delays in deliveries occurred more and more frequently. In order to hedge the exposure to price changes implied by these delays, merchants entered into forward contracts.

Because of the scarcity of ancient sources, it is hard to determine exactly where and when forward contracts were born. Transactions exhibiting forward features existed in ancient Mesopotamia, and during Greek and Roman times, although the precise terms of the contract often remain unclear (Poitras 2000). During the Middle Ages, the use of forward contracts experienced a huge growth. Urban merchants took long positions on the future delivery of fishes before they were caught, or crops before they were harvested. As time went by, the need for an organized market for forward contracts became acute. Their trade soon took place in the most active markets: Antwerp (16th century), Amsterdam (17th century) and London (18th century). By the end of the 19th century, these contracts were commonly quoted all around the world.

On the Antwerp stock exchange, the world's leading bourse in the 16th century, 'to arrive' contracts, for which records exist since 1511, were created. This business expended dramatically during the 16th century. The contracts were mainly on grain, but also on salt or herrings. From 1589 to 1594, the accounts of Della Faille, a large Antwerp-based firm, exhibit sales amounting to 385 524 guilders of which 75 per cent were forward contracts with a three to six-month maturity. The military fall of Antwerp in 1585 brought a local end to these activities and their transfer to Amsterdam (Poitras 2000).

In Amsterdam, forward contracts, called 'deals for time' or 'time bargains', were first carried out on grain and herrings, later on whale oil, brandy or colonial goods (spices, pepper, cocoa, coffee), and eventually on financial securities. During the 16th century, a market appeared where investors could bet on the future exchange rate and only pay the spread at maturity. The 17th century saw the emergence of an active market for securities. The most liquid ones, such as shares in VOC (the Verenigde Oostindische Compagnie, or Dutch East Indian Company, founded in 1602), were the first ones to experience forward trade. Speculation developed quickly. For example, in 1609 a 'bearish consortium' under the direction of Isaac le Maire tried to secure profits by shorting VOC shares and propagating negative rumours about the company (Van Dillen 1927; Gelderblom and Jonker 2005). Regarding the contract features, the delay to maturity was usually much longer than nowadays, often exceeding one year.

Because of the large amounts involved (the average cost of a VOC share was 3000 guilders), it was common to pay only the difference at maturity. A quarterly settlement operation, called 'rescontre', was held on February, May, August and November. During these meetings, bullish speculators could prolong their position to the next deadline by paying an interest ranging from one to three per cent. Without clearing houses and regulation, the execution of the forward contracts depended on the parties' good faith. In fact, the only sanction for repudiating investors was the exclusion from the exchange. However, deposits in cash were sometimes asked for.

In London, 'dealings for time' started at the end of the 17th century. A few years later, forward contracts were mostly traded at Jonathan's coffee house. The London stock exchange borrowed both terminology and practices from Amsterdam. Mortimer, a contemporaneous writer, provides a description of the terminology used at the time, as he tells the story of a broker wishing to buy 'd 1000 3 per cent annuities for the May "rescounter"' (Cope 1978). As in Amsterdam, rollover was frequent and forward contracts were largely dominated by speculators. Differences would nonetheless appear between the two exchanges. Around 1745, the concept of backwardation is mentioned in London. By this date, the frequency of the 'rescounters' was quarterly, but it would change to six-weekly less than 40 years later. Settlement arrangements were not organized and bargains had to be taken care of by individual brokers.

Shortly after their creation, forward contracts faced strong opposition. Considered as bets, they fell under gambling legislation and were stigmatized as immoral. The only contracts that escaped such laws were the ones for which the seller owned the underlying asset and intended to deliver it at maturity. Famous legal examples include the Dutch edict from 1610 banning 'windhandel' (Garber 2001; Gelderblom and Jonker 2005), or trade in the wind, and the British Barnard Act from 1734. Despite the economic importance of forwards, restrictions were still largely applied in several European countries during the 19th century.

Many references dealing with their historical evidence do not distinguish forwards from futures. Forwards are often labelled as futures, but the reverse is unusual since, before the 19th century, there were no futures markets in the modern sense of the term. (A possible exception is the Japanese Dojima rice markets under the Tokugawa period, 1603–1867.) The appearance of futures during the 19th century did not bring an end to forward contracts, which have remained widely exchanged.

See Also

- ► Futures Markets, Hedging and Speculation

Bibliography

Cope, S. 1978. The stock exchange revisited: A new look at the market in securities in London in the eighteenth century. *Economica* 45(177): 1–21.

Einzig, P. 1964. *The history of foreign exchange*. London: Macmillan.

Garber, P.M. 2001. *Famous first bubbles: The fundamentals of early manias*. Cambridge, MA: MIT Press.

Gelderblom, O., and J. Jonker. 2005. Amsterdam as the cradle of modern futures and options trading. In *The origins of value: The financial innovations that created modern capital markets*, ed. W.N. Goetzmann and K.G. Rouwenhorst. Oxford: Oxford University Press.

Poitras, G. 2000. *The early history of financial economics, 1478–1776: From commercial arithmetic to life annuities and joint stocks*. Cheltenham/Northampton: Edward Elgar.

Schaede, U. 1989. Forwards and futures in Tokugawa-period Japan: A new perspective on the Dojima rice market. *Journal of Banking and Finance* 13(4–5): 487–513.

Van Dillen, J.G. 1927. Termijnhandel te Amsterdam in de 16de en 17de eeuw. *De Economist* 76: 503–523.

Hobbes, Thomas (1588–1679)

C. B. Macpherson

The greatest English political theorist and philosopher, Hobbes was born at Malmesbury and died at Hardwick, the seat of the Earl of Devonshire, who had been Hobbes's patron for many years.

After attending Magdalen Hall, Oxford (BA 1608), Hobbes entered the Devonshire household as tutor to the son, and made several trips to the Continent, on one of which (in 1636) he conversed with Galileo, whose resolutive-compositive method Hobbes took over, and whose laws of motion he later carried over and applied to the motions, internal and external, of men. In 1640, fearing that his earliest work would offend the Long Parliament, he went into voluntary exile in Paris, where for a time (1646–8) he tutored the future Charles II in mathematics. He returned to England in 1651 and from then on lived as inconspicuously as he could.

Economic insights are to be found in his three main works of political theory, *The Elements of Law, Natural and Politic* (1640); *De Cive* (1642), translated (by Hobbes) as *Philosophical Rudiments Concerning Government and Society* (1651); *Leviathan* (1651); and in his history of the Long Parliament and the Civil War, *Behemoth* (1682). Hobbes's great work was his political science, of which his economic ideas seem to be only an incidental part. Yet we may notice that his political edifice rested on economic assumptions, in that his model of society was the atomistic bourgeois market society whose seismic rise in England in his own time he had certainly noticed. However, he did not attempt anything along the lines of the classical political economy of the 18th century, or even of the political arithmetic of his own century: he offered neither a general theory of exchange value nor a theory of distribution, that is, of the determinants of rent, interest, profits and wages, nor even a theory of the balance of trade or of foreign exchange. But he did set down a few general economic principles. One is a supply and demand theory of exchange value, as in: 'The value of all things contracted for, is measured by the Appetite of the Contractors' *(Leviathan,* ch. 15, p. 208) and in his more striking statement

> The *Value* or WORTH of a man, is as of all other things, his Price; that is to say, so much as would be given for the use of his Power: and therefore is not absolute; but a thing dependent on the need and judgement of another . . . And as in other things, so in men, not the seller, but the buyer determines the Price . . . (ibid., ch. 10, pp. 151–2).

The two statements are consistent only on the assumption of an endemic surplus of wage-labourers, an assumption which Hobbes did explicitly make. The able-bodied poor, who were expected to increase indefinitely,

> are to be forced to work: and to avoyed the excuse of not finding employment, there ought to be such Lawes, as may encourage all manner of Arts; as Navigation, Agriculture, Fishing, and all manner of Manifacture that requires labour. The multitude of poor, and yet strong people still encreasing, they are to be transplanted into Countries not sufficiently inhabited: where neverthelesse, they are not to exterminate those they find there; but constrain them to inhabit closer together, and not range a great deal of ground, to snatch what they find; but to court each little Plot with art and labour, to give them their sustenance in due season (Leviathan, ch. 30, p. 387; cf. Behemoth, p. 126).

Another general proposition is that 'a mans Labour also, is a commodity exchangeable for benefit, as well as any other thing' (*Leviathan,* ch. 24, p. 295).

More important than such general principles are his many policy recommendations to the Sovereign, all of which are designed to increase the wealth of the nation by promoting the accumulation of capital by private enterprisers seeking their own enrichment. Typical are his recommendations about taxation. Taxes are justified only because they provide the income which enables the sovereign power to maintain the conditions for private enterprise: 'the Impositions that are layd on the People by the Soveraign Power, are nothing else but the Wages, due to them that hold the publique Sword, to defend private men in the exercise of severall Trades, and Callings' (ibid., ch. 30, p. 386). Taxes on wealth are bad, for they discourage accumulation. The best taxes are those on consumption, which discourage 'the luxurious waste of private men' (p. 387). Hobbes's recommendations to the Sovereign all follow from his most general rule, as set out in the opening paragraph of chapter 30:

> The office of the Soveraign, (be it a Monarch, or an Assembly,) consisteth in the end, for which he was trusted with the Soveraign Power, namely the procuration of the safety of the people . . . But by Safety here, is not meant a bare Preservation but also all other Contentments of life, which every man by

> lawfull Industry, without danger, or hurt to the Common-wealth, shall acquire to himself (p. 376).

Most important of all was his insistence that the sovereign was above the law and could not be limited by any of the traditional rights of leasehold or copyhold tenants, or by any traditional limits on market transactions, or traditional protections of the poor: 'it belongeth to the Common-wealth, (that is to say, to the Soveraign), to appoint in what manner, all kinds of contract between Subjects, (as buying, selling, exchanging, borrowing, lending, letting, and taking to hire), are to bee made; and by what words and signes they shall be understood for valid' (ibid., ch. 24, p. 299).

In short, the job of the state was to clear the way for capitalism. It is evident that Hobbes's doctrine was particularly appropriate to the period of primary capital accumulation. It is scarcely too much to say that it was his perception of the needs of such a period which determined the main lines of his political theory. What was needed was a sovereign powerful enough to override all the protections of the common law, and, to justify such a power, a new, untraditional basis for political obligation. That is what Hobbes's doctrine provided. In effect, it is the legitimation of the early capitalist state.

Selected Works

1650. *Elements of law natural and politic,* ed. F. Tonnies. Cambridge: Cambridge University Press, 1928.

1651. *Philosophic rudiments concerning government and society.* Published as *De cive or the citizen*, ed. S.P. Lamprecht. New York: Appleton-Century-Crofts, 1949.

1651. *Leviathan*, ed. C.B. Macpherson. Harmondsworth: Penguin, 1968.

1682. *Behemoth; or the long parliament*, ed. F. Tonnies. London: Simpkin, Marshall & Co., 1889.

References

Macpherson, C.B. 1983. Hobbes's political economy. *Philosophical Forum* 14 (3–4).

Hobson, John Atkinson (1858–1940)

Peter Clarke, Roger E. Backhouse and P. J. Cain

Abstract

John Atkinson Hobson, a self-styled economic heretic, had a long and prolific career as an economist and political activist. His heresies included underconsumptionism and a critique of orthodox welfare economics based on ideas from John Ruskin, the former being elaborated into a theory of imperialism that influenced Lenin. He was belatedly recognized as a forerunner by Keynes in his General Theory, but this does not do justice to the range of Hobson's work.

Keywords

Accelerator; Clark, J. B.; Effective demand; Fabian economics; Forced gains; Hobson, J. A.; Imperialism; Involuntary unemployment; Keynes, J. M.; Lenin, V. I.; Living wage; Marginal productivity theory; Marginalism; Mill, J. S.; Pigou, A. C.; Productive and unproductive surplus; Progressive and regressive taxation; Protection; Quantity theory of money; Rent; Surplus value; Underconsumptionism

JEL Classifications

B31

John Atkinson Hobson was born in Derby in 1858 and died at home in Hampstead in 1940. He was educated at Derby School and Lincoln College, Oxford, where he read Greats from 1876 to 1880, but only gained a Third. He taught classics at Faversham and Exeter in 1880–81, before moving to London, where he supplemented his private income (from the Derby newspaper which his father had owned) with intermittent earnings from journalism, lecturing and his books (Clarke 1978). A prolific writer, he propagated his economic views through more than 50 books and

700 articles, many of them in a series of organs of radical liberal and socialist leanings. Hobson thus left an oeuvre which is not easy to assess and in which formal inconsistencies are not difficult to find: but he conveys, nonetheless, a general vision of the scope and nature of economics that is both distinctive and coherent. His reputation has been coloured by his supposed role as a predecessor not only of Lenin and his theory of imperialism but also of Keynes and his concept of effective demand. Neither connection is wholly factitious but both have been open to unhistorical distortions of Hobson's own concerns.

Hobson has long been best known as an underconsumptionist. His first book (Mummery and Hobson 1889) was written in collaboration with A.F. Mummery, a businessman, who seems to have been the senior partner. The book set out to expose fallacies in classical political economy as expounded by J.S. Mill. Its central proposition was that trade depression was caused by a deficiency in effective demand since it was the level of consumption in the immediate future that limited profitable production. It followed that there was a limit to the amount of useful savings which a community could make. Each individual could save with advantage to himself, but the overall result might be a position of underconsumption, for which over-saving was another name. Hobson was to seize on this self-defeating process as an example of what he called the protean fallacy of individualism – an idea that pervades his work in a far more general way than the particular concept of underconsumption. The polemical thrust of this early book was thus against the tendency of economists to extol thrift in so far as this neglected the crucial importance of maintaining sufficient demand. Hobson and Mummery provided an account (complete with a numerical example) of the accelerator, a concept commonly believed to have originated in the 20th century (1889, pp. 85–6; cf. Backhouse 1990). Though the book attracted hostile comment from established economists, it did not, as Hobson alleged, blight his career. He carried on teaching economics as a university extension lecturer, the job for which he was well suited temperamentally (Kadish 1990). Later, he was proud to proclaim himself an 'economic heretic' (Hobson 1938).

This early statement of the underconsumptionist case was reiterated in two further books (Hobson 1894, 1896) the second of which made use of the newly coined term 'unemployment', defining it in terms of involuntary leisure suffered by the working classes. He broadened rather than narrowed his dissent from neoclassical analysis through his distrust of marginalism, which he rejected on the ground that it rested upon an unreal individualism, marking a further breach with Marshallian orthodoxy (Hobson 1901b, 1926a). A later book (Hobson 1913), which was savagely reviewed by J.M. Keynes, sought to expose the errors of the quantity theory of money, recently popularized by Irving Fisher: this shows the extent to which Hobson was still thinking as a classical economist brought up on Mill, failing to fully take account of the innovations of his contemporaries such as Marshall and Fisher (Backhouse 1990).

Hobson was to supplement his account of underconsumption with a theory of distribution (Hobson 1900) which drew heavily upon the Fabian theory of rent. This theory built on a marginal productivity theory of distribution that had first been published in 1891 in the *Quarterly Journal of Economics,* alongside John Bates Clark's article on the same subject. Hobson distinguished the costs of subsistence for any factor of production from its rent element, and argued that in principle surplus value might accrue to land, labour or capital. He further introduced the idea of 'forced gains' as an assertion of superior bargaining power in this process, with the result that 'unearned income' accrued to certain individuals and classes. He also assumed that the proportion of income which was in this sense economically functionless varied directly with the absolute level of income received. It followed that progressive taxation would not in practice impair any necessary incentive to production.

This analysis was later elaborated (Hobson 1909b) to distinguish a 'productive surplus' that covered the costs of growth from an 'unproductive surplus', distributed according to no functional principle. Morally this was the property of the community which had created it. If redistributive taxation could restore it to its rightful possessors, over-saving by the rich would be curtailed and underconsumption by the poor rectified. This functional view of the proper working of the economic system, with effort matched to reward by rooting out parasitism, reappears constantly as a paradigm in Hobson's writings. He dignified it with the name 'the organic law' and often suggested an evolutionary provenance for it. But he also claimed the authority of John Ruskin, of whom he wrote an admiring study (Hobson 1898), for seeing consumption, not production, as the qualitative end of economic activity. He sought to unite these ideas in one of the most frequently reprinted of his books (Hobson 1894) by adopting the formula: 'From each according to his powers, to each according to his needs.'

Hobson's view, taken from Ruskin, that attention should be focused on the human cost of economic activity was the basis for Work and Wealth: A Human Valuation (Hobson 1914), which offered a systematic response to Pigou's welfare economics, the first systematic exposition of which had been published two years earlier. As in his writings on underconsumption, and distribution, he adopted terminology that emphasized, and possibly exaggerated, his differences with orthodoxy. Resting on clear value judgements about the worth of different activities, such an approach fell out of favour in the 1930s, and even before that failed to dislodge the Cambridge approach, especially in Britain. However, his work was much better received in the United States, where he had significant personal connections and where some institutionalists considered him the leading representative of English welfare economics.

In the early 1890s, Hobson was inclined to believe that protection and economic imperialism could mitigate underconsumption. As his political radicalism intensified, however, he dismissed protection as a device for safeguarding the incomes of the wealthy, thereby aggravating the problem of over-saving. In the wake of the scramble for China and the outbreak of the South African War (1899–1902) Hobson also developed a novel theory of economic imperialism. He identified speculative investment in undeveloped territories as a cause of imperialism and claimed that it arose from over-saving by a parasitic class at home. In this sense underconsumption was the economic taproot of imperialism (Hobson 1902). What he vigorously rejected was the proposition that there was sufficient profit to the country as a whole from trade and investment in Africa to counterbalance the costs of aggression. In contrast to Lenin, therefore, Hobson denied that imperialism was a structural necessity of the metropolitan economy. It could and should be checked at home by a policy of redistributive taxation, which would have the reciprocal effect of cutting the taproot (ending over-saving) and stimulating domestic demand (ending underconsumption).

The economic implication was that Britain could easily make up any loss on foreign trade by generating wealth at home – an argument that could be used by protectionists. Nonetheless, it was the Liberal and Labour Parties, with their commitment to free trade, to which Hobson looked for reformist amelioration. He was confident that imperialism could be beaten by democratic means precisely because it did not serve the interests of the majority but only of a privileged section of the nation. In his most famous book, therefore, Hobson devotes more than twice as much space to the politics than to the economics of imperialism (Hobson 1902). He needed to do so because the puzzle was how a policy that was bad business for the nation as a whole had come to be adopted. The answer was that finance was the 'governor' of an engine whose motor power came from the forces of nationalism and social psychology that fuelled the politics of self-assertion (Hobson 1901a). His analysis of imperialism changed over time and was often strongly coloured by passing political events. In at least one book (Hobson 1911) he commended cosmopolitan finance as a force for peace and saw imperialism as a step on the road to world economic development. During the First World War, he made a partial return to his earlier views and between the wars his position was often an uneasy compromise between the stances adopted in 1902 and 1911. The fact that he chose to republish *Imperialism: A Study* in 1938 virtually unaltered obscured the complexity of his response to empire (Cain 2002).

It will be apparent that Hobson was no single-minded underconsumptionist. In the early 1900s his energies were directed towards permeating the Liberal Party with a broad-based conception of economics that would justify it in rejecting the classical nostrums of laissez-faire in favour of interventionist policies designed to further social justice (Hobson 1909b). The publication of Hobson's *The Industrial System,* which consolidated much of his previous work, opportunely coincided with Lloyd George's People's Budget of 1909 and offered a defence of the policy of redistributive taxation via the concept of the surplus. This aspect overshadowed the restatement of Hobson's underconsumptionist position; though he now went further than before in analysing the dynamic process by which over-saving reduced all real incomes in the economy until automatic checks came into play (Hobson 1909b, ch. 18). One might call this Hobson's most accomplished exercise in macroeconomics.

It was in the context of the depression after the First World War that Hobson once more returned to this theme (Hobson 1922, 1930), and it was in this period that his economic views enjoyed greatest publicity. He was now loosely identified with the Labour Party and found a natural application for his ideas in mounting an economic case for a 'living wage' (Hobson 1926b). His central contentions on oversaving continued to be refined (King 1994) and, amid widespread unemployment, they found a more sympathetic response, even among professional economists who had previously accepted a full-employment assumption. In particular, by 1930 Hobson was on cordial terms with J.M. Keynes, who had in earlier years scorned his work. But Keynes was still anxious to keep his distance, as he made clear (Keynes 1930, pp. 160–1). The reason was that when Keynes wrote of over-saving he meant under-investment; whereas for Hobson saving and investment were two names for the same thing, and by over-saving he had always meant under-spending. It followed also that Keynes had more interest in policies of public works as a means of promoting investment, whereas Hobson concentrated on the case for redistribution as a means of stimulating consumption. It was not until Keynes had virtually finished the General Theory that he fully realized that he had done Hobson and Mummery an injustice; and so he paid them a handsome, if belated, tribute (Keynes 1936, pp. 364–71).

See Also

- ▶ Clark, John Bates (1847–1938)
- ▶ Keynes, John Maynard (1883–1946)
- ▶ Lenin, Vladimir Ilyich [Ulyanov] (1870–1924)
- ▶ Pigou, Arthur Cecil (1877–1959)
- ▶ Underconsumptionism

Selected Works

1889. (With A.F. Mummery.) *The physiology of industry.* London: Murray.

1894. *The evolution of modern capitalism*. London: Walter Scott.

1896. *The problem of the unemployed*. London: Methuen.

1898. *John Ruskin, social reformer*. London: Nisbet.

1900. *The economics of distribution*. New York: Macmillan.

1901a. *The psychology of Jingoism*. London: Nisbet.

1901b. *The social problem.* London: Nisbet.

1902. *Imperialism: A study.* London: Nisbet.

1909a. *The crisis of liberalism,* ed. P.F. Clarke. Brighton: Harvester Press, 1974.

1909b. *The industrial system.* London: Longman.

1911. *An economic interpretation of investment*. London: Financial Review of Reviews.

1913. *Gold, prices and wages.* London: Methuen.

1914. *Work and wealth.* London: Macmillan.

1919a. *Taxation in the new state*. London: Methuen.

1922. *The economics of unemployment*. London: Macmillan.

1926a. *Free thought in the social sciences*. London: Allen & Unwin.

1926b. (With H.N. Brailsford, A. Creech Jones, and E.F. Wise.) *The living wage.* London: Independent Labour Party.

1930. *Rationalisation and unemployment*. London: Allen & Unwin.

1938. *Confessions of an economic heretic.* With an introduction by M. Freeden. Brighton: Harvester Press, 1976.

Bibliography

Backhouse, R.E. 1990. J.A. Hobson as a macroeconomic theorist. In Freeden (1990).

Cain, P.J. 2002. *Hobson and imperialism: Radicalism, new liberalism and finance, 1887–1938*. Oxford: Oxford University Press.

Clarke, P.F. 1978. *Liberals and social democrats*. Cambridge: Cambridge University Press.

Freeden, M. 1990. *Reappraising J. A. Hobson*. London: Unwin Hyman.

Kadish, A. 1990. Rewriting the confessions: Hobson and the extension movement. In Freeden (1990).

Keynes, J.M. 1930. *A treatise on money, vol. 1: The pure theory of money.* London: Macmillan for the Royal Economic Society, 1971.

Keynes, J.M. 1936. *The general theory of employment, interest and money.* London: Macmillan for the Royal Economic Society, 1973.

King, J.E. 1994. J. A. Hobson's macroeconomics: The last ten years. In Pheby (1994).

Pheby, J. 1994. *J. A. Hobson after fifty years*. Basingstoke: Macmillan.

Hodgskin, Thomas (1787–1869)

N. W. Thompson

Thomas Hodgskin joined the navy at the age of twelve, rose to the rank of lieutenar and was then forcibly retired on half pay after a contretemps with his authoritarian captain. On the advice of Francis Place he subsequently embarked, in 1815, upon a continental tour with the object of collecting material on the social and economic conditions of post-Napoleonic war Europe. It was this that formed the basis of his *Travels in the North of Germany*, which was published in 1820.

It was in the early 1820s that his sympathy for the working classes was aroused through his involvement in the struggle for the repeal of the Combination Acts and the attempts to establish a London Mechanics' Institute. The former led to the publication of Hodgskin's most famous work, *Labour Defended against the Claims of Capital* (1825), while the latter led to a series of lectures, given at the Institute, which formed the basis of *Popular Political Economy* (1827). It was these two works which, together with his *Natural and Artificial Right of Property Contrasted* (1832), established his reputation as one of the major nineteenth-century anti-capitalist writers. After 1832 financial necessity forced Hodgskin to abandon his more serious intellectual labours and to concentrate upon a journalistic career that had begun in the early 1820s and which lasted until the early 1850s when he worked for the *Economist*.

While generally dubbed a 'Ricardian socialist', the single most important influence upon the thought of Thomas Hodgskin was Adam Smith. From Smith he believed he had derived the central tenet of his social and economic philosophy, namely that the material world was shaped by natural laws emanating from an omniscient and beneficent Providence, all interference with which was either superfluous or pernicious. Such views are apparent in Hodgskin's *Travels in the North of Germany*, where he attacked the malign interference of government with the natural laws that should be left to regulate trade and industry, and they were to provide the philosophical underpinning of all his major works.

Thus in *Labour Defended*, written in defence of trade union activity, Hodgskin attacked profit, the reward of the capitalist, as a violation of the natural laws of value and distribution. Here Hodgskin confronted the classical argument that the capitalist derived his entitlement to a share in labour's product from his ownership of fixed and circulating capital which he provided for the use of his workforce. The idea of a fund of circulating capital Hodgskin dismissed as a fiction. What labour depended upon during the period necessary to make and bring a commodity to market was co-existing labour, while fixed capital was simply the result of past exertions, utilized, maintained and ultimately replaced by present labour. Thus the capitalist's reward derived not from his exertions but from the economic power which allowed him to transform 'natural' into 'social' price through the addition of profit to natural value. Here Hodgskin distinguished Smith's additive explanation of the determination of exchange value under capitalism from Ricardo's labour embodied theory. For Hodgskin, while Ricardo had explained what *should* determine natural value. Smith had made clear what *were* the actual determinants of prices under capitalism namely wages plus profits plus rents. Thus

> Mr. R. appears to me to have confounded in the whole of his speculations real natural price with exchangeable value. The former is accurately measured by the quantity of labour necessary to obtain any commodity from nature, the latter on the contrary is the quantity of labour augmented by the amount of rent and profits.

It was, therefore, in Smithian rather than Ricardian terms that Hodgskin formulated the profit-upon-alienation theory of labour exploitation to which he adhered in his major works.

As a good Smithian it became a central concern of Hodgskin not only to ensure that the natural laws of economic life prevailed but also to demolish the arguments of those who would impugn their beneficence. Thus Hodgskin's *Popular Political Economy* challenged in particular Malthusian population theory, which seemed to suggest that there existed insuperable natural obstacles to material prosperity in the form of Nature's parsimony and Man's sexual incontinence. For Hodgskin, the obstacles which existed were not natural but the consequence of the coercive exercise of power bolstered by social regulations and artificial laws. Contrary to Malthus, he saw population increase not as a cause of indigence but of material improvement, with demographic pressure creating new demands and needs, stimulating the inventive faculties of Mankind and so enhancing society's capacity to produce.

Hodgskin was equally critical of the slur upon Nature cast by Ricardian rent theory. This he saw as implying the necessary advent of a poverty stricken stationary state due to the finite nature of land resources. For Hodgskin poverty was not a dictate of niggardly nature. On the contrary, it resulted from the unwarranted exactions of capitalist, landowner, State and Church; in particular it was 'the overwhelming nature of the demands of capital, sanctioned by the customs of men, enforced by the legislature . . . which keep the labourer in poverty and misery'.

In his last major work, *The Natural and Artificial Right of Property Contrasted* (1832), Hodgskin went on to consider the nature of the rights conferred by these positive laws as against those granted by Nature herself. As he saw it positive law generally legitimized the gains reaped by the exercise of coercive force, while at best it did no more than mirror the dictates of natural law, imposing upon the present a conception of rights which historical progress would rapidly render redundant.

Hodgskin's achievement was to integrate a teleological optimism based upon an anti-Malthusian conception of the consequences of population increase, with a penetrating critique of contemporary capitalism which turned to critical use the tools, concepts and analytical constructs of political economy. It is this critique which explains his categorization as a 'socialist'.

Yet, while their critical analyses are similar in many respects, Hodgskin's vision of the future never permitted any flirtation with the Owenites or the principles of Owenite socialism. His was an individualistic Utopia and he never doubted that private property was a *sine qua non* of material progress. His just and equitable society was essentially atomistic, an unplanned consequence of the spontaneous, unrestricted actions of individuals. It is, therefore, in the company of William Godwin and Herbert Spencer, rather than Owen, Bray and Marx that Hodgskin should be placed.

Marx did, of course, see much of worth in Hodgskin's political economy in particular, his theory of capital, but it would be entirely wrong to see Hodgskin as his intellectual precursor. For, while the Ricardianism of Marx led him to locate exploitation at the point of production, the Smithianism of Hodgskin led him to place it in the sphere of circulation or exchange. Thus while Marx preached the working-class seizure of the means of production, Hodgskin advocated the creation of equitable exchange relations through the liberation of market forces. This is the essence of Hodgskin's libertarian political economy which was eventually to evolve from the acerbic anti-capitalism of the 1820s to the 1840s when, writing for the *Economist* (1848–53), he began to deny the existence of any necessary antagonism between capital and labour. In the final analysis, Hodgskin wished to purify and generalize capitalism rather than to destroy it.

Selected Works

1820. *Travels in the north of Germany, describing the present state of the social and political institutions, the agriculture, manufactures, commerce, education, arts and manners in that country, particularly in the Kingdom of Hannover*, 2 vols. Edinburgh.

1825. *Labour defended against the claims of capital: Or, the unproductiveness of capital proved with reference to the present combinations amongst journeymen, by a labourer*. London.

1827. *Popular political economy, four lectures delivered at the London Mechanics' Institution*. London.

1832. *The natural and artificial right of property contrasted, a series of letters addressed without permission to H. Brougham, Esq*. London.

Bibliography

Beales, H.L. 1933. *The early English socialists*. London: Hamish Hamilton.

Beer, M. 1953. *A history of British socialism*, 2 vols. London: Allen & Unwin.

Cole, G.D.H. 1977. *A history of socialist thought*, 5 vols. Vol. 1: Socialist thought, the forerunners, 1789–1850. London: Macmillan.

Foxwell, H.S. 1899. Introduction to the English translation of A. Menger. In *The right to the whole produce of labour*. London: Macmillan.

Gray, A. 1967. *The socialist tradition, Moses to Lenin*. London: Longman.

Halévy, E. 1903. *Thomas Hodgskin*. Trans. with an introduction by A.J. Taylor. London: Benn, 1956.

Hollander, S.G. 1980. The post-Ricardian dissension: A case study of economics as ideology. *Oxford Economic Papers* 32: 370–410.

Hunt, E.K. 1977. Value theory in the writings of the classical economists. *History of Political Economy* 9: 322–345.

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.

Lowenthal, E. 1911. *The Ricardian socialists*. New York: Longman.

Marx, K. 1969. *Theories of surplus value*, 3 vols. Moscow: Progress.

Thompson, N.W. 1984. *The people's science: The popular political economy of exploitation and crisis 1816–34*. Cambridge: Cambridge University Press.

Hold-Up Problem

Yeon-Koo Che and József Sákovics

Keywords

Bargaining; Coase theorem; Contract failure; Fixed-price contracts; Hold-up; Incomplete contracts; Ownership structures; Relationship-specific investment; Risk neutrality; Underinvestment hypothesis; Vertical integration

JEL Classifications

C78; D4; D83; C70; D23; K12; L14; L22

Investments are often geared towards a particular trading relationship, in which case the returns on them within the relationship exceed those outside it. Once such an investment is sunk, the investor has to share the *gross* returns with her trading partner. This problem, known as hold-up, is inherent in many bilateral exchanges. For instance, workers and firms often invest in firm-specific assets prior to negotiating for wages. Manufacturers and suppliers often customize their equipment and production processes to the special needs of their partners, knowing well that future (re)negotiation will confer part of the benefit from customization to their partners. Clearly, the risk of the investor being held up discourages him or her from making socially desirable investments.

We first describe a simple model of hold up and illustrate the main *underinvestment* hypothesis (see Grout 1984, and Tirole 1986, for the first formal proof). A buyer and a seller, denoted B and S, can trade quantity $q \in [0\overline{q}]$, where $\overline{q} > 0$. The transaction can benefit from the seller's (irreversible) investment. The investment decision is binary, $I \in \{0, 1\}$, with $I = 1$ meaning 'invest' and $I = 0$ meaning 'not invest'. The investment I costs the seller $k\ I$, where $k > 0$. Given investment I, the buyer's gross surplus from consuming q is $v_I(q)$ and the seller's cost of delivering q is $c_I(q)$, where both v_I and c_I are strictly increasing with $v_I(0) = c_I(0) = 0$. Let $\varphi_I = \max_{q \in Q}[v_I(q) - c_I(q)]$ denote the efficient social surplus given S's investment, and let q_I^* be the associated socially efficient level of trade. The net social surplus is then $W\ (I) := \varphi_{I}\_kI$. Suppose that

$$\varphi_1 - k > \varphi_0, \tag{1}$$

so it is socially desirable for S to invest.

A crucial assumption is that S's investment decision, although observable to the parties, is not verifiable, and therefore it cannot be contracted upon. For the moment, assume as well that the nature of trade is sufficiently 'inchoate' so that the parties can contract on q only after S's investment decision has been made. We model the negotiation of this contract *à la* Nash, yielding the efficient trading decision q_I and splitting the gross surplus φ_I equally between the parties. The seller thus appropriates only a fraction (a half, in this case) of her investment return, while she bears the entire cost of investment, k, so her net payoff will be $U_S(I) := \frac{1}{2}\varphi_I - kI$, following her investment. Suppose

$$\frac{1}{2}\varphi_1 - k < \frac{1}{2}\varphi_0. \tag{2}$$

Then, even though the investment is socially desirable, S will not invest. Hence underinvestment arises.

Organizational Remedies

One interpretation of the inefficiency is the failure of the Coase Theorem. The parties cannot achieve the efficient outcome since the non-contractibility of S's investment decision prevents them from meaningfully negotiating over that decision *ex ante*. From this perspective, the hold-up problem entails a transaction cost of market/bargaining mechanisms, and, as Coase (1937) suggested, the transaction cost may be avoided or reduced via other organizational structures. Indeed, Klein et al. (1978) and Williamson (1979) suggested *vertical integration* as an organizational response.

Just how the hold-up problem disappears or at least diminishes through integration is not clear, however, and requires a theory of how a particular ownership structure affects the parties' exposure to hold up. This is precisely what Grossman and Hart (1986) and Hart and Moore (1990), hence forth GHM accomplish (see also Hart 1995, for an excellent synopsis). According to them, the ownership of an asset gives the owner the right to determine the use of the asset that is contractually not specifiable. The parties will still negotiate the terms of trade (presumably to achieve an efficient outcome), but this *residual right* – and thus ownership – matters, since it determines the status quo payoffs of the parties in the negotiation.

To illustrate how the status quo payoffs may affect the incentives, consider our model above and suppose that either B or S can own all assets necessary for the vertical operations. The former type of integration is called B-*integration* and the latter type is called S-*integration*. Fix *i*-integration and fix S's investment decision $I \in \{0, 1\}$. If they fail to agree on the trade decision, party i can unilaterally realize the (status quo) payoff of $\psi_i^i(I)$ and party $j \neq i$ can realize the payoff of $\psi_i^i(I)$. It is reasonable to assume that, for $i = 1, 2$

Assumption GHM: (i) $\psi_i^i(I) + \psi_j^i(I) \leq \varphi_I$, $I \in \{0, 1\}$;

(ii) $\psi_S^i(1) - \psi_S^i(0) < \varphi_1 - \varphi_0$; (iii) $\psi_i^i(1) > \psi_i^i(0)$ and $\psi_j^i(1) = \psi_j^i(0)$.

Assumption GHM-(i) means that the status quo is welfare dominated by efficient trade; (ii) means that S's investment is specific to the relationship; and (iii) means that the investment improves the owner's status quo payoff but not the non-owner's.

Given the assumption that the parties split the surplus over and above the status quo payoffs, S's payoff will be

$$U_S^i(I) = \psi_S^i(I) + \frac{1}{2}\left(\varphi_I - \psi_B^i(I) - \psi_S^i(I)\right) - kI$$
$$= \frac{1}{2}\varphi_I + \frac{1}{2}\left(\psi_S^i(I) - \psi_B^i(I)\right) - kI.$$

Hence, S's gain from investing under *i*-integration is

$$U_S^i(1) - U_S^i(0) = \frac{1}{2}(\varphi_1 - \varphi_0) + \frac{1}{2}\Delta^i - k, \quad (3)$$

Where

$$\Delta^i := \psi_S^i(1) - \psi_S^i(0) - \left[\psi_B^i(1) - \psi_B^i(0)\right].$$

Given assumption GHM-(ii) and -(iii), $\varphi_1 - \varphi_0 > \Delta^S > 0 > \Delta^B$. Hence,

$$U_S^S(1) - U_S^S(0) > U_S(1) - U_S(0)$$
$$> U_S^B(1) - U_S^B(0).$$

This shows that the S-*integration* is the optimal ownership structure, dominating symmetric (non-integrated) structure, which in turn dominates B-*integration* structure. In particular, if $U_S^S(1) - U_S^S(0) > 0 > U_S(1) - U_S(0)$ then the investment is sustainable if and only if the seller has the asset ownership. This result reveals the main tenet of GHM that asset ownership can serve to reduce the owner's exposure to hold up.

Remark 1 *The effects of alternative ownership structures may depend on the particular bargaining solution assumed. For example, the outside option bargaining or a Bertrand bidding solution may change the relative rankings of the alternative structures and may eliminate inefficiencies altogether. If the buyer's outside option is binding either from the buyer's owning more assets (that is, B-integration) or from the seller being subject to competition from another seller, then the seller is forced to make the buyer indifferent to that option, which causes the seller to internalize the social return of her investment. For this reason, B-integration may perform better than S-integration* (Chiu 1998; De Meza and Lockwood 1998)*, or competition/nonintegration may solve the hold-up problem* (Bolton and Whinston 1993; Che and Hausch 1996; Cole et al. 2001; Felli and Roberts 2001; MacLeod and Malcomson 1993).

Contractual Solutions

In the above model, the trade decision is contractible only after the investment decision has been made. While this assumption resonates with many real business situations, it is difficult to reconcile with the fact that the parties can accurately calculate the payoff consequences of their behaviour (Maskin and Tirole 1999). It is also crucial: if the parties *can* contract on q prior to the investment decision, the underinvestment problem may be solved, without requiring the organizational remedies discussed above.

To illustrate, suppose the parties sign a contract requiring them to trade $\hat{q}$ for the total price of $\hat{t}$. Unless renegotiated, this contract will give S a payoff of $\hat{t} - c_I(\hat{q}) - kI$if she chooses $I \in \{0, 1\}$. If $\hat{q} \neq q_I^*$, though, both parties will be better off by renegotiating to implement q_I^*. Given the assumption that this renegotiation splits the surplus equally, S's *ex ante* payoff will be

$$\hat{U}_S(I;\hat{q}) :$$
$$= \hat{t} - c_I(\hat{q}) + \frac{1}{2}[\varphi_I - (v_I(\hat{q}) - c_I(\hat{q}))]$$
$$- kI.$$

Hence, her net benefit from investing under this contract is

$$\hat{U}_S(1;\hat{q}) - \hat{U}_S(0\hat{q}) = \frac{1}{2}(\varphi_1 - \varphi_0)$$
$$- \frac{1}{2}(v_1(\hat{q}) - v_0(\hat{q}))$$
$$- \frac{1}{2}(c_1(\hat{q}) - c_0(\hat{q}))$$
$$- k. \quad (4)$$

Whether a contract like this can create a sufficient incentive for S to invest depends on the nature of the investment made. Suppose first that the investment is *selfish*, so that it only decreases S's cost but does not affect B's valuation (that is, $v_1(\cdot) = v_0(\cdot)$). In this case, the trade contract can indeed protect S's incentive for investment. Observe that

$$\begin{aligned} c_0\left(q_1^*\right) - c_1\left(q_1^*\right) &= v_1\left(q_1^*\right) - c_1\left(q_1^*\right) \\ &\quad - \left[v_0\left(q_1^*\right) - c_0\left(q_1^*\right)\right] \\ &\geq \varphi_1 - \varphi_0. \end{aligned}$$

By the same logic, $c_0\left(q_0^*\right) - c_1\left(q_0^*\right) \leq \varphi_1 - \varphi_0$. Since $c_I(\cdot)$ is continuous, there exists $\hat{q}^*$ between q_0^* and q_1^* such that $c_0(\hat{q}^*) - c_1(\hat{q}^*) = \varphi_1 - \varphi_0$. Consequently, $\hat{U}_S(1;\hat{q}^*) - \hat{U}_S(0;\hat{q}^*) = W(1) - W(0)$, so S will indeed invest whenever it is efficient to do so. Edlin and Reichelstein (1996) show that a fixed-price contract can provide efficient incentives for a selfish investment by either side and, with an additional condition, for selfish investments by both, in a more general environment with continuous investment. This result implies that, as long as the investments are selfish, the organizational remedies mentioned above will not be necessary.

Remark 2 *Aghion et al.* (1994) *and Chung* (1991) *have noted that efficiency can be achieved for investments by both sides via a contract that manipulates the status quo payoff of one party in the same way as above and gives the full bargaining power to the other party at the renegotiation stage, thus making that party a residual claimant of the social surplus in the marginal sense. The idea of contractual manipulation of bargaining powers also appears in Hart and Moore* (1988) *and Nöldeke and Schmidt* (1995).

Contract Failure

Contracts may not restore efficiency if the investments are not selfish. Suppose the investment is *cooperative*: $c_1(\cdot) = c_0(\cdot)$. So, S's investment increases B's valuation only, worsening the former's bargaining position. Such a cooperative nature of investments underlies many instances of the hold-up problem (for example, quality-enhancing R&D investment by a supplier and customization efforts by partners). In this case, any commitment to trade exacerbates rather than alleviates the investor's vulnerability to hold up. Formally, given $c_1(\cdot) = c_0(\cdot)$, S's *ex ante* payoff will be

$$\begin{aligned} \hat{U}_S(1;\hat{q}) - \hat{U}_S(0\hat{q}) &= \frac{1}{2}(\varphi_1 - \varphi_0) - \frac{1}{2}(v_1(\hat{q}) \\ &\quad - v_0(\hat{q})) - k \leq \frac{1}{2}(\varphi_1 - \varphi_0) - k \\ &= U_S(1) - U_S(0) < 0. \end{aligned}$$

for any $\hat{q}$. In other words, no such trade contract creates more incentives for S than the null contract. In fact, Che and Hausch (1999) demonstrated that all feasible contracts are worthless if investments are cooperative.

A similar result can be obtained if the investment is selfish, but it is difficult to predict the 'type' of trade that will benefit from the investment (Hart and Moore 1999; Segal 1999). Specifically, suppose that there are n potential goods the parties may wish to trade but that only one of them becomes a 'special' type and *only* the special type will benefit from an investment. Assume that each of the n goods has an equal chance of becoming that special type *ex post*, so the parties can predict the special type only with probability $1/n$. Adapted to our model, the surplus from trading the special type is φ_I given investment $I \in \{0, 1\}$, and the surplus from trading a 'generic' type is φ_0, regardless of the investment decision. Assume for simplicity that $q_I^* = 1$, for $I = 0, 1$. As the contract is renegotiable, under a contract requiring the parties to trade any good, S's *ex ante* payoff from choosing $I \in \{0, 1\}$, becomes

$$\begin{aligned} &\tilde{U}_S(I): \\ &\quad = \frac{1}{n}(\hat{t} - c_I(1)) + \frac{n-1}{n}(\hat{t} - c_0(1)) \\ &\quad + \frac{1}{2}\left[\varphi_I - \frac{1}{n}\varphi_I - \frac{n-1}{n}\varphi_0\right] - kI. \end{aligned}$$

In other words, S's investment influences her status quo payoff only when the good they contracted to trade turns out to be the special type, an event that arises with probability $1/n$. This feature weakens the ability of a contract to provide incentives, as can be seen from S's gain from investing:

$$\begin{aligned}\tilde{U}_S(1) - \tilde{U}_S(0) &= \frac{1}{n}(c_0(1) - c_1(1)) \\ &\quad + \frac{1}{2}\left[\varphi_1 - \varphi_0 - \frac{1}{n}(\varphi_1 - \varphi_0)\right] \\ &\quad - k \\ &= \frac{1}{2}\left(1 + \frac{1}{n}\right)(\varphi_1 - \varphi_0) - k.\end{aligned}$$

Further, as the environment becomes 'complex' in the sense that $n \to \infty$, S's incentive reduces to that under the null contract, thus rendering contracts virtually worthless.

Several implications can be drawn from these two results. First, the contract failure result implies that the true challenge of the hold-up problem may lie with the nature of specific investments – either the 'cooperative' nature or the 'unpredictability of investment benefit'. Second, the general failure of contracting to protect against hold up lends credence and relevance to the GHM analysis of the ownership structures or organizational theory in general based on the hold-up problem as a source of inefficiency. Third, for the above results it is crucial for the parties to be unable to commit not to renegotiate their contract. Were such commitment available, they could devise a contract that would induce them to reveal truthfully S's investment decision, say, by having both parties report simultaneously about the decision and penalizing both of them for any inconsistency via zero trade and zero transfer. Then, S can easily be induced to invest by a sufficient amount of bonus given to her only conditional on both parties reporting 'S has invested'. If a contract is renegotiable, such a costless revelation of information is impossible to achieve: Inconsistent reports do not reveal the identity of the liar, and both parties cannot be simultaneously punished, since they will renegotiate back to the Pareto frontier.

Remark 3 *Several elements are crucial for the contract failure result. First, it requires the existence of an opportunity to renegotiate following any contractspecified action. If there is some non-renegotiable action, then an efficient outcome may be achievable. Rogerson (1984) shows that liquidated damages achieve the efficient outcome if a contract can be breached non-renegotiably. Likewise, if in the last period of renegotiation the buyer can irrevocably determine the terms of trade, then buyer-option contracts can overcome the hold-up problem* (see Lyon and Rasmusen 2004). *Contract failure re-emerges, however, in the case of cooperative investment if the parties discount delayed exercise of the option* (Wickelgren 2007). *Second, risk neutrality is important for contract failure. If the parties were risk averse, then a lottery could be used to punish both parties even in the presence of renegotiation, and could achieve the first-best* (Maskin and Tirole 1999). *Third, it is important for the contract to be bilateral. If a third party can be involved, efficiency can be achieved even when the contract is subject to renegotiation or collusion* (Baliga and Sjostrom 2005). *Last,* Watson (2006) *gives a general treatment of how renegotiation opportunities arising at different stages interact with the technology of trade, and recognizes the relevance of modeling technological details of trade, i.e., whether the trade is individual or public.*

Dynamics

The basic hold-up model assumes that there is a single opportunity to invest, followed by the distribution of the surplus. Not too surprisingly, if the interaction is repeated, inefficiencies can be greatly reduced, in accordance with the Folk Theorem for repeated games (see, for example, Klein and Leffler 1981). More surprisingly, allowing for dynamic investment patterns can have a dramatic effect even in a oneshot interaction, as shown by Che and Sákovics (2004a). When the agents can continue to invest even after the negotiation of the terms of trade has started, the anticipated investment dynamics can influence the way the parties negotiate and improve the incentives for investment.

To see how this works, modify our running example by allowing S to invest in the following period if she has not invested in the past and no agreement has been reached yet. If the parties discount their future very little, S's 'invest' can be sustained in a subgame-perfect equilibrium. In this equilibrium, hold up still arises on the equilibrium path in that S receives only the fraction of the gross surplus commensurate with his bargaining power. Yet this does not stop S from investing. Suppose S does not invest today but is expected to invest tomorrow in case no agreement is reached today. Then, there will be more surplus to divide tomorrow than there is today. Since the cost of tomorrow's investment will be borne solely by the investor, the prospect of the investor raising his investment tomorrow causes his partner to demand more to settle today. The investment dynamics thus results in a worse bargaining position for the party upon not investing, and creates a stronger incentive for investing than would be possible if such investment dynamics – that is, the option to invest in the future – were not allowed. As a result, investment can be supported in equilibrium.

In sum, dynamics in the trading relationship and/or investment technology lessens either the risk of hold up or the degree of inefficiencies caused by it. This questions the relevance of the hold-up problem as a rationale for organization and/or contractual remedies. At the same time, the presence of dynamics alters the nature of the incentive problems and calls for different types of contractual or organizational prescriptions against hold up than those proposed based on the static models, as seen by Baker et al. (2002), Che and Sákovics (2004b) and Halonen (2002).

See Also

- ▶ Coase Theorem
- ▶ Contract Theory
- ▶ Incomplete Contracts
- ▶ Procurement

Bibliography

Aghion, P., M. Dewatripont, and P. Rey. 1994. Renegotiation design with unverifiable information. *Econometrica* 62: 257–282.

Baker, G., R. Gibbons, and K. Murphy. 2002. Relational contracts and the theory of the firm. *Quarterly Journal of Economics* 117: 39–83.

Baliga, S., and T. Sjostrom. 2005. *Contracting with third parties*. Mimeo: Northwestern University.

Bolton, P., and M. Whinston. 1993. Incomplete contracts, vertical integration and supply assurance. *Review of Economic Studies* 60: 121–148.

Che, Y.-K., and D. Hausch. 1996. *Cooperative investments and the value of contracting: Coase vs. Williamson*. Mimeo: University of Wisconsin-Madison.

Che, Y.-K., and D. Hausch. 1999. Cooperative investments and the value of contracting. *American Economic Review* 89: 125–147.

Che, Y.-K., and J. Sákovics. 2004a. A dynamic theory of holdup. *Econometrica* 72: 1063–1103.

Che, Y.-K. and Sákovics, J. 2004b. Contractual remedies to the hold up problem: A dynamic perspective. Social Science Research Institute Working Paper N. 2004-03, University of Wisconsin, and Economics Discussion Paper No. 100, University of Edinburgh.

Chiu, Y.S. 1998. Noncooperative bargaining, hostages, and optimal asset ownership. *American Economic Review* 88: 882–901.

Chung, T.-Y. 1991. Incomplete contracts, specific investment and risk sharing. *Review of Economic Studies* 58: 1031–1042.

Coase, R. 1937. The nature of the firm. *Economica* 4: 386–405.

Cole, H., G. Mailath, and A. Postlewaite. 2001. Efficient non-contractible investments in large economies. *Journal of Economic Theory* 101: 333–373.

De Meza, D., and B. Lockwood. 1998. Does asset ownership always motivate managers? Outside options and the property rights theory of the firm. *Quarterly Journal of Economics* 113: 361–386.

Edlin, A., and S. Reichelstein. 1996. Holdups, standard breach remedies and optimal investment. *American Economic Review* 86: 478–501.

Felli, L., and K. Roberts. 2001. Does competition solve the hold-up problem? STICERD Theoretical Economics Discussion Paper No. TE/01/414, London School of Economics.

Grossman, S., and O. Hart. 1986. The costs and benefits of ownership: A theory of lateral and vertical integration. *Journal of Political Economy* 94: 691–719.

Grout, P. 1984. Investment and wages in the absence of binding contracts: A Nash bargaining approach. *Econometrica* 52: 449–460.

Halonen, M. 2002. Reputation and the allocation of ownership. *Economic Journal* 112: 539–558.

Hart, O. 1995. *Firms, contracts, and financial structure*. Oxford: Clarendon Press.

Hart, O., and J. Moore. 1988. Incomplete contracts and renegotiation. *Econometrica* 56: 755–785.

Hart, O., and J. Moore. 1990. Property rights and the nature of the firm. *Journal of Political Economy* 98: 1119–1158.

Hart, O., and J. Moore. 1999. Foundations of incomplete contracts. *Review of Economic Studies* 66: 115–138.

Klein, B., R. Crawford, and A. Alchian. 1978. Vertical integration, appropriable rents, and the competitive contracting process. *Journal of Law and Economics* 21: 297–326.

Klein, B., and K. Leffler. 1981. The role of market forces in assuring contractual performance. *Journal of Political Economy* 89: 615–641.

Lyon, T., and E. Rasmusen. 2004. Buyer-option contracts restored: Renegotiation, inefficient threats, and the holdup problem. *Journal of Law, Economics, and Organization* 20: 148–169.

Macaulay, S. 1963. Non-contractual relations in business: A preliminary study. *American Sociological Review* 28: 55–70.

MacLeod, W., and J. Malcomson. 1993. Investments, holdup, and the form of market contracts. *American Economic Review* 83: 811–837.

Maskin, E., and J. Tirole. 1999. Unforeseen contingencies and incomplete contracts. *Review of Economic Studies* 66: 83–114.

Nöldeke, G., and K. Schmidt. 1995. Option contracts and renegotiation: A solution to the hold-up problem. *RAND Journal of Economics* 26: 163–179.

Rogerson, W. 1984. Efficient reliance and damage measure for breach of contract. *RAND Journal of Economics* 15: 39–53.

Segal, I. 1999. Complexity and renegotiation: A foundation for incomplete contracts. *Review of Economic Studies* 66: 57–82.

Tirole, J. 1986. Procurement and renegotiation. *Journal of Political Economy* 94: 235–259.

Watson, J. 2006. Contracts, mechanism design and technological detail. *Econometrica* (in press).

Wickelgren, A. 2007. The limitations of buyer-option contracts in solving the hold-up problem. *Journal of Law, Economics, and Organization* 23(1):(forthcoming).

Williamson, O. 1979. Transactions-cost economics: The governance of contractual relations. *Journal of Law and Economics* 22: 233–262.

Hollander, Jacob Harry (1871–1940)

A. W. Coats

Born in Baltimore, Maryland on 23 July 1871, Hollander spent his entire career at the Johns Hopkins University, studying under R.T. Ely and J.B. Clark, graduating AB in 1891, PhD in 1894, and joining the faculty immediately thereafter. A versatile scholar, his special fields were labour economics, the history of economic thought, and public finance. In the first of these he ran a notable seminar for several decades with his colleague George Barnett, and both were elected President of the American Economic Association, Hollander in 1921, Barnett in 1932. As a doctrinal historian Hollander is especially remembered for his discovery and editing of Ricardo's letters, and the latter's important *Notes on Malthus*. He also collected a major library of works on economics. As a tax and financial expert Hollander held numerous local, state, federal and international posts, especially in Puerto Rico (1900–1901) and in the Dominican Republic (1905–1907), where he continued to serve as financial adviser up to 1910. He was a pacifist, opposing US membership of the League of Nations, and a defender of Prohibition.

Selected Works

1895. *Letters of David Ricardo to John Ramsay McCulloch 1816–1823*. New York: Macmillan Co.

1899a. *The financial history of Baltimore*. Baltimore: Johns Hopkins Press.

1899b. (ed., with J. Bonar). *Letters of David Ricardo to Hutches Trower*. Oxford: Clarendon Press.

1905a. *Debt of Santo Domingo . . . . Report on the debt of Santo Domingo submitted to the President of the United States*. Washington, DC: Government Printing Office.

1905b. (ed., with G. Barnett). *Studies in American trade unionism*. New York: Holt and Co. Reprinted, 1912.

1910. *David Ricardo. A centenary estimate*. Baltimore: Johns Hopkins Press.

1914. *The abolition of poverty*. Boston/New York: Houghton Mifflin Co.

1919. *War borrowing: A study of Treasury certificates of indebtedness of the United States*. New York: Macmillan Co.

1928. (ed., with T.E. Gregory). *Notes on Malthus' 'Principles of political economy' by David Ricardo*. Baltimore/London: Johns Hopkins Press/Oxford University Press.

Homan, Paul Thomas (1893–1969)

Warren J. Samules

Homan was born on 12 April 1893 in Indianola, Iowa, and died on 3 July 1969 in Washington, DC. Educated at Williamette University, Oxford University and the Brookings Institution (then Graduate School of Economics and Government) (PhD, 1926), he taught at Cornell (1919–47), the University of California at Los Angeles (1950–69), and Southern Methodist University (1953–63). He was managing editor of the *American Economic Review*, 1941–52. He served with the War Production Board, UNRRA, UNESCO, and the Council of Economic Advisers. He was on the staff of the Brookings Institution and also was associated with Resources for the Future.

An expert on the National Recovery Administration, Homan later wrote on oil conservation regulation, estimating oil and gas reserves, costing in the petroleum industry, and, several years before the OPEC oil embargo, problems of Middle Eastern oil for the western world.

Homan's *Contemporary Economic Thought* (1928) was an influential interpretation of the state of the discipline at the time. He emphasized its enormous diversity, treating the heterodox work of John A. Hobson, Thorstein Veblen and Wesley C. Mitchell alongside the more orthodox doctrines of John Bates Clark and Alfred Marshall, all as serious inquiry within general economic theory. Although admiring economics as a science, he recognized that economics has the quality of a system of beliefs, both influenced by and influencing general philosophical and ideological points of view in society. He found the principal axis of diversity to lie between those who emphasized the static, deductive, mathematical individualist approach, and those who pursued realism, empiricism and holistic evolutionism. His personal view was complex. He clearly thought that value theory, logical deduction, mechanical analogy, and the study of the price system were central to economics, and that a framework of thought commanding more general assent might be desirable; but that diversity was not objectionable *per se*, there being room for the study of the price system, institutions, and the meaning of economic life.

Selected Works

1928. *Contemporary economic thought*. New York: Harper.

1934. (With others.) *The ABC of the NRA*. Washington, DC: Brookings Institution.

1935. (With others.) *The national recovery administration: An analysis and appraisal*. Washington, DC: Brookings Institution.

1945. (ed., with F. Machlup.) *Financing American prosperity.* New York: Twentieth Century Fund.

1964. (With W.F. Lovejoy.) *Problems of cost analysis in the petroleum industry.* Dallas: Southern Methodist University Press.

1965. (With W.F. Lovejoy.) *Methods of estimating reserves of crude oil. Natural gas, and natural gas liquids*. Baltimore: Johns Hopkins University Press for Resources for the Future.

1967. (With W.F. Lovejoy.) *Economic aspects of oil conservation regulation*. Baltimore: Johns Hopkins University Press for Resources for the Future.

1971. (With S.H. Schurr.) *Middle Eastern oil and the Western World; prospects and problems*. New York: Elsevier.

Home Production

Yongsung Chang and Andreas Hornstein

Abstract

Studying the incentives and constraints in the non-market sector – that is, home production – enhances our understanding of economic behaviour in the market. In particular, it helps us to understand (*a*) small variations of labour supply over the life cycle, (*b*) the low correlation between employment and wages over the business cycle, and (*c*) large income differences across countries.

Keywords

Capital accumulation; Cobb–Douglas functions; Elasticity of substitution; Home production; Hours worked; Income differences; Intertemporal substitution elasticity; IS–LM; Labour supply; Labour supply elasticities; Leisure; Michigan Time Use Survey; National income accounts; Neoclassical growth theory; Non-market activity; Real business cycles; Stochastic growth model; Technical progress; Total factor productivity

JEL Classification

D4; D10

Studies such as the Michigan Time Use Survey (Hill 1984; Juster and Stafford 1991) indicate that a typical married couple allocates only about one-third of its discretionary time working for paid compensation in the market. The allocation of time for non-market activities, such as home production or leisure, may be as important for economic welfare as is the time spent working. Starting with Becker (1965) and Mincer (1962), the value of non-market activity has been explicitly incorporated into economic analysis in terms of forgone earnings. Since household decisions on the allocation of time to market and non-market activities are undertaken jointly, studying the incentives and constraints in the non-market sector – home production – enhances our understanding of economic behaviour in the market sector. We discuss three examples where the inclusion of home production has improved our understanding of macroeconomic issues: (*a*) low estimates of the labour supply elasticity from panel data; (*b*) low correlation between return to working and hours worked over the business cycle; and (*c*) large differences in measured output across countries.

In a standard neoclassical growth model with home production, a household derives utility not only from the consumption of market goods but also from the consumption of non-market goods. Non-market goods are produced in a home production sector using work effort and capital. The household's utility also depends on the consumption of leisure, which is the household's time endowment minus work effort supplied to the market and the home production sector. One usually assumes that the economy's technology is such that investment goods that can be used to augment the capital stock in the market and non-market sectors are produced only in the market sector of the economy. Important factors in the determination of the dynamics of a neoclassical growth model with home production are the substitution elasticity between the consumption of market and non-market goods, the substitution elasticity between capital and labour in market and home production, the relative capital intensity of production in the market and the home production sectors, and the correlation of total factor productivity in the two sectors. Examples of the neoclassical growth model augmented with home production are Benhabib et al. (1991) and Greenwood and Hercowitz (1991).

Business Cycle Analysis

The allocation of hours worked – employment – is at the heart of business cycle analysis. Table 1 shows the standard deviations and correlation of the cyclical components of total hours worked and returns to working for the US economy, 1964–2003.

Two features are of great interest to macroeconomists. First, hours worked is substantially more volatile than the return to working. Second, hours worked is not highly correlated with the return to working. Employment in other countries also exhibits similar features (for example, Backus et al. 1992). These facts present a serious challenge to modern business cycle theory that builds on the idea of intertemporal substitution of work effort. Intertemporal substitution assumes that people work relatively more hours in some

Home Production, Table 1 Business cycle statistics of the US labour market, 1964–2003

σ_n/σ_w	$\sigma_n/\sigma_{y/n}$	$cor(n, w)$	$cor(n, y/n)$
1.51	1.72	.38	.01

Note: All variables are logged and de-trended with the use of the Hodrick–Prescott filter. Hours worked (n) represents the total hours employed in the non-agricultural business sector. Wages (w) are the real hourly earnings of production and non-supervisory workers. Labour productivity (y/n) is output divided by hours worked. The period covered is from 1964:I to 2003:II

Sources: DRI-WEFA Basic Economics Database; Global Insight

years than in others because the return from working in the market is unusually high in those years (for example, Lucas and Rapping 1969). According to Table 1, on the one hand it appears as if employment would have to be very elastic in its response to changes in the return to work, but on the other hand the returns to work appear to be only weakly correlated with the supply of work time.

Estimates of Labour Supply Elasticity

Business cycle theory that builds on the stochastic growth model – for example, Kydland and Prescott (1982) – indeed requires a large intertemporal elasticity of substitution in order to account for the relatively large fluctuations of hours worked. Yet a substantial empirical literature based on micro data finds that households' willingness to substitute hours is quite low – less than 0.5 (for example, MaCurdy 1981; Altonji 1986). Home production provides a potential resolution of this problem.

Most micro estimates of the intertemporal substitution elasticity rely on the variation of hours worked and wages over the life cycle of households. Rupert et al. (2000) show that these estimates may underestimate the true willingness to substitute hours across time if one does not take into account the fact that households simultaneously decide on the supply of hours for market and non-market activities. Essentially, conventional estimates of labour supply elasticities suffer from an omitted variable bias: home work is positively correlated with market work and should be included in the estimation. For simplicity we assume that households' preferences are log-linear in a consumption aggregator of market, c_{mt}, and home-produced consumption, c_{ht}, and work time, be it in the market, n_{mt}, or at home, n_{ht}:

$$u(c_{mt}, c_{ht}, n_{mt}, n_{ht}) = \log c(c_{mt}, c_{ht}) - B\frac{(n_{mt} + n_{ht})^{1+1/\gamma}}{1 + 1/\gamma}.$$

Then the optimal labour supply of a household that is t years old can be written as

$$\log w_t = (1/\gamma)\log(n_{mt} + n_{ht}) + A_t,$$

where w_t denotes the market wage rate, and A_t represents other terms that may depend on age. The parameter γ denotes the willingness to substitute total hours over time – intertemporal substitution elasticity. For conventional estimates of the labour supply elasticity, which ignore home production, time spent for home production activities represents an unobserved supply shifter for market labour.

A typical worker faces a hump-shaped wage profile in his life: wage rates rise, reach a peak at age 45–55, and decline from then on. It is not unreasonable to assume that the consumption of non-market goods, and therefore hours worked in home production, is correlated with the market wage profile over the life cycle. For example, high earning years tend to be around the years in which one buys a house or has children, both of which call for more time spent in home production. The fact that home work and market work are positively correlated over the life cycle, but home work is omitted from the estimation equation, implies that the estimated inverse labour supply elasticity $1 + \hat{\gamma}$ will be biased upward.

Wage–Employment Correlations

One of the primary empirical patterns that have puzzled many business cycle theorists is the lack of a systematic relationship between employment and wages. On the one hand, Keynesian IS–LM models assume that real wages and hours worked lie on a stable, downward-sloped marginal product of labour schedule, and predict a strong negative correlation between real wages and hours worked (for example, Dunlop 1938). On the other hand, real business-cycle models, such as that of Kydland and Prescott (1982), where productivity shocks shift the labour demand schedule along a relatively stable positively sloped market labour supply curve, tend to predict a strong positive correlation between wages and employment. Incorporating home production into the neoclassical growth model helps account for the low correlation between market work and wages as well as the large variation of employment.

Technical progress not only augments the marginal product of labour in the market sector but also affects the marginal product of labour in the home production sector. Consider, for example, technical progress that is embodied in consumer durables, such as vacuum cleaners and washers. This kind of technological progress often reduces the required work effort in the home sector for household chores, and thereby shifts the supply curve of market work outward along a negatively sloped market demand for labour curve. Thus, while technical progress in the market sector causes a positive correlation between market hours and wages, technical progress in the non-market sector can cause a negative correlation between market hours and wages. If technical progress in the market is positively correlated with that in the non-market sector, then market hours may fluctuate substantially without any accompanying changes in real wages.

In general, the allocation of hours between the market and home depends on (*a*) the covariance structure of productivity in the market and home, (*b*) the substitution elasticity between market goods and home-produced goods, and (*c*) the substitution elasticity between capital and labour in the home production function – in particular, if the purchase of home capital (for example, a home theatre system) requires or saves hours in home production. Recently, rich structures between the market and home production have been introduced to study the various features of business cycles – for example, McGrattan et al. (1997), Hornstein and Praschnik (1997), Fisher (1997), Einarsson and Marquis (1997), Ingram et al. (1997), Perli (1998), Chang (2000), Gomme et al. (2001).

Cross-Country Income Differences

There are enormous income differences across countries, and such disparity has persisted over time. According to Heston et al. (2002), the ratio of the average per capita GDP (based on purchasing power parity price) of the richest fifth of all countries to that of the poorest fifth of all countries was about 12 in 1960 and had doubled to almost 25 by 2000. In the standard neoclassical growth model, distortions to capital accumulation contribute to income differences. For a reasonably calibrated neoclassical growth model, the distortions that are required to account for the observed income differences are, however, unreasonably large. Parente et al. (2000) show that the required distortions are substantially reduced once we distinguish between an economy's market sector whose output is measured in the national income accounts and a home-production sector whose output is not measured. With home production, distortions to capital accumulation not only reduce the capital stock but also can reallocate economic activity from the market sector to the non-market sector. Moreover, the measured income differences overstate the true differences in welfare, and the unmeasured consumption from home production may explain how individuals in some countries can survive on the very low levels of reported income.

Consider the neoclassical growth model with log preferences in consumption, c_m, and leisure, l. Output, y_m, is produced using capital, k_m, and labour, n_m, as inputs to a constant returns to scale Cobb–Douglas production function, $y_m = k_m^{\alpha_m}(z_m n_m)^{1-\alpha_m}$. Output can be used for consumption and investment, x_m, to increase the capital stock: $k_{m,t+1} = (1 - \delta)k_{mt} + x_{mt}/\pi$, where δ is the depreciation rate. With capital accumulation distortions, investment increases the capital stock less than one for one: $\pi \geq 1$ (for example, Parente and Prescott 1994). It is easily conceivable that there are substantial inefficiencies in capital accumulation in less developed economies (for example, inefficient governments, ill-protected property rights). Given commonly assumed preferences and technology, the investment rate and work effort on the balanced growth path will be independent of the magnitude of capital distortions, but the capital stock and output will decline with the capital distortion. Two countries that look alike in terms of the investment rates may nevertheless have very different output levels. Conditional on a reasonable parameterization of the economy, we would, however, have to assume capital distortions, $\pi \geq 100$, in order to account for observed output differences of a factor of at least 10 (for example, Parente et al. 2000).

A straightforward extension of the neoclassical growth model that includes home production assumes that preferences are defined over a consumption aggregator that includes market consumption and non-market consumption, c_h, from the home-production sector. The home-production sector also uses capital, k_h, and work effort, n_h, as inputs to a Cobb–Douglas production function. The household's time endowment can now be used in the market and the non-market sectors, and market production can be used for investment in the market and the non-market sectors. If home production is less capital-intensive than market production, and market and non-market goods are sufficiently close substitutes, a higher capital distortion not only reduces total capital accumulation but also leads to a reallocation of the available capital and work effort from the market sector to the non-market sector. Parente et al. (2000) argue that, for reasonable substitution elasticities between market and home-production consumption and capital shares in the home-production sector, capital distortions as low as $\pi = 15$ can account for income differences in the market sector of a factor of ten.

See Also

- ▶ Business Cycle Measurement
- ▶ Economic Growth, Empirical Regularities In
- ▶ Labour Supply
- ▶ Real Business Cycles
- ▶ Time Use

Any opinions expressed in this paper are those of the authors and do not necessarily reflect those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

Bibliography

Altonji, J. 1986. Intertemporal substitution in labor supply: Evidence from micro data. *Journal of Political Economy* 94: S176–S215.

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

Becker, G. 1965. A theory of the allocation of time. *Economic Journal* 75: 493–517.

Benhabib, J., R. Rogerson, and R. Wright. 1991. Homework in macroeconomics: Household production and aggregate fluctuations. *Journal of Political Economy* 99: 1166–1187.

Chang, Y. 2000. Comovement, excess volatility, and home production. *Journal of Monetary Economics* 46: 385–396.

Dunlop, J. 1938. The movement of real and money wage rates. *Economic Journal* 48: 413–434.

Einarsson, T., and M. Marquis. 1997. Home production with endogenous growth. *Journal of Monetary Economics* 39: 551–569.

Fisher, J. 1997. Relative prices, complementarities and comovement among components of aggregate expenditures. *Journal of Monetary Economics* 39: 449–474.

Gomme, P., F. Kydland, and P. Rupert. 2001. Home production meets time-to-build. *Journal of Political Economy* 109: 1115–1131.

Greenwood, J., and Z. Hercowitz. 1991. The allocation of capital and time over the business cycles. *Journal of Political Economy* 99: 1188–1214.

Heston, A., R. Summers, and B. Aten. 2002. *Penn world table version 6.1*. Philadelphia: Center for International Comparisons, University of Pennsylvania.

Hill, M. 1984. Pattern of time use. In *Time, goods and well–being*, ed. F. Juster and F. Stafford. Ann Arbor: University of Michigan Press.

Hornstein, A., and J. Praschnik. 1997. Intermediate inputs and sectoral comovement in the business cycle. *Journal of Monetary Economics* 40: 573–595.

Ingram, B., N. Kocherlakota, and N. Savin. 1997. Using theory for measurement: An analysis of the cyclical behavior of home production. *Journal of Monetary Economics* 40: 435–456.

Juster, F., and F. Stafford. 1991. The allocation of time: Empirical findings, behavior models, and problems of measurement. *Journal of Economic Literature* 29: 471–522.

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

Lucas Jr., R., and L. Rapping. 1969. Real wages, employment, and inflation. *Journal of Political Economy* 77: 721–754.

MaCurdy, T. 1981. An empirical model of labor supply in a life-cycle setting. *Journal of Political Economy* 88: 1059–1085.

McGrattan, E., R. Rogerson, and R. Wright. 1997. An equilibrium model of the business cycle with household production and fiscal policy. *International Economic Review* 38: 267–290.

Mincer, J. 1962. On-the-job training: Costs, returns and its implications. *Journal of Political Economy* 70: 50–79.

Parente, S., and E. Prescott. 1994. Barriers to technology adoption and development. *Journal of Political Economy* 102: 298–321.

Parente, S., R. Rogerson, and R. Wright. 2000. Homework in development economics: Household production and the wealth of nations. *Journal of Political Economy* 108: 680–687.

Perli, R. 1998. Indeterminacy, home production, and the business cycle: A calibrated analysis. *Journal of Monetary Economics* 41: 105–125.

Rupert, P., R. Rogerson, and R. Wright. 2000. Homework in labor economics: Household production and intertemporal substitution. *Journal of Monetary Economics* 46: 557–579.

Homogeneous and Homothetic Functions

J.-P. Crouzeix

JEL Classifications
C0

Homothetic Orderings

Given a cone E in the Euclidean space $\mathbb{R}^n$ and an ordering $\preccurlyeq$ on E (i.e. a reflexive and transitive binary relation on E), the ordering is said to be homothetic if for all pairs $x, y, \in E$

$$x \leq y \Rightarrow \lambda x \leq \lambda y \text{ for all } \lambda > 0.$$

For each $x \in E$, denote by $L(x)$ the indifference surface

$$L(x) = \{y \in E : y \leq x \text{ and } x \leq y\}.$$

Hence, geometrically, if the ordering is homothetic, then for all $x \in E$ and $\lambda > 0$

$$L(\lambda x) = \{\lambda y : y \in L(x)\}.$$

Homothetic Functions

Recall that a real function f on a set E defines a complete (or total) ordering on E via the relation

$$x \leq y \text{ if and only if } f(x) \leq f(y).$$

By definition, f is said to be homothetic if the ordering is homothetic (implying that the domain E of f is a cone). Thus utility functions which represent a homothetic ordering are homothetic.

Assume, now, that f is a homothetic and differentiable function on an open cone E of $\mathbb{R}^n$. Assume also that $\nabla f(x) \neq 0$ for all $x \in E$. Hence for all $\lambda > 0$ and all $x \in E$ there exists $k > 0$ such that

$$\frac{\partial f}{\partial x_i}(\lambda x) = k \frac{\partial f}{\partial x_i}(x), \text{for } i = 1, 2, \ldots, n.$$

In economic terms, this property means that the marginal rate of substitution remains constant along any ray from the origin. In fact, under some suitable assumptions, this property characterizes homothety of functions.

Positively Homogeneous Functions

A real function f defined on a cone E of $\mathbb{R}^*$ is said to be positively homogeneous of order p if for all $x \in E$

$$f(\lambda x) = \lambda^p f(x) \text{ for all } \lambda > 0.$$

If $p = 1$, the function is said to be positively homogeneous or linearly homogeneous. If $p = 0$, then the definition becomes

$$f(\lambda x) = f(x) \text{ for all } \lambda > 0 \text{ and } x \in E.$$

Clearly, positively homogeneous functions of any order are homoethetic. Conversely, under some suitable assumptions on E and f (for instance E is the positive orthant in $\mathbb{R}^n$ and f is increasing on E) then, if f is homothetic there exist a positively homogeneous function g of order 1 on E and an increasing function k on $\mathbb{R}$ such that

$$f(x) = k[g(x)] \text{ for all } x \in E.$$

(This property is sometimes used as an alternative definition of homothety for functions.) As a consequence, under reasonable economic assumptions, a homothetic preference ordering can be represented by a linearly homogeneous utility function. Production functions are often assumed to be positively homogeneous of order p. For example, the so-called Cobb –Douglas function

$$f(x_1, x_2, \ldots, x_n) = Kx_1^{\alpha_1} x_2^{\alpha_2} \ldots x_n^{\alpha_n} x_j > 0,$$

where K, α_1, α_2, . . . , α_n are positive constants, is homogeneous of order $p = \alpha_1 + \alpha_2 + \ \ldots \ + \alpha_n$.

In consumer theory, demand functions are positively homogeneous of order zero in prices and wealth.

Positively Homogeneous Convex (or Concave) Functions

Since convexity is a fundamental concept in economics, special attention should be paid to positively homogeneous functions which are convex or concave.

Let E be a convex cone and f a real function on E. Then a necessary and sufficient condition for f to be convex (concave) and positively homogeneous of order 1 on E is that for all $x \in E$ and $\lambda \geq 0$

$$f(\lambda x) = \lambda f(x)$$

and for all pairs $x, y \in E$

$$f(x + y) \leq (\geq) f(x) + f(y).$$

The producer's cost function illustrates a concave positively homogeneous function: assuming that only one output is produced using n inputs, the cost function is given by

$$c(y,p) = \underset{x}{\text{Min}}[p^t x : F(x) \geq y]$$

where p_i, $i = 1,2,..., n$, is the unit price of input i and $F(x)$, the production function, is the maximal amount of output which can be produced with the input vector $x = (x_1, x_2,..., x_n)$. Then, for a fixed price vector p, $c(y, p)$ is the minimal cost of producing y units of the output. For y fixed, $c(y, p)$ is concave and positively homogeneous of order 1 in p. Similarly, in consumer theory, if F now denotes the consumer's utility function, the $c(y, p)$ represents the minimal price for the consumer to obtain the utility level y when p is the vector of utility prices.

A fundamental property is as follows. Let f be a real continuous function on a closed convex cone of $\mathbb{R}^n$. Then f is convex and positively homogeneous of order 1 if and only if there exists a closed convex set S of $\mathbb{R}^n$ such that

$$f(x) = \text{Sup}\left[y'x/y \in S\right]$$

This set S is unique and the function is called the support function of S (by symmetry, the same result holds when replacing convex by concave and Sup by Inf). Duality in consumer's (as well as in producer's) theory is based on this property.

We conclude with three examples of functions widely used in mathematics. A *semi-norm* on $\mathbb{R}^n$ is a convex positively homogeneous function f of order one on $\mathbb{R}^n$ such that $f(x) = f(-x)$ for all x (then $f(x) \geq 0$ for all x). A *norm* is a semi-norm for which $x = 0$ whenever $f(x) = 0$. Finally, given a convex set C which contains the origin, the *gauge* of C is the function f defined by

$$f(x) = \text{Inf } [\lambda \geq 0/x \in \lambda C]$$

A gauge function is convex and positively homogeneous of order one. Moreover, if the origin belongs to the interior of C and C is balanced (i.e. $x \in C$ implies that $x \in -C$), then the gauge is a norm.

Positively Homogeneous Quasi-Concave (Quasi-Convex) Functions

Let $\succcurlyeq$ be a preference ordering on a set E. In view of economic considerations, a common and reasonable assumption is the convexity of the ordering (i.e. for all $x \in E$, the set $\{y \in E/y \succcurlyeq x\}$ is convex). Then the utility functions which represent the ordering are quasi-concave but in general, a concave representation does not exist. However, in the case where the ordering is homothetic, it does. Indeed, a quasiconcave linearly homogeneous function which takes only positive (negative) values on the interior of its domain is concave [Newman] (by symmetry the same result holds for quasi-convex functions). It follows that a representable preference ordering which is homothetic and convex admits a representation by a concave linearly homogeneous utility function.

See Also

- ▶ Aggregate Demand Theory
- ▶ Cobb–Douglas Functions
- ▶ Euler's Theorem
- ▶ Quasi-Concavity
- ▶ Separability

Bibliography

Barten, A.P., and V. Bohrn. 1981. Consumer theory. In *Handbook of mathematical economics*, ed. K.J. Arrow and M.D. Intriligator, vol. 2, 381–429. New York: North-Holland Publishing Company.

Diewert, W.E. 1981. Duality approaches to microeconomic theory. In *Handbook of mathematical economics*, ed. K.J. Arrow and M.D. Intriligator, vol. 2, 353–399. New York: North-Holland Publishing Company.

Green, J., and W.P. Heller. 1981. Mathematical analysis and convexity with application to economics. In *Handbook of mathematical economics*, ed. K.J. Arrow and M.D. Intriligator, vol. 1, 1–52. New York: North-Holland Publishing Company.

Katzner, D.W. 1970. *Static demand theory*. New York: Macmillan.

Newman, P. 1969. Some properties of concave functions. *Journal of Economic Theory* 1: 291–314.

Horizontal and Vertical Equity

Jean-Yves Duclos

Abstract

This article describes the concepts of vertical and horizontal equity and provides some normative and positive justifications for them. It then outlines a few of the measures that have been proposed to assess whether government policies, and tax and transfer systems in particular, are vertically and horizontally equitable. It also points to useful references in the literature.

Keywords

Concentration curve; Discrimination; Equality of opportunity; Equality of resources; Hobbes, T.; Horizontal equity; Liability progression; Locke, J.; Lorenz curve; Nozick, R.; Positioning; Procedural equity; Progressive and regressive taxation; Rawls, J.; Redistribution of income; Relative deprivation; Residual progression; Sen, A.; Social justice; Utilitarianism; Veil of ignorance; Vertical equity; Well-being

JEL Classifications

D63

Two broad principles govern the redistributive analysis of government policies. The first one, *vertical equity,* helps assess the distributive equity of a policy's impact on individuals with differing initial levels of welfare. The second, *horizontal equity,* serves to evaluate the policy's impact across individuals who are similar in all relevant ethical aspects – including their initial level of welfare.

In terms of taxation, the principle of vertical equity (VE) requires that the net fiscal burden increase with individuals' capacity to pay (measured by pre-tax income, say). A strong form of this principle is usually accepted: it postulates that the capacity to pay increases more rapidly than income, and that the net tax burden should thus also rise faster than income, and should therefore be *progressive*. It can be shown that the application of this principle serves to decrease relative inequality in income, net of the tax burden. The principle of horizontal equity, in turn, stipulates that similar individuals should receive a similar tax treatment from the government. Application of this second principle also controls for the emergence of vertical disparities among initially similar individuals. Though the two principles are generally applied to the monetary dimension of the impact of government policies, they can also prove pertinent to the analysis of other dimensions thereof.

Vertical Equity

Concern about inequality in resource allocation has a long history in moral and political philosophy, and features prominently in all major religions. It is mostly based on a belief in the fundamental dignity that is equally shared by all human beings as well as on a natural social aversion to material and human deprivation. VE in government policies is one of the tools most often advocated to bring about greater equality in resource allocation. VE in resource distribution has also long been considered a condition for social cohesion and stability. Two thousand and four hundred years ago, Plato indeed expressed the following concern about equality:

> We maintain that if a state is to avoid the greatest plague of all – I mean civil war, though civil disintegration would be a better term –extreme poverty and wealth must not be allowed to arise in any section of the citizen-body, because both lead to both these disasters. That is why the legislator must now announce the acceptable limits of wealth and poverty. The lower limit of poverty must be the value of the holding. The legislator will use the holding as his unit of measure and allow a man to possess twice, thrice, and up to four times its value. (*The Laws,* Book V, quoted in Cowell 1995, pp. 21–2)

A utilitarian justification for a concern for VE is that surveys on the subjects of happiness and health suggest that the consumption of unnecessary goods essentially represents a consumption for ‘positioning’ vis-à-vis others. Such consumption improves the individual’s position relative to others but, in and of itself, yields little or no increase in the individual’s welfare and decreases others’ relative sense of wellbeing, causing anxiety, stress, and hostility. Individuals also appear to have difficulty dealing with feelings of relative deprivation and exclusion, which can be detrimental to the good functioning of markets and institutions. The purpose of VE is then to reduce inequality in the distribution of welfare so as to mitigate the effects of inequality’s negative externalities.

An influential ethical foundation for the principle of VE has also appeared since the 1970s in the writings of a number of philosophers, the most well-known probably being John Rawls and Amartya Sen (for example, Rawls 1971; Sen 1985). Rawls in particular has argued that in the absence of preferences and socio-economic interests (that is, behind a *veil of ignorance*; see, for instance, Harsanyi 1955), individuals would agree that social justice implies maximizing the set of opportunities and well-being of the least well-off group, namely, equalizing opportunities ‘upwards’ so that the greatest possible well-being be available to all.

Horizontal Equity

As already mentioned, the principle of horizontal equity (HE) stipulates that ethically similar individuals must be treated similarly by the government. ‘Ethically similar’ also implies having a similar level of well-being, since as seen above it can be ethically justified for governments to distinguish between poor and rich. Two initially similar individuals must therefore find themselves at approximately the same welfare level after the effect of a government policy has been accounted for, regardless of the individuals’ initial preferences or socio-economic characteristics. This is the *classical* formulation of the principle of horizontal equity. An important corollary is that government interventions should not reverse the ranking of individuals in the distribution of welfare, unless it can be shown that the initial ranking was unjust – this is the alternative and popular *reranking* formulation of the HE principle.

The rationale for HE is primarily borne of a concern for procedural equity. Unlike for VE, it is not the result that is judged, but the process. For example, it can be argued that a reranking of two individuals by the government (in which one of the two receives assistance) can reduce the income distance (and vertical inequality) between the two, but this reranking must be considered horizontally inequitable if the initial ranking was not demonstrably unjust.

The HE principle is not only universally simple to appreciate, but it generally also garners more support from philosophers than the VE principle (though see Kaplow 1989, for a critique). The most important ethical justification for HE is the avoidance of all forms of arbitrary discrimination in the government's treatment of citizens. Individuals of similar ethical worth should be treated and valued equally by the government. Notice that we are here dealing with individuals who are ethically similar, though not necessarily identical in all respects. Limiting the principle of HE to individuals who are identical in all points would strip it of virtually all practical relevance and would arguably leave governments too much latitude to practise arbitrary discrimination between individuals.

Drawing on the 17th-century social contract theories of Thomas Hobbes and John Locke, the foundations of this procedural justice were promoted *inter alia* by Nozick (1974), for whom the usual theories of justice place too much emphasis on outcomes in the redistribution of welfare, utility, or capacities. However, the bases for HE also follow from theories of vertical equity, since the unequal treatment of equals can only increase the distance between them. Robert Musgrave, an influential contributor to the development of the HE principle (see in particular Musgrave 1959), summarizes this as follows:

> The requirement of HE remains essentially unchanged under the various formulations of distributive justice, ranging from Lockean entitlement over utilitarianism and fairness solutions. That of VE, on the contrary, undergoes drastic changes under the various approaches. While HE is met by the various VE outcomes, this does not mean that HE is derived from VE. If anything, it suggests that HE is a stronger primary rule. (Musgrave 1990, p. 116)

There are also various utilitarian foundations for the principle of HE. Government policies that discriminate between ethically comparable individuals give rise to resentment and insecurity amongst them and can also lead to social and political unrest. Exclusion and discrimination can have an impact on both individual welfare and on feelings of social cohesion; this is particularly for policies that discriminate among those that are alike since individuals often specifically compare their treatment with that of others who enjoy a similar standard of living or characteristics.

There are two major sources of horizontal inequity (HI). The first is that the impact of public policy often varies purposefully with individual characteristics and preferences, and the second is that public policy is typically non-deterministic by design and/or in application. Instances of HI occur in practice because of the difficulties faced by policies to account appropriately for household heterogeneity, and because of informational problems, administrative errors, incomplete take-up, tax evasion, randomness in the effect of programs and policies, and outright or implicit discriminatory behaviour by the government.

Measurement

Local Measures of VE and Progressivity

Let X and N represent respectively pre-tax income and post-tax incomes, and let $T(X)$ be taxes, with $N = X - T(X)$ – and suppose for a moment that the tax system is deterministic (or non-stochastic) and differentiable. Denote the average rate of taxation at pre-tax income X by $t(X) = T(X)/X$, and the derivative of $t(X)$ and $T(X)$ at $X = x$ by $t'(x)$ and $T'(x)$. A tax $T(X)$ is said to be

- locally progressive at $X = x$ if the average rate of taxation increases with X, that is, if $t'(x) > 0$;
- locally proportional at $X = x$ if the average rate of taxation stays constant with X, that is, if $t'(x) = 0$;
- and locally regressive at $X = x$ if the average rate of taxation decreases with X, that is, if $t'(x) < 0$.

The elasticity of taxes with respect to X, also called *liability progression,* is then given by:

$$LP(X) = \frac{X}{T(X)} T'(X) = \frac{T'(X)}{t(X)}. \qquad (1)$$

$LP(X)$ is the *local* ratio of the marginal tax rate over the average tax rate at X. A second local measure of progression, $RP(X)$, called *residual progression,* is the elasticity of net income with respect to pre-tax income:

$$RP(X) = \frac{\partial(X - T(X))}{\partial X} \cdot \frac{X}{N} = \frac{1 - T'(X)}{1 - t(X)}. \quad (2)$$

A tax system is everywhere progressive if $RP(X) < 1$ everywhere.

Lorenz and Concentration Curves

For several reasons, we can expect the tax system to be stochastically linked to X, and can thus express taxes T as $T = T(X) + v$, where $T(X)$ and v are respectively a deterministic and stochastic tax determinant. The Lorenz curve $L_X(p)$ for X is the proportion of the total X that is held by those whose percentile in the distribution of X is p or lower. A frequent tool for measuring the VE of the tax T is the concentration curve, defined as $C_T(p) = \int_0^p \overline{T}(q)dq/\mu_T$, where $\overline{T}(q)$ is the expected tax paid by those at percentile q in the distribution of X, and where μ_T is the average of T in the entire population. $C_T(p)$ thus shows the proportion of total taxes paid by the p bottom proportion of the population. The concentration curve $C_N(p)$ for net incomes is analogously defined as the proportion of total N that is enjoyed by those whose percentile in the distribution of X is p or lower. Finally, let t be the average tax as a proportion of average pre-tax income: $t = \mu_T/\mu_X$. On the assumption of no reranking from the pre-tax to the post-tax distribution, the following conditions are then equivalent:

1. $t'(X) > 0$ for all X;
2. $LP(X) > 1$ for all X;
3. $RP(X) < 1$ for all X;
4. $L_X(p) > C_T(p)$ for all $p \in]0,1[$ and for any distribution of pre-tax income;
5. $L_N(p) > L_X(p)$ for all $p \in]0,1[$ and for any distribution of pre-tax income.

Progressive taxation thus makes the distribution of N unambiguously more equal than the distribution of X, in the sense that it pushes up the Lorenz curve for incomes *whatever* the distribution of pre-tax incomes. *Tax progressivity* and *vertical equity* can in that sense be used interchangeably. Analogous results can be obtained for the more general case in which T can be negative (in the context of a *tax and benefit* system, say; see Duclos and Araar 2006, for more details). In the presence of reranking (when $T'(X) > 1$ or when the tax system is stochastic), result 5 does not hold anymore.

Global Measures of VE and Progressivity

There are two major approaches to measuring *global* progressivity: the tax-redistribution (TR) approach, and the income-redistribution (IR) approach.

1. A tax T is TR-progressive if $C_T(p) < L_X(p)$ for all $p \in]0, 1[$.
2. A tax T is IR-progressive if $C_N(p) > L_X(p)$ for all $p \in]0, 1[$.

For two taxes, T_1 and T_2, if $LP_1(X) > LP_2(X)$ at all values of X, then the tax 1 is necessarily more TR-progressive than the tax 2; if $RP_1(X) < RP_2(X)$ at all values of X, then the tax 1 is necessarily more IR-progressive than the tax 2. In the absence of reranking, a more IR-progressive tax system is one which decreases inequality by more and is therefore more vertically equitable.

Horizontal Equity

The literature on the measurement of HE has evolved very significantly since around 1980. There have been two sub-periods, the first of which focused on the measurement of reranking using concentration and Lorenz curves and indices based thereon. One central result is that $C_N(p)$ will never be lower than the Lorenz curve $L_N(p)$, and will be strictly greater than $L_N(p)$ for at least one value of p if there is reranking in the redistribution of incomes. A tax T will thus cause reranking (and hence horizontal inequity) if and only if $C_N(p) > L_N(p)$ for at least one value of p. The difference between the Lorenz curve of post- and pre-tax incomes can then be expressed as:

$$L_N(p) - L_X(p) = \underbrace{C_N(p) - L_X(p)}_{\text{VE: progressivity}} - \underbrace{(C_N(p) - L_N(p))}_{\text{HI: Reranking}} \quad (3)$$

This shows why a progressive tax system that causes reranking can push the Lorenz curve down and therefore increase inequality.

A recent promising approach to measuring classical HE has been to estimate the impact of the variability of taxes conditional on some initial value of pre-tax income. Capturing the impact of this variability can be done using many of the popular social welfare and inequality indices; see *inter alia* Aronson et al. (1994); Aronson and Lambert (1994); Lambert and Ramos (1997); Duclos and Lambert (2000); and Auerbach and Hassett (2002). This has typically led to total redistribution being expressible as the difference between VE and HI components.

Further Reading

Classical texts on the concept and the measurement of VE and tax progressivity include Musgrave and Thin (1948); Slitor (1948); Blum and Kahen, Jr. (1963); Vickrey (1972); Fellman (1976); Jakobsson (1976); Kakwani (1977a, b); Suits (1977); Reynolds and Smolensky (1977); Atkinson (1979); Plotnick (1981, 1982), King (1983); and Pfahler (1987). Recent literature surveys on the meaning and the measurement of HE can be found in Jenkins and Lambert (1999); Lambert (2001); and Duclos and Araar (2006).

See Also

- ▶ Redistribution of Income and Wealth
- ▶ Tax Incidence
- ▶ Taxation and Poverty

Bibliography

Aronson, R., P. Johnson, and P. Lambert. 1994. Redistributive effect and unequal income tax treatment. *Economic Journal* 104: 262–270.

Aronson, R., and P. Lambert. 1994. Decomposing the Gini Coefficient to reveal the vertical, horizontal, and reranking effects of income taxation. *National Tax Journal* 47: 273–294.

Atkinson, A. 1979. Horizontal equity and the distribution of the tax burden. In *The economics of taxation*, ed. H. Aaron and M. Boskin. Washington, DC: Brookings Institution.

Auerbach, A., and K. Hassett. 2002. A new measure of horizontal equity. *American Economic Review* 92: 1116–1125.

Blum, W., and H. Kahen Jr. 1963. *The uneasy case for progressive taxation*. Chicago: University of Chicago Press.

Cowell, F. 1995. *Measuring inequality*. London: Prentice Hall, Harvester Wheatsheaf.

Duclos, J.-Y., and A. Araar. *2006. Poverty and equity: Measurement, policy and estimation with DAD*. Boston: Springer.

Duclos, J.-Y., and P. Lambert. 2000. A normative approach to measuring classical horizontal inequity. *Canadian Journal of Economics* 33: 87–113.

Fellman, J. 1976. The effect of transformations on Lorenz curves. *Econometrica* 44: 823–824.

Harsanyi, J. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy* 63: 309–321.

Jakobsson, U. 1976. On the measurement of the degree of progression. *Journal of Public Economics* 5: 161–168.

Jenkins, S., and P. Lambert. 1999. Horizontal inequity measurement: A basic reassessment. In *Handbook of income inequality measurement. With a foreword by Amartya Sen*, ed. J. Silber. Boston/London: Dordrecht/Kluwer.

Kakwani, N. 1977a. Applications of Lorenz curves in economic analysis. *Econometrica* 45: 719–728.

Kakwani, N. 1977b. Measurement of tax progressivity: An international comparison. *Economic Journal* 87: 71–80.

Kaplow, L. 1989. Horizontal equity: Measures in search of a principle. *National Tax Journal* 42: 139–154.

King, M. 1983. An index of inequality: With applications to horizontal equity and social mobility. *Econometrica* 51: 99–116.

Lambert, P. 2001. *The distribution and redistribution of income*. 3rd ed. Manchester/New York: Manchester University Press ; distributed by Palgrave, New York.

Lambert, P., and X. Ramos. 1997. Horizontal inequity and vertical redistribution. *International Tax and Public Finance* 4: 25–37.

Musgrave, R. 1959. *The theory of public finance*. New York: McGraw-Hill.

Musgrave, R. 1990. Horizontal equity, once more. *National Tax Journal* 43: 113–122.

Musgrave, R., and T. Thin. 1948. Income tax progression 1929–48. *Journal of Political Economy* 56: 498–514.

Nozick, R. 1974. *Anarchy, state and utopia*. Oxford: Basil Blackwell.

Pfahler, W. 1987. Redistributive effects of tax progressivity: Evaluating a general class of aggregate measures. *Public Finance/Finances Publiques* 42: 1–31.

Plotnick, R. 1981. A measure of horizontal inequity. *Review of Economics and Statistics* 62: 283–288.

Plotnick, R. 1982. The concept and measurement of horizontal inequity. *Journal of Public Economics* 17: 373–391.

Rawls, J. 1971. *A theory of justice*. Cambridge, MA: Harvard University Press.

Reynolds, M., and E. Smolensky. 1977. *Public expenditure, taxes and the distribution of income: The United States, 1950, 1961, 1970*. New York: Academic Press.

Sen, A. 1985. *Commodities and capabilities*. Amsterdam: North-Holland.

Slitor, R. 1948. The measurement of progressivity and built-in flexibility. *Quarterly Journal of Economics* 62: 309–313.

Suits, D. 1977. Measurement of tax progressivity. *American Economic Review* 67: 747–752.

Vickrey, W. 1972. *Agenda for progressive taxation*. 1st ed. New York: Ronald Press.

Horner, Francis (1778–1817)

Donald Winch

Horner was born in Edinburgh in 1778, the son of a merchant. He was educated at the Royal High School and the local university and was a member of the group of former students of Dugald Stewart who in 1802 founded the *Edinburgh Review* – a Whig quarterly which became the main reviewing periodical in political economy during the first third of the 19th century. Horner was the expert on political economy within the founding group, and his advice on books and reviewers was often crucial to the editor, Francis Jeffrey. Horner's other early claim to be of note is that he was probably one of the world's first *students* of political economy, having attended Stewart's pioneering course of lectures on the subject on no less than three occasions. The record of his studies during this period reveals him to have been a close but by no means uncritical student of the *Wealth of Nations,* and an admirer of the work of Turgot, whose writings he hoped to translate, having already translated Euler's *Elements of Algebra* from the French in 1797.

After graduation Horner joined the Scottish Bar, but in 1803 decided to move to the English Bar. He entered Parliament in 1806 under the patronage of a Whig magnate, but did not follow the party line on all matters, especially on foreign policy, where, for example, he opposed any attempt to restore the monarchy in France after Napoleon's defeat. He was one of the prime movers in calling for the establishment of the Bullion Committee in 1810, and his reputation as one of the leading parliamentary experts on political economy made him the obvious candidate for its chairmanship. Although Ricardo was a member of this committee and was later to become an outspoken advocate of its main recommendation in favour of resumption of cash payments by the Bank of England, the report was chiefly written by Horner, Huskisson and Thornton. Horner's efforts in 1816 to gain acceptance of the report's views by means of a commitment to return to convertibility in two years' time were unsuccessful; but he had already played a major part in the process which led to acceptance by the Bank of England of its public responsibilities as lender of last resort and guardian of monetary orthodoxy.

Horner's best-known article for the *Edinburgh Review* was the generally appreciative one he wrote on Henry Thornton's *Inquiry into the Nature and Effects of the Paper Credit of Great Britain* (1802), a review that is credited with being more systematic than the book itself. It corrects Thornton's erroneous opinion that domestic inflation consequent upon the over-issue of inconvertible paper money would make goods dearer abroad rather than generate a gold outflow when the market price of gold rose above the mint price. He contributed to the debate on the corn bounty in 1804 in an article which upholds Smith's conclusions against the bounty system, but not the theory on which they were based. Horner has also received attention for the article which he persistently *failed* to write, namely a review of Malthus's *Essay on Population* – possibly because he found himself in disagreement with someone with whom he had become friendly as a result of common interests and Whig sympathies. Horner's letters to Malthus on the Corn Laws reveal a proto-Ricardian response to Malthus's heresy in giving his support to the retention of a measure of protection.

Horner died of consumption in Pisa in 1817, aged 38. It was widely thought that he had good chances of becoming Chancellor of the Exchequer in a future Whig ministry; and one of the arguments used by James Mill to convince Ricardo that his services were needed in Parliament was that he would replace Horner as the spokesman for 'correct principles'.

Selected Works

1853. *Memoirs and correspondence of Francis Horner, M.P.*, 2 vols, ed. L.J. Horner. Boston.
1957. *The economic writings of Francis Horner in the Edinburgh review, 1802–6*, ed. F.-W. Fetter. London: LSE Selected Reprints.

Hot Money

Brendan Brown

Hot money describes large-scale international movements of short-term capital under a fixed exchange rate system driven either by speculation on an imminent devaluation (or revaluation) or by interest rate differentials apparently greater than exchange risk. In the two decades prior to World War I, hot money flows were rare – so great was the level of confidence in the maintenance of the gold standard in the major countries. It was quite different in the interwar years. Major episodes of hot money flows included the flood of foreign funds into France in 1926–8 on speculation that the franc would be revalued. Then in the mid-1930s, there were huge outflows of hot money from the gold bloc currencies into London and New York.

Hot money flows reached a new crescendo in the final years of the Bretton Woods system. The Sterling devaluation of 1967, the devaluation of the French franc and revaluation of the Deutsche mark in 1969, and the floating of the mark in 1971 were all preceded by huge speculative flows of capital. The biggest ever movement of hot money was in the first quarter of 1973. Speculation was rife that the Smithsonian Agreement would break down. The Nixon Administration was pursuing a prices and wages policy whilst US interest rate were held at low levels. In contrast, the Bundesbank was seeking to combat inflationary pressures by instituting a monetary squeeze and pushing interest rates to much higher levels. A general devaluation of the dollar in mid-February failed to arrest the hot money flows out of the dollar (principally into the mark). Finally, on 12 March, the EEC currencies were jointly floated. In the era of floating exchange rates, the main examples of hot money flows have been within the Snake (and its successor, the EMS) and into or out of the British pound during periods when its rate has been temporarily stabilized (either against the dollar or some weighted basket).

Hot money flows are usually a source of instability in the domestic economy, in that they induce sudden and occasionally perverse changes in monetary conditions. The country losing funds suffers deflation, sometimes intensely, as interest rates are pushed to high levels in defence of the currency. The deflationary cost of sticking to a parity in defiance of market pressure often proves unacceptable politically, even where policy-makers strongly believe that the present parity is consistent with 'fundamental equilibrium'. Thus hot money flows may produce selffulfilling prophecies. The same is true in the opposite direction. Hot money inflows into a country on speculation of a revaluation may force such action, or else the continued swelling of the domestic money supply would threaten an outbreak of inflation.

Governments have turned to a variety of weapons to combat hot money flows in order not to be deflected from their chosen policy course. One has been direct controls on capital movements. For example, banks may be restricted in their covered interest arbitrage operations. In consequence, some speculative pressure would be absorbed by the differential between the forward exchange rate and its interest rate parity level, and less pressure would fall directly on interest rates. An alternative option is the introduction of a dual exchange market, whereby capital flows are channelled through a financial tier, in which the rate floats freely. Then speculation on a change in the official rate gives rise to a change in the free rate rather than to a loss or gain of reserves together with interest rate changes. In practice, though, it is difficult to prevent leaks between the two tiers. Central banks subject to large-scale money inflows from abroad may impose raised reserve requirements on domestic banks' external liabilities and on domestic corporations' borrowing from abroad.

An alternative to direct controls as a method of insulating domestic monetary conditions from speculative pressures are policies of sterilization, where the central bank seeks to offset the effect of foreign reserve changes on the money supply by undertaking open market or swap operations. In practice, sterilization policies have rarely been applied forcefully – mainly because they tend to aggravate the flow of hot money and the amount of foreign exchange intervention necessary to support the parity. A central bank which desists from raising interest rates when its currency is under attack not only fails to increase the cost of speculation but also confirms suspicions that it is set on an easy money policy, inconsistent with exchange rate stability.

See Also

- Capital Flight
- Exchange Control
- International Capital Flows

Bibliography

Emminger, O. 1977. The D-Mark in the conflict between internal and external equilibrium 1948–75. In *Essays in international finance*, vol. 122. Princeton: Princeton University.

McKinnon, R. 1974. Sterilization in four dimensions: Major trading countries, Euro-currencies and the United States. In *National monetary policies and the international financial system*, ed. R.Z. Aliber. Chicago: University of Chicago Press.

Swoboda, A.K. 1974. The dual exchange-rate system and monetary independence. In *National monetary policies and the international financial system*, ed. R.Z. Aliber. Chicago: University of Chicago Press.

Hotelling, Harold (1895–1973)

Kenneth J. Arrow

Abstract

Harold Hotelling was devoted mainly to mathematical statistics but had a deep influence on economics. His famous 1929 paper on stability in competition introduced the notions of locational equilibrium in duopoly, with implications for political competition. His application of the calculus of variations to the allocation of a fixed stock over time formed the basis of subsequent work on the subject. In his 1938 presidential address to the Econometric Society he argued that marginal-cost pricing was necessary for Pareto optimality even for decreasing-cost industries, and showed that suitable line integrals were a generalization of consumers' and producers' surplus for many commodities.

Keywords

Accademia Nazionale dei Lincei; American Economic Association; Arrow, K. J.; Calculus of variations; Competition; Confidence intervals; Consumer surplus; Depreciation; Dummy variables; Dupuit, A.-J.-L.; Econometric society; Econometrics; Exhaustible resources; Game theory; Hotelling, H.; Institute of Mathematical Statistics; Local equilibrium; Marginal cost pricing; Market socialism; Mathematical economics; Mathematical statistics; National Academy of Sciences; Political competition; Producer surplus; Royal Statistical Society; Simultaneous equations models; Statistics and economics; Subgame perfection; Welfare economics

JEL Classifications

B31

Harold Hotelling, a creative thinker in both mathematical statistics and economics, was born in

Fulda, Minnesota, on 29 September 1895 and died in Chapel Hill, North Carolina, on 26 December 1973. His influence on the development of economic theory was deep, though it occupied a relatively small part of a highly productive scientific life devoted primarily to mathematical statistics; only ten of some 87 published papers were devoted to economics, but of these six are landmarks which continue to this day to lead to further developments. His major research, on mathematical statistics, had, further, a generally stimulating effect on the use of statistical methods in different specific fields of application, including econometrics.

His early interests were in journalism; he received his BA in that field from the University of Washington in 1919. Later in classes, he would illustrate the use of dummy variables in regression analysis by a study (apparently never published) of the effect of the opinions of different Seattle newspapers on the outcome of elections and referenda. The mathematician and biographer of mathematicians, Eric T. Bell, discerned talent in Hotelling and encouraged him to switch his field. He received an MA in mathematics at Washington in 1921 and a PhD in the same field from Princeton in 1924; he worked under the topologist, Oswald Veblen (Thorstein Veblen's nephew), and two of his early papers dealt with manifolds of states of motion.

The year of completing his PhD, he joined the staff of the Food Research Institute at Stanford University with the title of Junior Associate. In 1925 he published his first three papers, one on manifolds, one on a derivation of the F- distribution, and one on the theory of depreciation. Here, apparently for the first time, he stated the now generally accepted definition of depreciation as the decrease in the discounted value of future returns. This paper was a turning-point both in capital theory proper and in the reorientation of accounting towards more economically meaningful magnitudes.

In subsequent years at Stanford he became Research Associate of the Food Research Institute and Associate Professor of Mathematics, teaching courses in mathematical statistics and probability (including an examination of Keynes's *Treatise on Probability*) along with others in differential geometry and topology. In 1927, he showed that trend projections of population were statistically inappropriate and introduced the estimation of differential equations subject to error; he returned to the statistical interpretation of trends in a notable joint paper (1929a) with Holbrook Working, largely under the inspiration of the needs of economic analysis.

The same year he published the famous paper on stability in competition (1929b), in which he introduced the notions of locational equilibrium in duopoly. This paper is still anthologized and familiar to every theoretical economist. As part of the paper, he noted that the model could be given a political interpretation, that competing parties will tend to have very similar programmes. Although it took a long time for subsequent models to arise, these few pages have become the source for a large and fruitful literature.

The paper was in fact a study in game theory. In the first stage of the game, the two players each chose a location on a line. In the second, they each chose a price. Hotelling sought what would now be called a subgame perfect equilibrium point. However, there was a subtle error in his analysis of the second stage, as first shown by d'Aspremont et al. (1979). Hotelling indeed found a local equilibrium, but the payoff functions are not concave; if the locations are sufficiently close to each other, the Hotelling solution is not a global equilibrium. Unfortunately, this is the interesting case, since Hotelling concluded that the locations chosen in the first stage would be arbitrarily close in equilibrium. In fact, the optimal strategies must be mixed (Dasgupta and Maskin 1986, pp. 30–32).

His paper on the economics of exhaustible resources (1931a) applied the calculus of variations to the problem of allocation of a fixed stock over time. All of the recent literature, inspired by the growing sense of scarcity (natural and artificial), is essentially based on Hotelling's paper. Interestingly enough, according to his later accounts, the *Economic Journal* rejected the paper because its mathematics was too difficult (although it had published Ramsey's papers earlier); it was finally published in the *Journal of Political Economy.*

In 1931, he was appointed Professor of Economics at Columbia University, where he was to remain until 1946. There he began the organization of a systematic curriculum in theoretical statistics, which eventually attained the dignity of a separate listing in the catalogue, though not the desired end of a department or degree-granting entity. Toward the end of the 1930s, he attracted a legendary set of students who represented the bulk of the next generation of theoretical statisticians. His care for and encouragement of his students were extraordinary: the encouragement of the self-doubtful, the quick recognition of talent, the tactfully made research suggestion at crucial moments created a rare human and scholarly community. He was as proud of his students as he was modest about his own work.

He also gave a course in mathematical economics. The general environment was not too fortunate. The predominant interests of the Columbia Department of Economics were actively anti-theoretical, to the point where no systematic course in neoclassical price theory was even offered, let alone prescribed for the general student. Nevertheless, several current leaders in economic theory had the benefit of his teaching. But his influence was spread more through his papers, particularly those (1932, 1935) on the full development of the second-order implications for optimization by firms and households (contemporaneous with Hicks and Allen) and above all by his classic presidential address (1938) before the Econometric Society on welfare economics. Here we have the first clear understanding of the basic propositions (Hotelling, as always, was meticulous in acknowledging earlier work back to Dupuit), as well as the introduction of extensions from the two-dimensional plane of the typical graphical presentation to the calculation of benefits with many related commodities. He argued that marginal-cost pricing was necessary for Pareto optimality even for decreasing-cost industries, used the concept of potential Pareto improvement, and showed that suitable line integrals were a generalization of consumers' and producers' surplus for many commodities. Here also we have the clearest expression in print of Hotelling's strong social interests which motivated his technical economics. His position was undogmatic but in general it was one of market socialism. He had no respect for acceptance of the *status quo* as such, and the legitimacy of altering property rights to benefit the deprived was axiomatic with him; but at the same time he was keenly aware of the limitations on resources and the importance in any human society of the avoidance of waste.

One of Hotelling's contributions which has had very extensive practical use is not contained in a paper. In 1947, the Director of the National Park Service asked a number of economists how to evaluate the benefits to visitors to national parks. Since the fee is small, the net benefit is undoubtedly considerable. Hotelling observed in a letter (Hotelling 1947) that individuals incur considerable travel costs in coming to a park. Those individuals with the largest distance travelled can be assumed to receive zero net benefits, so that their gross benefits equal their travel costs. Nearer individuals receive a surplus that can easily be calculated.

Important as was his contribution to economics, most of his effort and his influence were felt in the field of mathematical statistics, particularly in the development of multivariate analysis. In a fundamental paper (1931b), he generalized Student's test to the simultaneous test of hypotheses about the means of many variables with a joint normal distribution. In the course of this paper, he gave a correct statement of what were later termed 'confidence intervals'. In two subsequent papers (1933, 1936) he developed the analysis of many statistical variables into their principal components and developed a general approach to the analysis of relations between two sets of variates. The statistical methodologies of these papers and in particular the last contributed significantly to the later development of methods for estimating simultaneous equations in economics.

In 1946, he finally had the long-desired opportunity of creating a department of mathematical statistics, at the University of North Carolina, where he remained until retirement. He continued his active interest in economics there.

Space forbids more than the brief mention of his important work in the foundation of two learned societies, the Econometric Society and the Institute of Mathematical Statistics, both of which he served as President at a formative stage. He received many formal honours during his lifetime, including honorary degrees from Chicago and Rochester; he was the first Distinguished Fellow of the American Economic Association when that honour was created, as well as member of the National Academy of Sciences and the Accademia Nazionale dei Lincei, Honorary Fellow of the Royal Statistical Society and Fellow of the Royal Statistical Society.

Selected Works

1925. A general mathematical theory of depreciation. *Journal of the American Statistical Association* 20: 340–53.

1927. Differential equations subject to error. *Journal of the American Statistical Association* 22: 283–314.

1929a. (With H. Working). Applications of the theory of error to the interpretation of trends. *Journal of the American Statistical Association* 24: 73–85.

1929b. Stability in competition. *Economic Journal* 39: 41–57.

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

1931b. The generalization of student's ratio. *Annals of Mathematical Statistics* 21: 360–378.

1932. Edgeworth's taxation paradox and the nature of supply and demand functions. *Journal of Political Economy* 40: 577–616.

1933. Analysis of a complex of statistical variables with principal components. *Journal of Educational Psychology* 24: 417–441, 498–520.

1935. Demand functions with limited budgets. *Econometrica* 3: 66–78.

1936. Relation between two sets of variates. *Biometrika* 28: 321–77.

1938. The general welfare in relation to problems of taxation and of railway and utility rates. *Econometrica* 6: 242–269.

1947. Letter of 18 June to Newton B. Drury. Included in R.A. Prewitt, *The economics of public recreation: An economic study of the monetary evaluation of recreation in the national parks.* Mimeo. Washington, DC: Land and Recreational Planning Division, National Park Service, 1949.

Bibliography

d'Aspremont, C., Gabszewicz, J.-J. and Thisse, J. 1979. On Hotelling's 'Stability in Competition'. *Econometrica* 47: 1145–50.

Dasgupta, P., and E. Maskin. 1986. The existence of equilibrium in discontinuous economic games, II: Applications. *Review of Economic Studies* 53: 27–41.

Hours Worked (Long-Run Trends)

Jeremy Greenwood and Guillaume Vandenbroucke

Abstract

From 1830 to 2000 hours worked fell on two accounts: a drop in the market workweek and a decline in housework. The end result was that leisure rose. What caused this? The answer is technological progress. First, rising living standards implied that people could work less. Second, the introduction of new forms of leisure goods enhanced the value of time off. Third, time-saving household products reduced the need for housework. The time released allowed women to switch from home into market production. These points are illustrated with the use of historical evidence, economic theory, and numerical examples.

Keywords

Edgeworth–Pareto complements and substitutes; Elasticity of substitution; Hours worked; Household production; Housework; Income effects; Leisure; Non-market goods; Real wage rates; Recreation; Substitutes and complements; Taxation of labour income;

Technological progress; Wealth effect; Women's work and wages

JEL Classifications

D4; D10

Between 1830 and 2000, the average number of hours worked per worker declined, both in the marketplace and at home. Technological progress is the engine of such transformation. Three mechanisms are stressed:

- the rise in real wages and its corresponding wealth effect;
- the enhanced value of time off from work, due to the advent of time-using leisure goods; and
- the reduced need for housework, due to the introduction of time-saving appliances.

These mechanisms are incorporated into a model of household production. The notion of Edgeworth–Pareto complementarity/substitutability is key to the analysis. Numerical examples link theory and data.

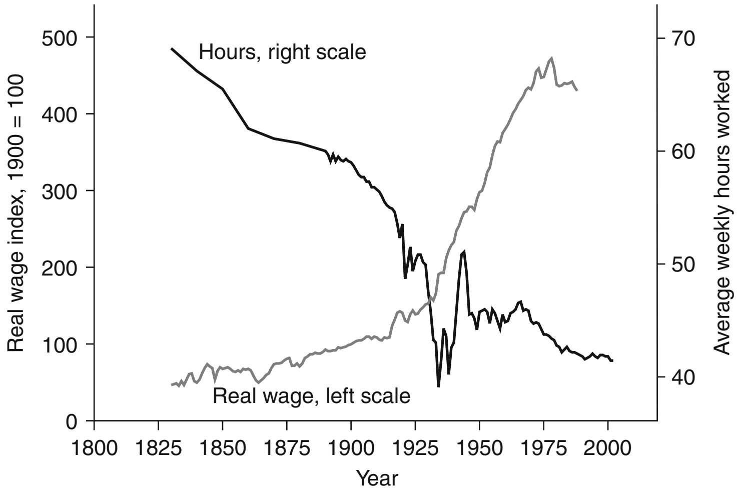
Facts

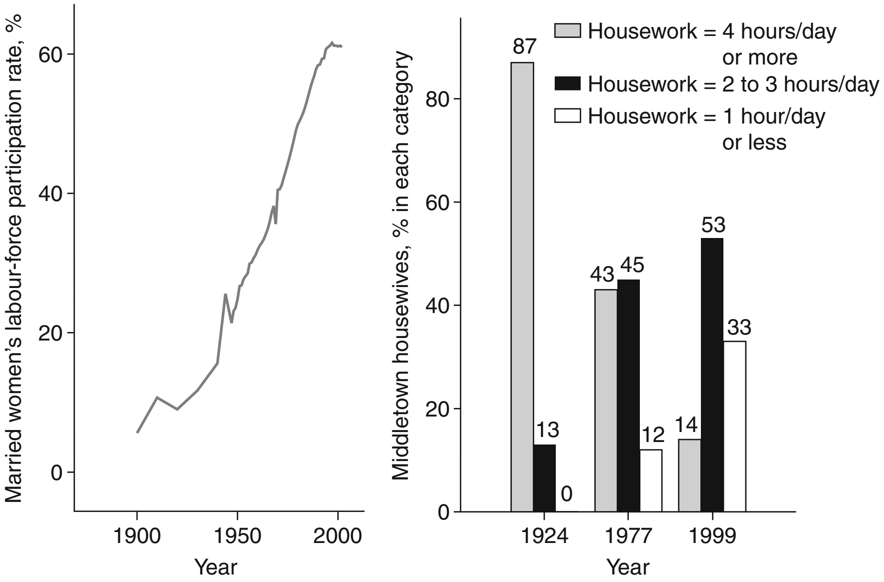
Hours worked dropped precipitously over the course of the 19th and 20th centuries, both in the marketplace and at home. In 1830 the average workweek for an American worker in the marketplace was 70 hours. This had plunged to just 41 hours by 2002. At the same time there was a ninefold gain in real wages. Figure 1 shows the shrinkage of the market workweek and the leap forward in real wages. Likewise, the amount of time spent on housework dropped. A famous study of Middletown, Indiana, documented that in 1924 87 per cent of housewives spent more than four hours per day on housework (see Fig. 2). None spent less than one hour. By 1999 only 14 per cent toiled more than four hours per day in the home, while 33 per cent spent less than one hour.

This decline in hours worked, both in the market and at home, was met by a rise in leisure. One implication of the increase in leisure is the uptrend in the share of personal consumption expenditure spent on recreation. This rose from three per cent in 1900 to 8.5 per cent in 2001, as Fig. 3 illustrates. Additionally, the amount of time that a person needs to work in order to buy the goods used for leisure has fallen by at least 2.2 per cent a year – real wages grew at an annual rate of 1.65 per cent over the 1901–88 period. This price decline neglects the fact that many new forms of leisure goods have become available over time, or that old forms have improved. As the workweek – or the time spent on work both in the market and at home – dropped, more and more women entered the marketplace to work. This may seem a little paradoxical. Only four per cent of married women worked in 1890 as compared with 49 per cent in 1980 – again, see Fig. 2.

What can explain these facts? The answer is nothing mysterious: technological progress. Three channels of effect are stressed here. First, technological progress increases wages. On the one hand, an increase in real wages should motivate more work effort since the price of consumption goods in terms of forgone leisure has fallen. On the other hand, for a given level of work effort a rise in wages implies that individuals are wealthier. People may desire to use some of this increase in living standards to enjoy more leisure. Second, the value of not working rises with the advent of new leisure goods. Leisure goods by their very nature are *time using*. Think about the impact of the following products: radio, 1919; Monopoly, 1934; television, 1947; videocassette recorder, 1979; Nintendo and Trivial Pursuit, 1984. Third, other types of new household goods reduce the need for housework. These household goods are *time saving*. Examples are: electric stove, 1900; iron, 1908; frozen food, 1930; clothes dryer, 1937; Tupperware, 1947; dishwasher, 1959; disposable diaper (Pampers), 1961; microwave oven, 1971; food processor, 1975. Some goods can be both time using and time saving, depending on the context: the telephone, 1876; IBM PC, 1984. A model is now developed to analyse the channels through which technological progress can affect hours worked in the market and time spent at home.

Hours Worked (Long-Run Trends), Fig. 1 The fall in the US market workweek and the gain in real wages, 1830–2002 (*Sources*: Average weekly hours data for 1830–80: Whaples (1990), Table 2.1). 1890–1970: *Historical Statistics of the United States: Colonial Times to 1970* (Series D765 and D803). 1970–2002: *Statistical Abstract of the United States.* Wage data: Williamson (1995, Table A1.1))





Hours Worked (Long-Run Trends), Fig. 2 The ascent of US female labour-force participation and the reduction in housework, 20th century (*Sources*: Time spent on housework in Middletown: Caplow et al. (2001, p. 37). Female labour-force participation: *Statistical Abstract of the United States)*

Analysis

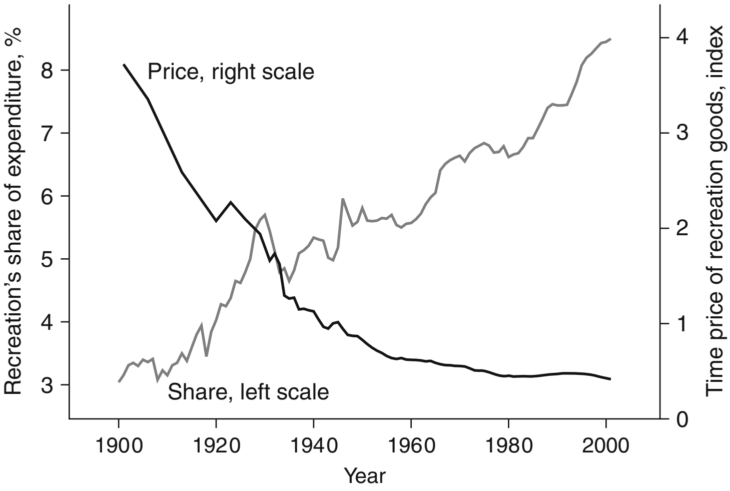
Setup

Let tastes be represented by

$$U(c) + V(n), \text{with } U_1, V_1 > 0 \text{ and } U_{11}, V_{11} < 0.$$

Here the utility functions U and V are taken to have the standard properties, while c and n represent the consumption of a market good and a non-market good. Now, suppose that the non-market good is produced in line with the constant-returns-to-scale production function

$$n = H(l, d) = dH\left(\frac{l}{d}, 1\right), \quad \text{with } H_1, \quad H_2 > 0 \quad \text{and} \quad H_{11}, \quad H_{22} < 0,$$



Hours Worked (Long-Run Trends), Fig. 3 The increase in recreation's share of expenditure and the decline in the time price of leisure in the US, 20th century (*Sources*: Recreation's share of expenditure for the years 1900–29: Lebergott (1996), Table A.1). 1929–2000: *Statistical Abstract of the United States.* Time price of leisure goods: Kopecky (2005))

where H has standard properties, d represents purchased household inputs, and l is time spent in household production. The idea that non-market goods are produced by inputs of time and goods, just as market ones are, was introduced in classic work on household production theory by Becker (1965) and Reid (1934). Assume for simplicity that there is some indivisibility associated with d. The household must use the quantity $d = \delta$. (This assumption is innocuous. Greenwood, Seshadri and Yorukoglu, 2005, Section 6, and Vandenbroucke, 2005, illustrate how it can easily be relaxed.) This fixed quantity of the household input sells at price q, which is measured in terms of time. Last, an individual has one unit of time that he can divide between working in the market and using at home. The market wage rate is w.

Now, define the function

$$X(l,d) = V\left(dH\left(\frac{l}{d},1\right)\right).$$

Household time, l, and purchased household inputs, d, are Edgeworth–Pareto complements in utility when $X_{12} > 0$ and substitutes when $X_{12} < 0$ (cf. Pareto 1906, Eqs. (63) and (64)). When l and d are Edgeworth–Pareto complements in utility, an increase in d raises the marginal utility from l, or X_1, and likewise more l increases the marginal utility from d, or X_2.

The individual's optimization problem is

$$W(w,q) = \max_{l}\{U(w(1-l) - qw) + X\,(l,\delta)\}.$$

The upshot of this maximization problem is summarized by the first- and second-order conditions written below.

$$\begin{aligned} wU_1(w(1-l)-qw) &= X_1(l,\delta) \\ &= V_1\left(\delta H\left(\frac{l}{\delta},1\right)\right)H_1\left(\frac{l}{\delta},1\right), \end{aligned} \tag{1}$$

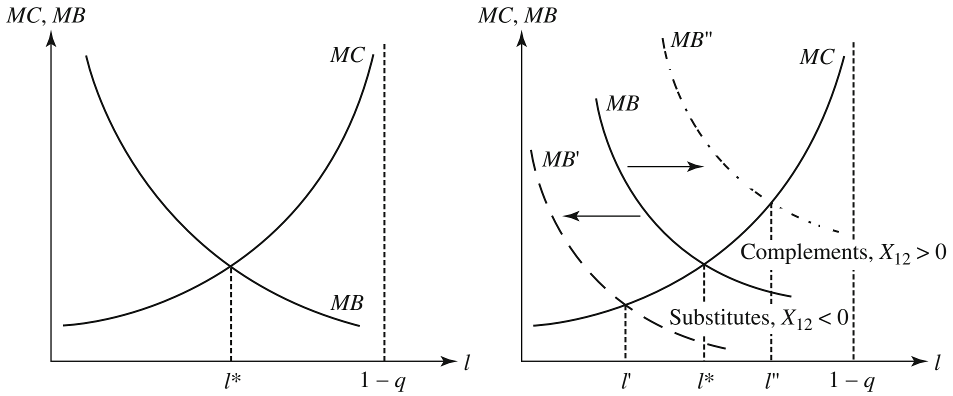
and

$$\Sigma \equiv w^2U_{11} + X_{11} < 0$$

The left-hand side of (1) represents the marginal cost of an extra unit of time spent at home. An extra unit of time spent at home results in a loss of wages in the amount w. This is worth $wU_1(w(1-l)-qw)$ in terms of forgone utility. The right-hand side gives the marginal benefit derived from spending an extra unit of time at home, $X_1(l,\delta)$. The solution for l is portrayed in Fig. 4.

Effect of Technological Progress in Household Goods

Now, suppose that there is technological progress in household goods. In particular, let this be manifested by an increase in the amount of home inputs, δ, that can be purchased for q forgone units



Hours Worked (Long-Run Trends), Fig. 4 The determination of time spent at home, l

of time. How will this affect the amount of time spent at home? It is easy to calculate that

$$\frac{dl}{d\delta} = -\frac{X_{12}}{\Sigma} \gtreqless 0 \text{ as } X_{12} \lesseqgtr 0.$$

Therefore, time spent on household activities will rise or fall depending on whether time and goods are complements or substitutes in household utility. When time and purchased inputs are complements in utility, an extra unit of d raises the worth of staying at home. So, time spent at home should rise. Leisure goods, such as television, fall into this category. Such goods have contributed to the decline in work (either in the marketplace or at home) by both men and women. A detailed account of how this mechanism can contribute to the long-run drop in hours worked is provided by Vandenbroucke (2005). This case is shown in Fig. 4 by a rightward shift in the marginal benefit curve from MB to MB'', causing time spent at home to rise from l to l''. The opposite is true when d and l are substitutes. This is portrayed in the figure by the leftward movement in the marginal benefit curve from MB to MB'. Time-saving household appliances, such as the microwave oven, are an example of this case. Such products have reduced the need for housework and have contributed to the increase in market work by women. Greenwood et al. (2005) show how the increase in female labour-force participation can be explained along these lines. Therefore, technological advance in household products is consistent with the long-run decline in the market workweek (leisure goods) and the rise in female labour-force participation (time-saving appliances and goods).

When are two goods Edgeworth–Pareto complements or substitutes? From (1) the marginal benefit of time spent at home, $X_1(l, \delta)$, is the product of two terms, the marginal utility from non-market goods, $V_1(\delta H(l/\delta,1))$, and the marginal product of household time, $H_1(l/\delta,1)$. The marginal utility of housework is decreasing in δ, while the marginal product of household time is increasing in it. Thus, the net effect of an increase in δ will depend upon whether the former falls faster with an increase in δ than the latter rises. Specifically,

$$X_{12} = -V_{11}H_1^2(l/\delta) - V_1H_{11}l|\delta^2 + V_{11}HH_1,$$

so that

$$X_{12} \lesseqgtr \text{ as } \frac{-(l/\delta)_{11}}{H_1} \lessgtr \frac{-nV_{11}}{V_1}\frac{\delta(H - H_1 l/\delta)}{n}.$$

In other words, whether or not $X_{12} \lesseqgtr 0$ depends on whether the elasticity of the marginal product of labour with respect to the time–goods ratio, $-(l/\delta)H_{11}/H_1$, is smaller or larger than the elasticity of marginal utility with respect to the home good, $-nV_{11}/V_1$, weighted by the share of purchased inputs in output, $\delta(H - H_1 l/\delta)/n$. Thus, l and δ are likely to be substitutes in utility when: (a) the responsiveness of the marginal product of l/δ is small with respect to a change in δ; (*b*) the marginal utility of home goods declines quickly with more consumption; (*c*) when purchased inputs are important in production.

Example 1 (The impact of leisure goods on hours worked) Let $U(c) = \varphi \ln (c)$ and $V(n) = (1 - \varphi) \ln (n)$. Represent the household technology by the constant-elasticity-of-substitution production function $H(l, \delta) = (\delta^{\rho} + l^{\rho})^{1/p}$. The household's budget constraint is $c = w(1 - l - q)$. Given this set-up, the first-order condition (1) can be rewritten as

$$\frac{\varphi}{1-\varphi} = \frac{1-l-q}{\delta^{\rho}+l^{\rho}} l^{\rho-1}. \qquad (2)$$

Observe that a change in wages, w, does not affect hours worked in the market, $1 - l$. The length of the workweek in the 1890s was about 42 per cent above that of the 1990s. In 1995 the typical worker spent about one-third of his available time working in the market. So, set $1 - l_{1995} = 1/3$ and $1 - l_{1895} = 1.42 \times 1/3$. Let $\delta_{1895} = 0.1$. The share of leisure goods in expenditures, s, is given by $s = q/(1 - l)$. Costa (1997) reports that this share was two per cent in the 1890s and six per cent in the 1990s. Thus, the time-price q is given by $q_t = (1 - l_t)s_t$, for $t = 1895$ and 1995. Finally, pick $\rho = -0.6$, which implies an elasticity of substitution between leisure time and leisure goods of 0.63. Proceed now in two steps. First, use (2) to back out the value of φ that is consistent with $l = l_{1895}$, $q = q_{1895}$, and $\delta = \delta_{1895}$. This results in $\varphi = 0.19$. Second, use this equation to find the value of δ_{1995} that is in agreement with $l = l_{1995}$, $q = q_{1995}$, and $\varphi = 0.19$. This leads to $\delta_{1995} = 0.69$. Voilà, an example has now been constructed where the change in market hours matches exactly the corresponding figure in the US data. Additionally, the share of expenditure spent on leisure is in line with the data. In physical units, households in 1995 had 6.90 times more leisure goods than did households in 1895. This number depends upon the elasticity of substitution between leisure time and leisure goods. The higher the degree of complementarity (or the smaller is ρ), the less is the required increase in δ.

Remark: An example can be constructed in very similar fashion to show that laboursaving household inputs (or the case of Edgeworth–Pareto substitutes) can account for the rise in female labour-force participation. The interested reader is referred to Greenwood and Seshadri (2005, Example 5, p. 1256).

Effect of an Increase in Wages

How will rising wages impact hours worked? It's easy to calculate that

$$\frac{dl}{dw} = \frac{U_1 + w(1-l-q)U_{11}}{\Sigma} \gtreqless 0 \text{ as } U_1 \lesseqgtr - w(1-l-q)U_{11}.$$

On the one hand, a boost in wages increases the opportunity cost of staying at home. This should reduce the time spent at home, l, and is represented by the substitution effect term, $U_1/\Sigma < 0$. On the other hand, higher wages make the individual wealthier. The individual should use some of this extra wealth to increase his time spent at home. This income effect is shown by the term, $w(1 - l - q)U_{11}/\Sigma > 0$. Thus, time spent at home can rise or fall with wages depending on whether the income effect dominates the substitution effect. In general, then, anything can happen, as the following two specialized cases for U make clear.

1. Let $U(c) = \ln (c)$, the macroeconomist's favourite utility function. Here, $U_1 = 1/c$ and $w(1 - l - q)U_{11} = -1/c$. Therefore, the substitution and income effects from a change in wages exactly cancel each other out. Long-run changes in wages have no impact on hours worked in the market, $1 - l$.
2. Suppose $U(c) = \ln(c - \mathcal{C})$, where $\mathcal{C} > 0$ is some subsistence level of consumption. Now, $U_1 = 1/(c - \mathcal{C})$ and $w(1 - l - q)U_{11} = -c/(c - \mathcal{C})^2$. Therefore, $dl/dw = -\mathcal{C}/\left[(c - \mathcal{C})^2\Sigma\right] > 0$. Consequently, rising wages lead to a fall in market hours, $1 - l$. The intuition is simple. At low levels of wages an individual must work hard to meet his subsistence level of consumption, $\mathcal{C}$. Achieving the subsistence level of consumption becomes

easier as wages rise and this allows the individual to ease up on his work effort. Thus, this form for the utility function is in accord with a long-run decline in hours worked. Additionally, it is consistent with the observation reported in Vandenbroucke (2005) that unskilled workers laboured longer hours in 1900 than did skilled ones, while today they work about the same.

Can an increase in wages explain the decline in the workweek? The answer is ‘yes’, as the following example makes clear.

Example 2: (The impact of rising wages on hours worked) Let $U(c) = \ln(c - \mathcal{C})$ **and** $\boldsymbol{V(n) = \alpha n}$. Represent the household technology by $\boldsymbol{H(l, d) = l}$. Equation (1) appears as

$$1 - l = \frac{1}{\alpha} + \frac{\mathcal{C}}{w}, \tag{3}$$

which gives a very simple solution for hours worked, $1 - l$. Let the time period for this example be 1830 to 1990. The real wage rate in 1990 (actually in 1988) was 9.15 times the wage rate of 1830 (Williamson 1995). So, set $w_{1830} = 1$ and $w_{1990} = 9.15$. Following the discussion in Example 1, fix hours worked in 1830 and 1990, or $1 - l_{1830}$ and $1 - l_{1990}$, using the equations $1 - l_{1830} = 1.65 \times 1/3$ and $1 - l_{1990} = 1/3$. Employing these restrictions in conjunction with (3) leads to a system of two equations in the two unknown parameters α and $\mathcal{C}$. Specifically, one obtains

$$1 - l_{1830} = \frac{1}{\alpha} + \frac{\mathcal{C}}{w_{1830}},$$

and

$$1 - l_{1990} = \frac{1}{\alpha} + \frac{\mathcal{C}}{w_{1990}}.$$

Solving yields $\alpha = 3.26$ and $\mathcal{C} = 0.24$. The subsistence level of consumption, $\mathcal{C}$, amounts to 44 per cent of consumption in 1830, and eight per cent in 1990.

The 20th century saw the advent of labour income taxation. So perhaps the previous example should have focused on the rise of after-tax wages. This is easy to amend.

Example 3: (The effect of higher labour income taxation on hours worked) Take the setup from Example 3 with one modification, to wit the introduction of labour income taxation. In particular, suppose that wages are taxed at rate τ. A fraction θ of the revenue the government receives is rebated back to the worker via lump-sum transfer payments, t. The rest goes into worthless government spending on goods and services, g – or equivalently one could assume that it enters into the consumer’s utility function in a separable manner. Hence, the worker’s budget constraint reads $c = (1 - \tau)w(1 - l) + t$, while the government’s appears as $g + t = \tau w(1 - l)$. The first-order condition for this setting is

$$\frac{(1 - \tau)w}{c - \mathcal{C}} = \alpha.$$

Combining the worker’s and government’s budget constraints yields $c = [1 - \tau(1 - \theta)]w(1 - l)$. Using this fact in the above first-order condition results in

$$1 - l = \frac{1 - \tau}{\alpha[1 - \tau(1 - \theta)]} + \frac{\mathcal{C}}{w[1 - \tau(1 - \theta)]}. \tag{4}$$

Observe that when $\mathcal{C} = 0$ and $\theta = 0$ (no rebate) an increase in the tax rate will have no impact on hours worked, because the substitution and income effects exactly cancel each other out. When $\mathcal{C} = 0$ and $\theta = 1$ (full rebate) higher taxes will dissuade hours worked since only the substitution effect is operational. Alternatively, if $\mathcal{C} > 0$ and $\theta = 0$ (no rebate), then it transpires that a rise in taxes will cause hours worked to move up. Here the negative income effect from the increase in government spending, which will result in more hours being worked, outweighs the substitution effect. Therefore, in general the effect of labour income taxation on hours worked is ambiguous. The result will depend on how the government uses the revenue it raises, and the functional forms and parameter values used for tastes and technology.

Take labour income taxes to be zero in 1830. Assume a rate of 30 per cent in 1990, in line with numbers reported by Mulligan (2002). Fix $\theta = 0.33$, its value for 1990 as measured by the National Income Product Accounts. By following the procedure in Example 3, it can be deduced that the observed fall in hours worked is occurs when $\alpha = 2.86$ and $\mathcal{C} = 0.20$. Furthermore, it can be inferred that the rise in wages accounts for 93 per cent of the fall, while the increase in taxes explains the remaining seven per cent. (For those interested, the decomposition is done as follows: Represent the right-hand side of (4) by $L(w,\tau)$. Then,

$$\begin{aligned}(1-l') - (1-l) = {} & [L(w',\tau') - L(w,\tau') \\ & + L(w',\tau) - L(w,\tau)]/2 \\ & + [L(w',\tau') - L(w,\tau') \\ & + L(w,\tau') - L(w,\tau)]/2.\end{aligned}$$

The first term in brackets is a measure of the change in hours worked, $(1 - l') - (1 - l)$, due to the shift in wages from w to w', while the second term gives the change due to a movement in taxes from τ to τ'.)

All of the above examples are intended solely as illustrations of some secular forces that potentially influence hours worked. A quantitative assessment of the impact that taxes have on hours worked will depend upon the particulars of the model used. A serious study is conducted in Prescott (2004).

The real world seems to have experienced two conflicting trends: a decline in market work and a rise in female-labour participation. A more general model could be consistent with both of these facts. To see this, imagine a framework with two types of labour, male and female. There is a division of labour in the home. Men work primarily in the market. Females do housework and, time permitting, market work. Households purchase both time-saving and time-using household inputs. Female labour-force participation would rise as labour-saving goods economize on the amount of housework that has to be done. Simultaneously, the market workweek would decline, due either to the introduction of leisure goods or to an income effect associated with a rise in wages. The value of leisure would rise for both men and women. Interestingly, Aguiar and Hurst (2006) document a dramatic increase in leisure for both men and women over the period 1965–2003. They construct various measures of leisure. They all showed a gain over the period under study. The narrowest definition rose by 6.4 hours a week for men and 3.8 hours for women, after adjustment for demographic changes in the population. This measure included time spent on activities such as entertainment, recreation, and relaxing. The authors' preferred measure increased by 7.9 hours a week for men and 6.0 hours for women. This broader definition also included activities such as eating, sleeping, personal care, and childcare. Another manifestation of the rise in the value of leisure is the increase in the fraction of life spent retired. Kopecky (2005) relays that a 20-year-old man in 1850 could expect to spend about six per cent of his life retired, while one in 1990 should enjoy about 30 per cent of his life in retirement. She shows how the trend towards enjoying more retirement can be analysed in much the same way as the decline in the workweek.

See Also

- ▶ Household Production and Public Goods
- ▶ Labour Supply
- ▶ Leisure
- ▶ Technical Change
- ▶ Time Use

Bibliography

Aguiar, M., and E. Hurst. 2006. Measuring trends in leisure: The allocation of time over five decades. Working paper no. 06-2, Federal Reserve Bank of Boston.

Becker, G. 1965. A theory of the allocation of time. *Economic Journal* 75: 493–417.

Caplow, T., L. Hicks, and B. Wattenberg. 2001. *The first measured century: An illustrated guide to trends in America, 1900–2000*. Washington, DC: AEI Press.

Costa, D. 1997. Less of a luxury: The rise of recreation since 1888. Working paper no. 6054. Cambridge: NBER.

Greenwood, J., A. Seshadri, and M. Yorukoglu. 2005. Engines of liberation. *Review of Economic Studies* 72: 109–133.

Greenwood, J., and A. Seshadri. 2005. Technological progress and economic transformation. In *Handbook of economic growth*, ed. P. Aghion and S. Durlauf, Vol. 1B. Amsterdam: North-Holland.

Kopecky, K. 2005. The trend in retirement. Economie d'avant garde Research Report No. 12, Department of Economics, University of Rochester.

Lebergott, S. 1996. *Consumer expenditures: New measures and old motives*. Princeton: Princeton University Press.

Mulligan, C. 2002. A century of labor–leisure distortions. Working paper no. 8774. Cambridge: NBER.

Pareto, V. 1906. *Manual of political economy*, 1971. New York: Augustus M. Kelly.

Prescott, E. 2004. Why do Americans work so much more than Europeans? *Federal Reserve Bank of Minneapolis Quarterly Review* 28: 2–13.

Reid, M. 1934. *Economics of household production*. New York: John Wiley.

US Bureau of the Census. 1975. *Historical statistics of the United States: Colonial Times to 1970*. Washington, DC: US Department of Commerce.

US Bureau of the Census. 2003. *Statistical abstract of the United States*, Mini Historical Statistics (suppl.). Washington, DC: US Department of Commerce.

Vandenbroucke, G. 2005. A model of the trends in hours. Economie d'avant garde Research Report No. 11, Department of Economics, University of Rochester.

Whaples, R. 1990. The shortening of the American work week: An economic and historical analysis of its context, causes, and consequences. Ph.D. dissertation, University of Pennsylvania.

Williamson, J. 1995. The evolution of global labor markets since 1830: Background evidence and hypotheses. *Explorations in Economic History* 32: 141–196.

Household Budgets

A. P. Barten

The earliest known example of systematically collected household budgets can be found in *The State of the Poor* by Eden (1797). To assess the living conditions of the lower classes Eden wanted to know, in addition to other matters, the 'Earnings and expenses of a labourer's family for a year: distinguishing the number and ages of the family; and the price and quantity of their articles of consumption' (Preface, p. iv). He obtained this information for households from some 50 parishes in England. Eden reports for these families their earnings by type of income (mostly wages) and income earner, and their expenses by type of expenditure (food, rent, fuel, clothing). Prices and quantities are only rarely given but the composition of the family and the occupation of its head are usually precisely described. Another well known early example is the collection of 199 budgets for Belgian labouring class families in 1853, published by Ducpétiaux (1855), which provided the statistical material for the formulation of Engel's Law (Engel 1857). Ducpétiaux used a uniform classification of expenditures to facilitate comparison of consumption patterns across families. The 19th century has seen a gradual extension of such household budget surveys mostly conducted by private (groups of) persons on an incidental basis. In more recent times official institutions organize these surveys more or less regularly as part of their normal operations. They may cover thousands of families.

Descriptive Aspects

In current usage a household budget is a summary of how a particular household allocates its expenditure over well defined items or groups of items during a given period (month, year). Usually, the items are grouped according to a uniform system of classes for all households participating in the same survey.

The emphasis is on expenditure rather than on earnings, contrary to the early examples. Information on earnings is more sensitive than that on expenses and is in general less accurately reported by the participants. The breakdown of expenditure into a quantity and a price component is whenever possible desirable but not always realized.

The unit, the household, consists of members of a family and others sharing living and eating arrangements. There are basically two ways in which the information on its expenditure is collected. One is to ask the household to record all expenditure as soon as it is made in a specially provided notebook. The other is to ask the household to recall its expenditure over a given period of time in the past. The first method is more demanding on the household and excludes those that have not enough literacy or discipline. The second one is clearly less reliable.

The type of household considered depends on the purpose of the survey. Many of the early surveys were conducted to obtain information about poverty and concentrated therefore on low income households. Surveys are sometimes used to obtain appropriate weights for cost of living indexes. If these indexes are used to gauge the real income of wage earners, the collected budgets are those of families with a wage earner as head. If a household budget survey is meant to provide detailed information on consumer habits in general one will try to have a more or less representative sample from the population. Because certain types of consumer units like one-member households, collective units (for example, boarding schools), illiterate and irregular families, families changing residence and composition tend to be excluded from the sample one should not expect full representativity.

Still, as a source of detailed information on consumer behaviour with respect to very finely detailed commodities and services there is no alternative of the same quality to the budget survey. It can provide useful information about the extent of the market for a certain product or about the type of families that have special interest in certain expenditures.

The degree of detail has a limit, however. To keep track of all available shades of quality is virtually infeasible. Some aggregation over qualities is unavoidable, which might cause apparent differences in unit prices owing to differences in the quality composition of the aggregate.

Other problems are the value of gifts and of the consumed own production of farmers. They are usually solved in a pragmatic way.

As such a budget survey gives synchronic information. It observes a group of families during the same, rather short, period of time. It provides no information about the changes in behaviour over time. For example, usually prices do virtually not change during the period of observation, which rules out the possibility to study responses to price changes. Clearly, repetition of a budget survey adds a time dimension and opens the possibility to analyse time dependent changes. So-called *panel studies* are such repeated surveys where in each new round a part of the participants is replaced by other households with similar characteristics. The combination of diachronic and synchronic information such a panel offers is of great value.

The value of budget surveys is also increased if the expenditure behaviour of a household can be related to its various characteristics like residence, race, degree and type of labour participation, composition of the family (sex, age), education, owned or rented housing, ownership of durables, hobbies, pets, and so on. Such additional information is not always fully collected, or frequently not made available in detail to the public in order to avoid identification of the participants by outsiders.

Budget surveys have their limits. They are usually costly. The participants will not (accurately) respond to certain questions for various reasons. As already mentioned some of the collected information is not published to protect the participants.

Normative Budgets

The early interest on household budgets had a humanitarian motive. The actual expenditure of a family was compared with what a family of that type needs. These needs were determined in the form of a set of minimal quantities of various items (usually foodstuffs). Given corresponding actual prices the total means to purchase these quantities can be calculated and constitutes a normative budget. A household with an income below the norm qualifies for support. The selection of the minimal quantities is not without ambiguity. A norm very close to a physical survival level leaves little choice, but norms corresponding to social viability in a modern society are difficult to define in an indisputable way.

Analysis of Household Budgets

Differences in the expenditure patterns across families can be attributed to three factors: (i) variation in available means, (ii) variation in relative prices, (iii) differences in other family characteristics. These three topics will be taken up one by one in what follows. Among differences in family characteristics differences in size and composition of the household have historically played an important role. This justifies their discussion in a special section.

Engel Curves

The relation between demand for (or expenditure on) a good and the means of a consumer unit is frequently named *Engel curve* after Ernst Engel (1857) who on the basis of his analysis of the budget data of Ducpétiaux stated his law that the share of food in total expenditure is a decreasing function of the level of prosperity of the family. Engel's law appears to hold almost universally – see, for example, Houthakker (1957). One would like to amend it somewhat in the sense that it pertains to the share of staple food items in the budget rather than to that of all types of food. This distinction is less relevant for families in the lowest income groups than for more well-to-do ones.

A generalization of Engel's law states that with increasing prosperity the budget share of any good initially increases (except for some basic subsistence good) and later on decreases. Increasing budget shares correspond with the 'luxury' status of a good. Its budget or income elasticity of demand is larger than one. A 'necessity' has a decreasing budget share. Its budget or income elasticity is smaller than one. Note that decreasing budget share does not imply decreasing quantity, although this might occur. A necessity with a diminishing quantity is an 'inferior' good. One with a constant or growing quantity as prosperity increases is a 'normal' or 'superior' good. A commodity may go through a prosperity cycle, being a luxury and normal commodity for the very poor, a necessity and normal commodity for the better-off and perhaps a necessity and inferior commodity for the very rich.

For empirical research the issue of the measurement of 'available means' or 'level of prosperity' arises. Total wealth of a family, defined as the market value of its real and financial assets plus the present value of expected future income from other sources, might be the most appropriate concept. It escapes direct measurement, however, because it is based on subjective expectations while also the ownership of assets is not well observed. The same holds for the concept of permanent income which is the amount of money which may be consumed leaving total wealth unchanged. Current income is apt to include transitory components and is presumably anyway not faithfully recorded. The amount of total expenditure is usually readily available and might be closely related to total wealth or permanent income. This makes it an attractive proxy for available means or prosperity when explaining the pattern of expenditure.

The explanation of expenditure patterns as a function of total expenditure is a typical allocation model. Let q_i denote the quantity bought of item i ($i = 1,\ldots, n$) and p_i its price. Then

$$e_i = p_i q_i \tag{1}$$

is the amount paid. By definition total expenditure m is given by

$$\sum_{i=1} e_i = m \tag{2}$$

The Engel curves $E_i(m)$ should satisfy the following adding-up condition

$$\sum_{i=1} \mathrm{E}_i(m) = m \tag{3}$$

Condition (2) is automatically satisfied by the data. Property (3) depends in part on the functional form of the Engel curves used in the explanation. Linear Engel 'curves' easily satisfy (3) but cannot deal with the possibility that goods change from superior to inferior commodities over the prosperity cycle. They can also not guarantee non-negative consumption. An Engel curve system satisfying (3), allowing for a prosperity cycle and excluding non-negative consumption is still not available.

Zero consumption for certain items, a common phenomenon, provides another complication of Engel curve analysis. Of the full range of commodities on the market only a limited number will be bought by a given family. The statistical distribution of e_i conditional on m is then a bimodal one with zero for the smaller mode. Least-squares regressions using all data will not estimate either of these modes. Leaving out observations with a zero consumption level estimates correctly the non-zero mode. This leads to problems, however, if one wants to estimate a full system of Engel curves simultaneously, because only a very few families will report non-zero expenditures for all items.

In the case when zero consumption is owing to the fact that the amount of money needed to acquire the smallest available quantity of a desirable item is more than the household can afford there is a link with prosperity, because for a sufficiently high level of prosperity the household will buy the commodity. Still there is a statistical complication due to the mixture of discreteness (zero versus non-zero) and continuity (if non-zero how much?) of the relationship. Tobin (1958) gives an elegant statistical solution to this problem using a combination of qualitative and quantitative response modelling – see also Maddala (1983).

Price Responses

Differences in expenditure pattern may be also due to differences in prices paid by the households. As already mentioned budget surveys do not usually provide a good data base to observe and analyse price responses. Sometimes, however, there is price variation owing to geographical distance or other supply factors. If the prices are reported their effects can be analysed. The same is true for panel data with price variations over time (and space).

If price information is available it can be employed to explain variation in expenditure patterns. One may write

$$q_i = fi(m, p_1, \ldots, p_n) \qquad (4)$$

to express that the quantity purchased of commodity i does depend on total means m or total expenditure (see above) and the prices (p_i) of all goods in the budget. On the assumption that given his budget m the consumer will select the set of quantities that satisfies him best economic theory has specified several properties of price responses – see, for example, Deaton and Muellbauer (1980). These can be taken into account in the estimation or can be tested on their empirical validity.

Effects of Household Characteristics

Variation in available means and relative prices account for a relatively small part of the variation in expenditures over various items across families. The remaining variation is to be attributed to differences in preferences, tastes. Such differences are not necessarily random. Engel (1857) previously pointed out the importance of differences in climate. Later (1883, 1895), he elaborated the effects of household composition. Next to such physiological factors there may be cultural ones like race and religion. The profession of the head of the household appears to have explanatory power – see Prais and Houthakker (1955). It is a mixture of a physiological (physical effort) and a sociological (reference group) effect. The urban/rural difference matters too. In part this difference can reflect differences in price structure, in part differences in proximity of shops and availability of public transportation. There may be also a sociological aspect to this difference. More economic in nature are differences in ownership of household appliances and in the extent to which the mother participates in the labour market. A variable like years of school education overlaps largely with the factors already mentioned.

As far as these factors are purely qualitative they can be taken into account in two ways. One is to split the sample into cells which are qualitatively homogeneous and estimate for each cell the effect of quantitative determinants. Some of these cells might be sparsely populated. Another possibility is to treat the qualitative factors as dummies (covariance analysis). A formal test can then supply the answer to the question whether a certain quality makes a significant difference for the expenditure pattern.

The effects of household characteristics on expenditure patterns can be analysed in economic terms as follows. Consider two households with the same m and the same price system. They belong to a different social class and have a different consumption behaviour. They can afford each others' life style but clearly prefer their own. By way of a system of positive and negative subsidies one can induce one household to purchase the same set of quantities as the other household. This change involves in part a reduction in real income because it cannot any more afford the originally preferred set of purchases and in part a change in relative prices. Since the argument is symmetric in both families one cannot say that one is better off than the other. For welfare comparison one needs other information than that on purchasing behaviour supplied by a budget survey.

Household Size and Composition

The treatment of household size and composition is a subject of long standing in household budget analysis. As a first approximation the effect of differences in family size can be taken into account by considering average expenditure per member as the variable to be explained and average total expenditure per member as the appropriate explanatory variable. Obviously, this approach ignores the possibility that members of the same household have different basic needs. To handle this issue one has experimented with a rescaling of the number of members into a number of equivalent members. Engel (1883) proposed as unit the 'quet', which corresponded with a newborn baby. The normal weight of a person of given age and sex divided by that of the infant defined the number of quets for that person. Family size was measured by the sum of quets of the members.

More recent approaches, however, take the male adult as the unit and the members of the household are converted into an equivalent male adult. The Amsterdam scale for example, assigns a factor 0.9 to a female adult and a factor of 0.1 to a one-year-old child.

The early equivalent adult scales reflected mostly physiological differences. They were usually meant to correct for household size effects on consumption of food and were established *a priori*. Many different scales have been introduced. Sydenstricker and King (1921) introduced commodity specific equivalent adult scales together with an overall scale. This can be formalized by writing the Engel curve as

$$e_{ih}/s_{ih} = \mathrm{E}_i(m_h/s_h) \qquad (5)$$

where i denotes commodity, h household h, s_{ih} is the size of household h using weights specific for i and s_h, is the size of household h using overall weights. According to (5) the addition of a member to the household will have a direct, usually positive, effect on the expenditure on i by way of its impact on s_{ih}, and an indirect, usually negative, effect by way of its reduction of total means per equivalent adult. Sydenstricker and King also suggested the estimation of the weights of the scales along with the other parameters of the Engel curves. These contributions went largely unnoticed until reintroduced by Prais and Houthakker (1955).

There are at least two problems with formulation (5). The first one is that it is nonlinear in the size variables causing estimation problems. The second problem is the one of (in) compatibility of the overall size variable s_h with the commodity specific scales. On the basis of (3) and (5) one has that

$$m_h = \sum_i s_{ih}\mathrm{E}_i(m_h/s_h) \qquad (6)$$

which may be seen as an implicit definition of s_h, involving m_h, the total means of the household. To put this another way, given the specific weights the overall weights are determined and they are generally not independent of m_h. Estimating the overall weights independently of the specific ones leads to problems.

Another approach using commodity specific scales is to estimate

$$e_{ih} = \pi_{ih} fi(m_h, \pi_{1h}, \ldots, \pi_{nh}) \quad (7)$$

where f_i ( ) is a demand function with as prices $\pi_{jh} = p_j s_{jh}$. The family size effect is in this way assimilated with a price effect.

Also here, an increase in the family results in a direct, positive, effect via the π_{ih} factor right after the equality sign in (7) and an indirect effect which the changes in the relative prices exert on demand. This latter effect takes the place of the overall effect in (5). This reformulation is formally justified by redefining the utility function of the consumer unit in terms of $x_{ih} = q_{ih}/s_{ih}$, that is in quantities per equivalent adult. This function is then maximized subject to the budget definition written as

$$\sum_i \pi_{ih} x_{ih} = m_h \quad (8)$$

The optimal x_{ih}, are given by $f_i()$. Multiplying by π_{ih} yields (7). This approach, proposed by Barten (1964), avoids incompatibility problems.

There is an identification issue, however – see Muellbauer (1975, 1980). One needs additional information on the weights of the specific scales in order to identify them from the observations. This prior information can take the form of assigning for each age–sex class the value of the weight in one of the n specific scales (for example 1 for male adult in the scale for tobacco; 1 for infants in the scale of babyfood; 1 for female adults in the scale for cosmetics). One can also formulate a restriction, again for each age–sex class, involving weights in more than one specific scale (for example, equality of the weights of teenagers in the scales for bread and for meat).

The specification of the scales deserves some further discussion. Define

$$s_{ih} = \sum_j b_{ij} c_{jh} \quad (9)$$

with c_{jh} being the number of members of household h in age–sex class j, while the b_{ij}, are the corresponding weights. The linear discrete specification treats the b_{ij}, as constants which are either estimated or fixed extraneously. The continuous scale approach of Friedman (1952) makes the b_{ij} a continuous function of age and sex: hm_i, (age j) for male members and hf_i (age j) for female members. Various restrictions on these functions result in scales which are smooth at the end points and parsimonious in parameters. There may be a problem, however, in obtaining a proper monotone behaviour.

A related issue is that of incorporating scale effects into the family size measure. Kapteyn and van Praag (1976) let the weight depend on the age rank (r) of the member of class j in the family. Following this approach one could specify as weight

$$b_{ijr} = b_{ij0} + b_{ij1}(r-1) + b_{ij2}(r-1)^2$$

The measurement of family size effects has sometimes been motivated by the desire to obtain a more objective, empirical basis for family allowance schemes. The welfare implications of varying family composition are not unambiguous, however. As already stated above when discussing the impact of differences of family characteristics in general one cannot conclude directly from observable behaviour that such differences imply being better or worse off. What holds in general is then also true for differences in family composition.

See Also

- ▶ Characteristics
- ▶ Consumer Expenditure
- ▶ Demand Theory
- ▶ Engel's Law
- ▶ Hedonic Functions and Hedonic Indexes
- ▶ Separability

Bibliography

Barten, A.P. 1964. Family composition, prices and expenditure patterns. In *Econometric analysis for national economic planning*, ed. P.E. Hart, G. Mills, and J.K. Whitaker. Butterworths: London.

Deaton, A., and J. Muellbauer. 1980. *Economics and consumer behavior*. Cambridge: Cambridge University Press.

Ducpétiaux, A.E. 1855. *Budgets économiques des classes ouvriéres en Belgique*. Brussel: Hayez.

Eden, F.M. 1797. *The state of the poor*, 3 vols. London: J. Davis. Facsimile edn, London: Frant Cass & Co., 1966.

Engel, E. 1857. Die Productions- und Consumtions verhältnisse des Königreichs Sachsen. *Zeitschrift des Statistischen Büreaus des Königlich Sächsischen Ministerium des Innern* 8(9). Reprinted in Bulletin de l'Institut International de la Statisque 9 (1985): 1–54.

Engel, E. 1883. Der Werth des Menschen; I. Teil: Der Kostenwerth des Menschen. *Volkswirtschaftliche Zeitfragen*, vols. 37–38, 1–74. Berlin: L. Simion.

Engel, E. 1895. Die Lebenskosten belgischer Arbeiter-Familien früher und jetzt. *Bulletin de l'Institut International de Statistique* 9: 1–124.

Friedman, M. 1952. A method of comparing incomes of families differing in composition. *Studies in Income and Wealth* 15: 9–24.

Houthakker, H.S. 1957. An international comparison of household expenditure patterns, commemorating the centenary of Engel's law. *Econometrica* 25: 532–551.

Kapteyn, A., and B. Van Praag. 1976. A new approach to the construction of family equivalence scales. *European Economic Review* 7: 315–335.

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

Muellbauer, J. 1975. Identification and consumer unit scales. *Econometrics* 43: 807–809.

Muellbauer, J. 1980. The estimation of the Prais-Houthakker model of equivalence scales. *Econometrica* 48: 153–176.

Prais, S.J., and H.S. Houthakker. 1955. *The analysis of family budgets*. Cambridge: Cambridge University Press.

Sydenstricker, E., and W.I. King. 1921. The measurement of the relative economic status of families. *Journal of the American Statistical Association* 17: 842–857.

Tobin, J. 1958. Estimation of relationships for limited dependent variables. *Econometrica* 26: 24–36.

Household Portfolios

Michael Haliassos

Abstract

This entry presents recent work on portfolio behaviour of households and its possible departures from optimal behaviour. Topics include the role of household characteristics in influencing participation in stockholding and portfolio shares conditional on participation; portfolio implications of housing and housing debts; and portfolio coexistence of consumer debt, liquid assets and illiquid assets, with emphasis on credit card debt.

Keywords

Age effects; Asset allocation; Asset ignorance; Asset location; Asset trading; Bankruptcy; Bequest motives; Borrowing constraints; Business equity; Cohort effects; Computational methods; Conditional portfolio share; Consumer debt; Consumption risk; Correlation between income risk and stock returns; Credit card debt; Credit cards; Debt; Debt refinancing; Delinquency; Diversification; Earnings shocks; Elasticity of substitution; Epstein–Zin preferences; Equities; Equity premium; Financial wealth; Fixed entry costs; Fixed-rate mortgages; Participation costs; Home equity loans; Homeownership; Household finance; Household portfolios; Housing; Housing collateral; Hyperbolic discounting; Interest-rate wedge; Marginal investors; Mortgages; Mutual funds; Real estate; Refinancing; Retirement; Retirement accounts; Risk aversion; Social interactions; Stockholding; Stockholding participation rate; Stockholding puzzle; Stockholding risk; Stocks; Strategic default motive; Time effects; Trust; Wealth distribution

JEL Classifications

G11

Household portfolios comprise the array of assets – financial (such as liquid accounts, stocks, bonds, and shares in mutual funds) and real (such as primary residence, investment real estate, and private businesses) – as well as liabilities held by a household, such as mortgages and consumer debt. This article focuses on three areas of active research – stockholding, housing, and credit cards – with respective household participation rates for the United States of the order of 50 per cent, two- thirds, and two- thirds. European participation rates vary. Stockholding participation approaches 60 per cent in Sweden and 40 per cent in the UK, but it is less than 20 per cent in France, Germany, and Italy. Homeownership rates are closer to that of the United States, but in some countries, such as Germany, the majority does not own a home. The features of credit cards vary across European countries. In some countries, households have only debit cards linked to accounts with overdraft facilities.

The study of household portfolios, or 'household finance', is a partner to corporate finance and asset pricing, and it bridges economics and finance by extending analyses of saving to incorporate portfolio choice. It has grown considerably since the early 1990s, along with the complexity of household portfolios, in the face of 'supply side' developments encouraging risky asset holding. Privatization of public utilities in Europe was often accompanied by broad campaigns to educate households on the nature and benefits of stockholding. The demographic transition encouraged introduction of tax-deferred retirement accounts, promoted through educational campaigns, first in the United States and subsequently in Europe. The internet facilitated provision of information, opening of accounts, and trading internationally.

The development of household-level databases has in turn facilitated empirical research by allowing study of overall portfolios and their links to demographics and attitudes. Modern computational methods have enhanced understanding of behaviour towards non-diversifiable, background risk regarding income or health expenditures. Observed portfolio behaviour often differs from predictions of standard models, creating puzzles variously attributed to inadequate models or 'investment mistakes'.

Stockholding

Understanding household stockholding is important, as it embodies key aspects of behaviour towards risk. In most countries, the majority of households holds no stocks, even indirectly through mutual funds, retirement, or managed accounts (Guiso et al. 2001, 2003). Exceptions were Sweden and the United States in 2001 (57 per cent and 52 per cent, respectively), but the United States fell back to 48 per cent in 2004. Non-participation despite an expected return premium ('equity premium') is inconsistent with standard expected utility maximization and constitutes the 'stockholding puzzle' (Mankiw and Zeldes 1991; Haliassos and Bertaut 1995). For a non-stockholder, stocks dominate bonds in expected return and do not contribute to consumption risk as they have zero covariance with consumption.

Various explanations have been proposed for limited participation in stock markets, given its widespread nature. Restrictions preventing borrowing at the riskless rate and short sales of stock yield zero stockholding, but only for poor households with no assets (Haliassos and Michaelides 2003). Positive correlation between labour income risk and stock returns, coupled with short sales constraints, could justify zero stockholding among households intending to short stocks to hedge income risk, but is exhibited in practice by households likely to hold stocks – for example, the more educated and entrepreneurs.

The most widely accepted cause of limited participation is fixed entry or participation costs, actual or perceived, that discourage small potential investors. Costs can be wide-ranging, from brokerage costs to costs of one's time devoted to monitoring the stock market. In their presence, factors contributing to higher costs or lower desired stockholding, such as risk aversion or low resources, become relevant for non-participation. An interest-rate wedge between borrowing and saving rates coupled with an empirically based assumption that borrowing rates are roughly equal to the expected return on equity also generates limited stock demand. Although Davis et al. (2006) offered this as an alternative to fixed costs for explaining non-participation, it could usefully serve also as a complement. Empirical estimates by Paiella (2001) and Vissing Jorgensen (2002), and numerically computed costs in Haliassos and Michaelides (2003) imply that relatively small fixed costs could justify observed patterns of non-participation.

The empirical participation literature provides various findings consistent with the presence of fixed costs (see contributions in Guiso et al. 2001; Rosen and Wu 2004). More educated, financially alert, healthy households that belong to ethnic or education groups traditionally targeted by the financial sector are likely to face lower entry costs and to be more likely to participate, consistent with empirical findings. Similarly, households with greater expressed willingness to bear risk and those who do not perceive binding borrowing constraints are more likely to plan sizeable stock holdings and thus to overcome any given entry costs.

Empirical studies also point to other, often ignored, factors, which seem relevant for non-participation by those unlikely to be small investors, such as the rich. Limited social interactions and associated opportunities to exchange stockholding experiences, or lower expressed willingness to trust others, contribute to non-participation (Hong et al. 2004; Guiso et al. 2005). This can justify non-stockholding by some rich, in addition to possible substitution of private businesses for stocks (Heaton and Lucas 2000). Non-participation also arises naturally if there is widespread ignorance of certain assets. Guiso and Jappelli (2005) found that only one-third of Italian households have simultaneous knowledge of stocks, mutual funds, and managed accounts. Moreover, although most of the literature has largely ignored tax considerations, tax laws have been shown to affect asset allocation, asset location, and trading (Bergstresser and Poterba 2004).

Given that stockholding participation has increased, it is important to understand its economy-wide implications, as well as its future prospects in the face of changing stock market conditions. The limited existing theoretical literature already points to ambiguous effects of increased participation on wealth distribution (Peress 2004; Guvenen 2006). Since certain characteristics were empirically found to encourage participation, the composition of the stockholder pool is likely to change as participation spreads. If increased participation means progressive entry of ‘marginal’ investors with more limited resources and investment ability, it can contribute to lower stockholding levels, overtrading that lowers realized returns, and possibly greater wealth inequality. Households with lower education and resources have been shown to be more prone to ‘investment mistakes’ in terms of (non) participation, (under)diversification, and lack of debt refinancing (Campbell 2006). Bilias et al. (2005, 2006) find evidence that the 1990s upswing attracted to the US stockholder pool households with characteristics, attitudes, and practices conducive to small stockholding levels, but this was reversed by entry and exit following the downswing. Overtrading characterizes households with brokerage accounts, but not the general population.

Households that do clear the participation hurdle need to decide what portfolio share to hold in stocks. Theory generates strong predictions on how this conditional portfolio share should be affected by household characteristics, but these are often not confirmed by the data. For example, under expected utility, constant relative risk aversion, and income risk, the share is predicted to fall with age or with the ratio of cash on hand to permanent income (Cocco et al. 2005). Either factor causes households to rely more on assets rather than on human wealth for financing consumption, and this reduces willingness to invest heavily in stocks. Yet the wealthy have conditional portfolio shares of risky assets about double those for the remaining population (Carroll 2001). Although it is impossible to identify separately age, time, and cohort effects using cross-sectional data (Ameriks and Zeldes 2005), regressions setting cohort effects to zero fail to find consistent dependence of conditional portfolio shares on age or resources (Guiso et al. 2003). Data from retirement accounts show great inertia in changing portfolio shares (Ameriks and Zeldes 2005), while studies using discount brokerage accounts find overtrading (Barber and Odean 2000). Representative data imply inertia in the population at large (Brunnemeier and Nagel 2005; Bilias et al. 2006).

Gomes and Michaelides (2005) exploited the additional flexibility of departures from expected utility maximization in the form of Epstein-Zin preferences to approximate observed portfolio shares more closely. Under expected utility maximization and preferences exhibiting constant relative risk aversion, the risk aversion, prudence, and intertemporal elasticity of substitution parameters are linked. Lowering risk aversion (which increases the risky portfolio share) lowers prudence (thus precautionary wealth) and raises elasticity of substitution (thus saving for retirement). Epstein-Zin preferences allow simultaneous lowering of risk aversion and elasticity parameters. Households with low risk aversion, prudence, and elasticity parameters smooth earnings shocks with small assets, and almost never invest in equities in the presence of fixed costs. Those who clear the participation hurdle have higher parameters and moderate portfolio shares in stocks.

Housing

Although stocks are an interesting part of a household's portfolio, housing is the largest part, and it is both important and challenging to understand how homeownership interacts with the rest of the portfolio. Due to housing investment, younger and poorer investors have limited wealth to invest in stocks (Cocco 2005). Payment commitments on mortgages may also discourage risky asset holding. Renters accumulating down payments for a house may be unwilling to jeopardize their accumulations by assuming financial risk. On the other hand, homeowners have access to home equity loans and other collateralized loans not available to renters, and ability to borrow may encourage financial risk taking.

Understanding housing as a portfolio element cannot be accomplished without studying the structure of mortgages and their risk implications, on which there is surprisingly little research. Campbell and Cocco (2003) show that adjustable-rate mortgages are attractive to households that face no binding borrowing constraints but large inflation risk relative to real interest rate risk, and to potentially borrowing- constrained households with low risk aversion. They are unattractive to constrained, highly risk-averse households. Sluggishness to refinance despite significant rate drops has been found, especially among households with less wealth or financial sophistication (Campbell 2006).

Credit Card Debt

Having discussed some household assets, financial and real, let us now turn to household debt and its coexistence with assets, which received considerable attention as participation rates and median levels of indebtedness grew. Credit card debt behaviour is topical at the time of writing (2006), given increases in bankruptcy and delinquency rates that cannot be attributed to changes in debtor characteristics or supply factors (Gross and Souleles 2002a). Gross and Souleles (2002b) documented two US credit card debt puzzles: (a) coexistence of high-interest card debt with substantial asset accumulation for retirement, suggesting a combination of short-run impatience with considerable patience for longer run objectives; and (b) coexistence of credit card debt with sizeable low-interest liquid assets that could have been used to pay it off.

The nature of these puzzles and the wide perception that credit cards make it difficult to control spending have led researchers mainly to behavioural explanations. Laibson et al. (2003) showed that a single rate of time preference has problems generating the former coexistence, and proposed hyperbolic discounting. The current self borrows because of short-run impatience. Accumulating illiquid assets is a way to control the future self, who will be impatient as retirement approaches.

The second puzzle seems to run against usual notions of arbitrage. Bankruptcy law allows households to rescue some assets, and this creates strategic default motives that discourage paying off debt, but strategic defaulters could avoid interest costs by buying exempt assets right before filing (Gross and Souleles 2002b).

Bertaut and Haliassos (2002) and Haliassos and Reiter (2005) propose an 'accountant-shopper' model that generates both types of coexistence. The accountant self (or household member) revolves debt (partly) to constrain the amount charged by the impatient credit-card shopper, but this is not inconsistent with accumulating assets for retirement or other purposes. Caplin and Leahy (2004) model an absent-minded consumer who does not keep track of his spending. Credit cards may lead to overspending because they provide less information on spending flows than cash transactions.

Household portfolios entail numerous research challenges. They include further understanding of: interactions between real and financial assets and debts; sources of international differences in portfolio structure, especially around retirement; which part of unexplained portfolio behaviour is due to investment mistakes rather than model shortcomings; how labour market behaviour influences portfolios; the role of intra-household bargaining and risk sharing; the role of inattention and financial advice in the face of agency; and other incentive problems.

See Also

- ▶ Consumption-Based Asset Pricing Models (Theory)
- ▶ Credit Card Industry
- ▶ Financial Market Anomalies
- ▶ Household Surveys
- ▶ Inheritance and Bequests
- ▶ Intertemporal Choice
- ▶ Non-expected Utility Theory
- ▶ Precautionary Saving and Precautionary Wealth
- ▶ Recursive Preferences
- ▶ Risk Aversion

Bibliography

Ameriks, J., and S. Zeldes. 2005. *How do household portfolio shares vary with age?* Mimeo: Columbia University.

Barber, B., and T. Odean. 2000. Trading is hazardous to your wealth: The common stock investment performance of individual investors. *Journal of Finance* 55: 773–806.

Bergstresser, D., and J. Poterba. 2004. Asset allocation and asset location: Household evidence from the Survey of Consumer Finances. *Journal of Public Economics* 88: 1893–1916.

Bertaut, C., and M. Haliassos. 2002. *Debt revolvers for self-control*. Federal Reserve Board: Mimeo.

Bilias, Y., D. Georgarakos, and M. Haliassos. 2005. Equity culture and the distribution of wealth. Working paper no. 2005/20. CFS, University of Frankfurt.

Bilias, Y., D. Georgarakos, and M. Haliassos. 2006. Portfolio inertia and stock market fluctuations. Working paper no. 2006/14. CFS, University of Frankfurt.

Brunnermeier, M., and S. Nagel. 2005. *Do wealth fluctuations generate time-varying risk aversion? Micro-evidence on individuals' asset allocation*. Mimeo: Princeton University.

Campbell, J.Y. 2006. Household finance. *Journal of Finance* 61: 1553–1604.

Campbell, J., and J. Cocco. 2003. Household risk management and optimal mortgage choice. *Quarterly Journal of Economics* 118: 1449–1494.

Caplin, A. and Leahy, J. 2004. The absent-minded consumer. Working paper no. 10216. Cambridge, MA: NBER.

Carroll, C. 2001. Portfolios of the rich. In *Household portfolios*, ed. L. Guiso, M. Haliassos, and T. Jappelli. Cambridge, MA: MIT Press.

Cocco, J. 2005. Portfolio choice in the presence of housing. *Review of Financial Studies* 18: 535–567.

Cocco, J., F. Gomes, and P. Maenhout. 2005. Consumption and portfolio choice over the life-cycle. *Review of Financial Studies* 18: 491–533.

Davis, S., F. Kubler, and P. Willen. 2006. Borrowing costs and the demand for equity over the life cycle. *Review of Economics and Statistics* 88: 348–362.

Gomes, F., and A. Michaelides. 2005. Optimal life-cycle asset allocation: Understanding the empirical evidence. *Journal of Finance* 60: 869–904.

Gross, D., and N. Souleles. 2002a. An empirical analysis of personal bankruptcy and delinquency. *Review of Financial Studies* 15: 319–347.

Gross, D., and N. Souleles. 2002b. Do liquidity constraints and interest rates matter for consumer behavior? Evidence from credit card data. *Quarterly Journal of Economics* 117: 149–185.

Guiso, L., and T. Jappelli. 2005. Awareness and stock market participation. *Review of Finance* 9: 537–567.

Guiso, L., M. Haliassos, and T. Jappelli. 2001. *Household portfolios*. Cambridge, MA: MIT Press.

Guiso, L., M. Haliassos, and T. Jappelli. 2003. Stockholding in Europe: Where do we stand and where do we go? *Economic Policy* 36: 117–164.

Guiso, L., P. Sapienza, and L. Zingales. 2005. Trusting the stock market. Working paper no. 11648. Cambridge, MA: NBER.

Guvenen, F. 2006. Reconciling conflicting evidence on the elasticity of intertemporal substitution: A macroeconomic perspective. *Journal of Monetary Economics* 53: 1451–1472.

Haliassos, M., and C. Bertaut. 1995. Why do so few hold stocks? *Economic Journal* 105: 1110–1129.

Haliassos, M., and A. Michaelides. 2003. Portfolio choice and liquidity constraints. *International Economic Review* 44: 143–178.

Haliassos, M., and M. Reiter. 2005. Credit card debt puzzles. Working paper no. 2005/26. CFS, University of Frankfurt.

Heaton, J., and D. Lucas. 2000. Asset pricing and portfolio choice: The importance of entrepreneurial risk. *Journal of Finance* 55: 1163–1198.

Hong, H., J. Kubik, and J. Stein. 2004. Social interaction and stock market participation. *Journal of Finance* 59: 137–163.

Laibson, D., A. Repetto, and J. Tobacman. 2003. *A debt puzzle*. In *Knowledge, information, and expectations in modern economics: In honor of Edmund S. Phelps*, ed. P. Aghion et al. Princeton: Princeton University Press.

Mankiw, N., and S. Zeldes. 1991. The consumption of stockholders and non-stockholders. *Journal of Financial Economics* 29: 97–112.

Paiella, M. 2001. Limited financial market participation: A transaction cost-based explanation. Working paper no. 01/06. London: IFS.

Peress, J. 2004. Wealth, information acquisition and portfolio choice. *Review of Financial Studies* 17: 879–914.

Rosen, H., and S. Wu. 2004. Portfolio choice and health status. *Journal of Financial Economics* 72: 457–484.

Vissing Jorgensen, A. 2002. Towards an explanation of household portfolio choice heterogeneity: Nonfinancial income and participation cost structures. Working paper no. 8884. Cambridge, MA: NBER.

Household Production

Richard A. Berk

Even a casual survey of recent developments in neoclassical economics will reveal a self-conscious intellectual imperialism. Substantive areas traditionally the private preserve of other social science disciplines have experienced significant incursions: fertility, voting behaviour, crime, education, and others. But perhaps the most visible and influential expansion of neoclassical economics has been into the formation, functioning, and dissolution of families. Under the banner of the 'new home economics', conventional utility maximization with fixed preferences claims to provide explanations for an enormous variety of decisions made by households and their members.

There is little doubt that these developments have gained a number of adherents. Among economists, much of this success can be attributed to the obvious appeal of creatively moving a well-known theoretical apparatus to a novel setting. But, there have also been converts from outside economics. For them, perhaps more important than the merits of a new home economics is the absence of persuasive alternatives; the new home economics is effectively directed at the soft underbellies of other social-science disciplines.

In particular, family sociology has conventionally applied conceptual frameworks placing instrumental activities in the market and expressive activities in the home (e.g. Blood and Wolf 1960). It does not occur to most sociologists, therefore, to think about households as 'productive', nor to see household activities as the concrete manifestation of production functions; home life is about affect. In addition, many sociologists are inductively inclined, preferring to work up from data not down from theory. Their literature, as a result, is rich in facts that are not easily placed under a single theoretical rubric; perhaps knowing more has meant knowing less.

Do Families Really Optimize?

General criticisms of utility maximization are well known well known and not be reviewed here (e.g. Hollis and Nell 1975; Leibenstein 1976; Lesourne 1977; Simon 1978). For at least two reasons, however, utility maximization may be especially problematic within the family setting.

First, all neoclassical economic perspectives on households require that households optimally allocate their resources. This assumption has be supported in part with the argument that inefficient households either will not form or will not survive (Becker 1981, pp. 40–2, 66–82, 219–36); households form and dissolve within a 'marriage market', which performs the same functions as any other free market.

However, as Blaug (1980, p. 119) has observed in a somewhat different context,

> to survive, it is only necessary to be better adapted to the environment than one's rivals, and we can no more establish from natural selection that surviving species are perfect than we can establish from economic selection that surviving firms are profit maximizers.

At best, therefore, only family partnerships better suited to the environment need survive; the survivors are not required to be optimally adapted. In other words, the assumption of optimization cannot be justified by recourse to market forces.

Second, there is some scepticism about whether families can adjust quickly to a changing environment. Schultz (1974, p. 6) observes,

> The typical family that we observe, especially in rich countries, lives in an economy in which economic conditions are and have been changing substantially over time. As these changes occur, thinking in terms of economics, there are presumably responses – responses in the age at which marriage occurs, responses in spacing and number of children, and responses in the amount of family resources devoted to investment in children. Furthermore, before these families have fully adjusted and have arrived at an equilibrium with respect to any given economic change, additional and unexpected changes will have occurred. Thus, the families we observe are seldom, if ever, in a state of economic equilibrium.

Whose Wellbeing Is Being Maximized?

Almost all neoclassical perspectives on the family assume that the decision-making unit is the family as a whole, and that there is a single household utility function. In the face of considerable skepticism (e.g. Nerlove 1974; Mancer and Brown 1980; McElroy and Horney 1981; Witte et al. 1984), Becker's use of 'altruism' (1981, pp. 172–201) is perhaps the best justification.

However, according to Ben-Porath (1982, p. 54), Becker's formulation requires some very strong assumptions, such as perfect information, despite powerful incentives for household members not to reveal accurately how well off they are. In a similar manner, Pollak (1985, p. 599) argues that Becker's results do not depend on altruism per se, but on 'implicit assumptions about power, or equivalently, about the structure of the bargaining game'. Perhaps most important, there is lots of evidence that *ongoing* conflict and coercion characterize a significant number of households. For example, one is a very long way from a single utility function when a recent report from the United States Attorney General's Office asserts (Hart 1984, p. 11).

> Battery is a major cause of injury to women in America. Nearly a third of female homicide victims are killed by their husbands or boyfriends. Almost 20 percent of all murders involve family relationships. Ascertainable reported cases of child abuse and neglect have doubled from 1976 to 1981. In addition to one million reported cases of child maltreatment, there may be another million unreported cases. Untold numbers of children are victims of sexual abuse, and uncounted older persons suffer abuse.

What About Joint Production?

Given the linear budget constraint, Pollak and Wachter (1975) point out that joint production is effectively excluded from the recent neoclassical approaches to the family. Thus, it is impossible to obtain psychic gratification and a concrete household commodity from the same household activity (e.g. cooking a meal). Berk and Berk (1983, p. 388) observe that joint production could be incorporated with a nonlinear budget constraint, but *additional assumptions* would have to be made. For example, one would need to specify through the appropriate elasticities how responsive to changes in family money income each of the joint products happened to be. It is very unlikely that data could be found to inform meaningfully such an exercise.

The key question, therefore, is whether joint production is common, and what little research that exists (e.g. Berk and Berk 1979, pp. 237–250), coupled with everyday experience, suggests that it is widespread. One has only to introspect a bit about the nature of child care.

Are There Constant Returns to Scale?

The assumption of constant returns to scale also creates difficulties. In recent statements (e.g. Becker 1981), household commodities are rather general entities such as prestige, health, esteem and the like. There is no reason to assume that for these outputs, constant returns to scale hold. Indeed, common experience suggests quite the opposite.

For example, doubling the amount of food one ingests will affect one's health in rather different ways depending on how much food one ordinarily ingests. For malnourished individuals, a rather dramatic improvement in health will probably be seen. For well-fed individuals, little improvement will result, and depending on the kind of food eaten, health could actually decline. In short, the linear budget constraint is once again inappropriate so that the usual formulations of the household production function no longer yields signed results. And again, the use of a nonlinear budget constraint requires new assumptions that are very unlikely to have any meaningful justification.

What About Transaction Costs?

Despite an explicit interest in household production, the recent neoclassical economics of the household so abstracts the production process that it become difficult to recognize the daily activities in which we all engage. Berk (1980, p. 136) has observed, 'One of the ironies of the New Home Economics is that with all the talk about the household production function, scant attention is paid to the actual production processes implied.' More recently, Pollak (1985, p. 582), has noted that

> Since neoclassical economics identifies firms with their technologies and assumes that firms operate efficiently and frictionlessly, it precludes any serious interest in the economizing properties and internal structure and organization of firms. The new home economics, by carrying over this narrow neoclasical view from firms to households, thus fails to exploit fully the insight of the household production approach.

Pollak goes on to propose a transactions cost approach to households in which the family is conceptualized as a governance structure rather than a preference ordering. Special emphasis is placed on how families are able to provide incentives to their members and monitor their performance. For example, because important instrumental and expressive activities are carried out in the same setting, families are able to apply rewards and punishments not readily available to other institutions. Yet at the same time, the intermingling of economic and personal relationships means that quarrels initiated in one sphere may carry over into another. Whatever the merits of Pollak's perspectives, they emphasis how much of family life has been lost in the neoclassical abstraction.

Model Specification in Empirical Work

The ultimate validation of any theory must come from how it performs in the empirical world. By and large, the empirical work done to date within the new home economics has been roughly consistent with theoretical predictions. However, the effects of key variables are often very small and/or statistically indistinguishable from zero (e.g. Layard and Mincer 1985). More important, as Pollak asserts (1985, p. 584), 'because of the central role of unobservable variables (e.g., preferences, household technology, genetic endowments), the new home economics view of the family does not lead simply or directly to a model capable of empirical implementation'.

For example, it is one thing to 'hold constant' the role of a priori preferences when extracting the essentials for theory development, but quite another to omit sound measures of tastes from one's econometric models (Berk and Berk 1983, pp. 380–1). Unless the omitted taste variables are uncorrelated with either the outcome variable or the explanatory variables that are included, based estimates will result. Hence, even when statistical results appear consistent with economic theory, it is not clear what has been demonstrated. And to date, the empirical literature has typically failed to introduce reasonable measures of family members' preferences.

Conclusions

Given the current state-of-the-art, economists probably ask far too much of their theories.

Nowhere is this more true than in the recent applications of neoclassical microeconomics to families. In the search for signed results, enormous simplifications and abstractions have been introduced. One is left with a perspective that if taken literally will probably fail.

First, the requisite assumptions, if accepted at face value, make the theory of dubious relevance for most households. Consequently, one is in practice reduced to arguing about how closely the theory approximates reality, and almost any empirical findings may be dismissed. If, for example, in certain developing countries women's labour force participation does not respond in expected ways to increases in market wages, one may simply claim that the market economy is insufficiently mature.

Second, many of the theory's key concepts are typically unobserved in practice and perhaps even unobservable in principle. This means that all empirical efforts are undermined by errors in variables and model misspecification. Once again, therefore, virtually any empirical finding may be discarded. For example, if women with more education spend fewer hours caring for their children than women with less education, it may be that one is witnessing the substitution effects (via greater market wages) predicted by economic theory. Alternatively, with greater education comes a preference for market activities. Or, women who already prefer market activities to home activities obtain more education. However, *all* of these interpretations may be easily dismissed. Neither the going, occupationally specific wage nor preferences for market activities are directly measured.

In contrast, the sensitizing role of recent efforts by neoclassical economists to understand family life has been extraordinarily useful. The new home economics force one to address seriously the nature of household production and the degree to which concepts from neoclassical economics can be instructive. In other words, we are told where to look and given some initial tools to aid in that process. These are major accomplishments.

See Also

- ▶ Family
- ▶ Gender
- ▶ Housework
- ▶ Women and Work

Bibliography

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

Ben-Porath, Y. 1982. Economics and the family – match or mismatch? *Journal of Economic Literature* 20(1), 52–64.

Berk, R.A. 1980. The new home economics: an agenda for sociological research. In *Women and household labor*, ed. S.F. Berk. Beverly Hills: Sage.

Berk, R.A., and S.F. Berk. 1979. *Labor and leisure at home*. Beverly Hills: Sage Publications.

Berk, R.A., and S.F. Berk. 1983. Supply-side sociology of the family: The challenge of the New Home Economics. In *Annual Review of Sociology*, vol. 9. Palo Alto: Annual Reviews Inc.

Blaug, M. 1980. *The methodology of economics*. Cambridge: Cambridge University Press.

Blood, R.O., and D.W. Wolf. 1960. *Husbands and wives: The dynamics of married living*. New York: Macmillan.

Hannan, M.T. 1982. Families, markets and social structure. *Journal of Economic Literature* 20(1), 65–72.

Hart, W.L. 1984. *Attorney general's task force on family violence*. Washington, DC: US Department of Justice.

Henderson, J.M., and R.E. Quandt. 1980. *Micro-economic theory: A mathematical approach*, 3rd ed. New York: McGraw Hill.

Hollis, M., and E. Nell. 1975. *Rational economic mann*. Cambridge: Cambridge University Press.

Layard, R. and Mincer, J. (guest eds.) 1985. Trends in women's work, education and family building. *Journal of Labor Economics* 3(1), i–iii.

Leibenstein, H. 1976. *Beyond economic man*. Cambridge: Harvard University Press.

Lesourne, J. 1977. *A theory of the individual for economic analysis*. New York: North Holland.

McElroy, M.B., and M.J. Horney. 1981. Nash – bargained household decisions: Toward a generalization of the theory of demand. *International Economic Review* 22(June): 333–49.

Mancer, M., and M. Brown. 1980. Marriage and household decision-making. *International Economic Review* 21, (February): 31–44.

Nerlove, M. 1974. Toward a new theory of population and economic growth. In *Economics of the family*, ed. T.W. Schultz. Chicago: University of Chicago Press.

Pollak, R.A. 1985. A transaction cost approach to families and households. *Journal of Economic Literature* 23(2), 581–608.

Pollak, R.A. and Wachter, M.L. 1975. The relevance of the household production function and its implications for the allocation of time. *Journal of Political Economy* 83(2), 255–277.

Schultz, T.W. 1974. Fertility and economic values. In *Economics of the family*, ed. T.W. Schultz. Chicago: University of Chicago Press.

Simon, H.A. 1978. Rationality as process and as product of thought. *American Economic Review* 68(2), 1–16.

Witte, A.D., Tauchen, H.V. and Long, S.K. 1984. *Violence in the family: A non-random affair. Working paper no. 89*, Department of Economics, Wellesley College.

Household Production and Public Goods

Reuben Gronau

Abstract

Home production constitutes even in modern economies about one-third of GNP. The article discusses Becker's theory of home production and its critiques. It develops a general model where welfare is a function of market and home goods, market work, work-at-home and leisure, focusing on problems of its identification arising from the fact that home output is not traded in the market. These problems are aggravated in the multi-person household framework, since intra-household allocation is unobserved. These difficulties have serious ramifications for the measurement of adult equivalent scales, productivity at home and home output.

Keywords

Barten method; Barten, A.; Becker's household production model; Children; Equivalence scales; Family; Fertility; Home goods vs. market goods; Household production and public goods; Intra-household distribution; Kuznets, S.; Leisure; Marginal productivity; Psychic income; Real business cycles; Schooling; Shadow prices; Time use; Value of time; Women's work and wages

JEL Classifications

D13

The concept of household production (or home production) is not new to economics. It is often used synonymously with 'cottage industries' – production taking place at home – and is generally associated with less-developed economies. Mincer (1962) emphasized the importance of the substitution between work at home and work in the market in developed economies for the understanding of married women's labour supply decisions. He was also the first to point out the importance of time scarcity for the analysis of fertility decisions, the demand for maids, and the choice of transport modes (Mincer 1963). It was, however, Becker's seminal paper (1965) that made the concept of household production an integral part of economic theory.

Becker's Theory of Household Production and Its Critiques

Becker introduced two novel elements into classical consumption theory. Whereas the classical consumer maximizes welfare subject to the budget constraint, and the object of welfare is the goods consumed, in Becker's analysis the object of welfare is the household's activities ('commodities', in Becker's terminology), where each activity is a combination of market goods and time inputs. The household maximizes welfare subject to two constraints – the budget constraint and the time constraint (the fact that the different time uses, at home and in the market, cannot exceed total time available). In this model the household's decisions can be divided into two stages: (a) the production stage (how to 'produce' each activity?) and (b) the consumption stage (what is the optimal activity bundle that will maximize welfare?). The 'household production' decisions are determined by the household technology and the relative factor prices, and the consumption decisions are determined by the activity 'shadow' prices and by the total resource constraint.

The new theory diverges from classical consumption theory in several important respects. Whereas in classical theory all households face the same prices, in the new framework different households place different values on their time, choose a different input mix, and consequently face different activity prices. Different consumption bundles consumed by households with identical incomes do not attest necessarily to differences in preferences, but may be traced to differences in home technology or in the implicit value of time. Specifically, when time can be moved freely from home uses to work in the market, and when work in the market does not involve any direct disutility, the implicit value of time will equal the (marginal) wage rate. Consumers who earn higher wages are expected to produce each activity using a more goods-intensive input mix – conserving on time. The more time-intensive the activity is (for example, sleep or watching television) the more expensive it becomes, and the less favourable it becomes for high-wage earners, who are expected to choose a more goods-intensive set of activities. In the Becker framework the theory of consumption is integrated with labour supply analysis.

The model of household production was instrumental in the development of demand analysis for fertility (Willis 1973; Becker and Lewis 1973), health (Grossman 1972), transport (Gronau 1970) and other applications. The popularity of the model can be traced to the insights gained by combining consumption and production theory to explain household behaviour. One of the few dissenting voices was that of Pollak and Wachter (1975). In Becker's original model the shadow prices of the activities are independent of the amount of the activities consumed. The authors point out that this assumption is satisfied only under very restrictive conditions. For this to hold, the marginal inputs of time and goods cannot vary with 'output', and the shadow price of time has to be constant. The first assumption requires that the production process be subject to constant returns to scale, and the second assumption requires that the time inputs do not generate any direct utility per se (that is, in Becker's model one enjoys the commodity 'children' but not the childcare going into their 'production'). The first assumption rules out the existence of increasing returns to scale, often mentioned as one of the economic motives for the establishment of multi-person households, while the second is at odds with the standard distinction (emphasized by Mincer) between leisure and work at home. In this formulation a meal is a meal, regardless whether one worked on it for two hours and ate it in five minutes, or worked on it for five minutes and ate it in two hours.

A Three-Way Allocation of Time: Market Work, Work at Home and Leisure

The distinction between the two types of home time was resurrected by Gronau (1977), who proposed a model where the consumer allocates his time between three time uses: leisure, work in the market, and work at home, the last of these serving as an input in the production of 'home goods'. In the most general formulation welfare is defined over the three time uses and the two types of goods (home and market goods) $U = U(X_m, X_h, T_m, T_h, L)$, where X_m denotes market goods, X_h home goods, T_m market time, T_h work at home time, and L leisure. The home production function is $X_h = F(T_h)$. The constraints confronting the person are the budget constraint $X_m = wT_m + V$, where w is the real wage rate, and V non-labour sources of income; and the time constraint $T_m + T_h + L = 1$. The first order conditions for an interior solution (that is, $T_m > 0$, $T_h > 0$) are

$$(U_L - U_{Tm})/U_{Xm} = w \quad \text{and} \quad (U_L - U_{Th})/U_{Xh} = F',$$

where F' denotes the marginal productivity of work at home. Combining the two equations, one obtains the familiar factor demand equation

$$(U_{Xh}/U_{Xm})F' = [(U_L - U_{Th})/(U_L - U_{Tm})]w,$$

stating that the value of marginal productivity of work at home equals the 'shadow' price of time at home. (U_{Xh}/U_{Xm}) denotes the 'shadow' price of home goods, and the 'shadow' price of time is corrected for the differential in direct utilities of work in the market compared with work at home $[(U_L - U_{Th})/(U_L - U_{Tm})]$.

Unfortunately, three out of the four terms in this equation are unobserved (the ‘shadow’ price of home goods, the marginal productivity of work at home, and the price of time correction factor), limiting the applicability of this equation for empirical research. Thus, changes in the observed variable, the wage rate, can be used to trace the parameters of any of the unobserved terms, but only if the parameters of the other two unobserved terms are arbitrarily restricted.

Gronau (1977) assumed that home and market goods are perfect substitutes ($U_{Xh} = U_{Xm}$), as are home and market work, yielding the dual condition for an interior optimum $(U_L - U_T)/U_X = F' = w$. The existence of two separate margins allows the tracing of the slope of the production function and the contours of the indifference curve between work time and goods. In this scheme the choice between leisure and goods is governed by preferences, and the allocation of work time between home and market is determined by technology.

Other studies tried to isolate other components of the equation, imposing a different set of restrictions. For example, Kerkhofs and Kooreman (2003), following Graham and Green (1984), estimated the psychic income from work at home by assuming perfect substitution between home and market goods and restricting the marginal productivity F' to be a linear function of work at home. Rupert et al. (1995) focused on the elasticity of substitution between home and market goods; but in order to obtain credible estimates of this parameter they had to assume that home and market work are perfect substitutes and impose specific values on the home production elasticity.

Home Production and Intra-Household Distribution

Becker’s original model strictly applies only to a one-person household. Several attempts have been made to adapt it to a multi-person environment (and specifically to the husband-wife case). The multi-person household models add to the household decisions a third dimension – the intra-household distribution. Given the difficulties encountered in separating consumption from production, adding a third set of unobservables does not contribute to the tractability of the models.

The models agree that each spouse’s leisure should appear separately in the welfare function, and that the spouses’ work at home is mutually substitutable in the home production function. There is, however, disagreement over whether home goods are private or public goods. The specific formulation of the welfare function varies depending on whether the researcher belongs to the ‘unitary’ or the ‘collective’ camp.

The empirical analysis reflects the difficulty in separating consumption (that is, the shadow price of home goods, the psychic income from work at home), household production technology (that is, the marginal productivity of work at home) and intra-household distribution effects. The most important ‘output’ of home production in most households is their children. Children (in particular when they are young) are associated universally with increased work at home and childcare. It is, however, impossible to tell how much of the increased time input is due to the increased shadow price of home goods, and how much should be attributed to the increased psychic income derived from work at home.

Similar difficulties affect the analysis of the factors affecting home productivity. A central theme in this analysis is the estimation of the returns to scale in home production or, alternatively, how important is the public-goods component of home goods (for example, home repair, house cleaning, laundry, cooking, shopping).

The analysis of returns to scale is one of the oldest chapters in empirical economics, dating back, under the heading of ‘Adult equivalence scales’, to the studies of Engel at the end of the 19th century. Equivalence scales are index numbers intended to allow comparisons of welfare (or real incomes) across households of different size and composition. The discussion of the methods of estimation of these index numbers on the basis of observed consumption patterns generated an extensive literature. The literature is unanimous in concluding that there exist substantial returns to scale in consumption. According to a survey paper by Van Praag and Warnaar (1997), discussing 76 studies, it is found that on average a two-person household ‘needs’ per person only 80 per cent of the resources ‘needed’ by a one-person household, and a three-person household ‘needs’ per person only two-thirds of a single-person household. Unfortunately, these estimates suffer from the shortcomings common to all multi-person models of household production: they cannot separate unobserved household technology from unobserved intra-household distribution rules. This difficulty is, perhaps, best demonstrated by one of the more sophisticated methods of estimation of the equivalence scales – the Barten method.

Barten (1964), recognizing that demographic changes may have different effects on different goods, allowed for goods-specific scales. In his formulation welfare is a function of the deflated quantity of goods (X_i/M_i), where the value of the goods- specific deflator M_i reflects the returns to scale in its consumption. Barten’s formulation looks very similar to that of Becker’s household production model, where time inputs are omitted, and where it is assumed that the marginal productivity of goods in the production of activities (‘commodities’) is constant and equal to $1/M_i$. If we follow this analogy, retrieving M_i should yield an estimate of the parameters of household technology.

Various suggestions have been made on how to estimate the deflators M_i by comparing the consumption patterns of households of different size. However, when total consumption is given, differences in consumption patterns between a single-person household and a multi-person household reflect both the difference in the consumption patterns of the household’s members and the resources each of them commands (if all members allocate their resources identically between all goods, the single-person household and a multi-person household will have the same consumption patterns, and the comparison will generate only ‘noise’). Hence, preferences (or technology) and distribution are inseparably entangled, and there is no way to separate returns to scale from the distribution rule (Gronau 1988).

Home Production Productivity

Productivity is positively correlated with physical capital investments, and there is unanimity that married women’s increased productivity at home, due to increased investment in home equipment, has been an important factor explaining their increase in labour force participation since the 1950s. Investments in human capital (schooling, health and on-the-job training) have been shown to increase a person’s productivity in the market. Do these investments have side benefits at home? Michael (1973), who studied the consumption patterns of households with different schooling, concluded that schooling significantly increases productivity in the use of goods in home production. Gronau (1973) studied the impact of schooling on the productivity of time use. He focused on the schooling effect on married women’s reservation wage, where the unobserved reservation wage is imputed from their labour-force participation decisions. He found that college education raises the value women place on their time at home by 20 per cent compared with high-school graduates (about half the effect schooling has on their productivity in the market). Finally, Gronau and Hamermesh (2001) argued that schooling makes people more productive at home by allowing them to squeeze more ‘leisure activities’ into a smaller amount of ‘free’ time.

Home Production and the Macroeconomy

Inspired by Becker’s original analysis, the orientation of most of the studies of household production was microeconomic, emphasizing the behaviour of individual households. This orientation changed following Becker’s 1987 AEA presidential address (1988), demonstrating the implications of family economics and household production for growth and the macroeconomy. The challenge was met by Benhabib et al. (1991)

and Greenwood and Hercowitz (1991). The two teams tried to explain some irregularities in the traditional model of the real business cycle (RBC) in terms of shifts from the market economy to home production. Benhabib, Rogerson and Wright tried to obtain a better explanation for the fluctuation in labour inputs over the business cycle, whereas Greenwood and Hercowitz focused on capital formation. Both topics have become the theme of several sequels, while other authors use home production to explain a wide range of additional macro phenomena: endogenous growth, development, fiscal policy and the business cycle, and the welfare cost of inflation. The common technique employed in the new generation of models is calibration, the models sharing the common assumption that market and home goods are close, though not perfect, substitutes. (An early survey of the topic and the literature is contained in Cooley 1995, chs. 1, 5, and 6.)

The Measurement of Household Output

While the new breed of macroeconomic studies is purely theoretical, the emphasis of an older macroeconomic branch, closely related to the national income accounting family, is purely on measurement. The exclusion of the output of the home sector has long been recognized as a major omission in national accounting (Kuznets 1944), an omission that can seriously bias international comparisons of standards of living and estimates of growth rates. Several attempts have been made to correct this lapse.

The value of output in the home sector, as that of other non-market sectors (such as the government and non-profit organizations) is measured in terms of the value of inputs used in the production process. There is, however, an inherent difference between the home sector and the other non-market sectors, namely, that the time inputs used in home production do not carry an explicit price tag. Two methods have been suggested to circumvent this difficulty: (*a*) the market opportunity cost method, and (*b*) the market alternative method. According to the first method, the time inputs are evaluated according to the price they can command in the market. The second approach tries to evaluate home services at their market prices. Both methods are vulnerable to serious conceptual objections.

The objection to the ‘opportunity cost’ method stems from the fact that the same service (say, childcare) is evaluated at different prices if the provider gets a different wage in the market. The objection to the ‘market alternative’ method is that the household could have bought the home services at these prices but has rejected this option. These difficulties can be traced again to the inherent problem of identification of the work-at-home demand equation.

If one could assume that work at home yields no psychic income, then the ‘opportunity cost’ method should be employed, since in this case differences in the market wage attest to differences in the evaluation of home goods (for example, because of differences in the conceived quality of service). On the other hand, if women with different market wages perform the same service at home merely because of differences in psychic income, then the ‘market alternative’ approach should be preferred.

Even if the conceptual difficulties could be resolved, some technical difficulties remain. The ‘opportunity cost’ method has to cope with the problem that a substantial fraction of home output is produced by ‘full-time’ workers in the home production ‘industry’ (that is, house-persons) who receive no market wage (Gronau 1973). Moreover, these workers should be regarded as self-employed, and the evaluation of their output should incorporate the returns to their entrepreneurial capacity (Gronau 1980). The ‘market alternative’ approach advocates are undecided whether to use the cost of a maid as the market alternative or whether each home service should be priced separately.

Given the often-heated debate between the proponents of the two methods, the imputation outcomes show a surprising degree of similarity. Hawrylyshyn (1976), who compared nine international studies of both types, found that the average estimate of the value of home production is 35 per cent of GNP, with the estimates ranging from 32 to 39 per cent.

See Also

- ▶ Becker, Gary S. (Born 1930)
- ▶ Childcare
- ▶ Collective Models of the Household
- ▶ Equivalence Scales
- ▶ Intrahousehold Welfare
- ▶ Kuznets, Simon (1901–1985)
- ▶ Leisure
- ▶ Real Business Cycles
- ▶ Time Use
- ▶ Value of Time

Bibliography

Barten, A. 1964. Family composition, prices and expenditure patterns. In *Econometric analysis for national economic planning*, ed. P. Hart, G. Mills, and J. Whitaker. London: Butterworth.

Becker, G. 1965. A theory of the allocation of time. *Economic Journal* 75: 493–517.

Becker, G. 1988. Family economics and macro behavior. *American Economic Review* 78: 1–13.

Becker, G., and H. Lewis. 1973. On the interaction between quantity and quality of children. *Journal of Political Economy* 81: S279–S288.

Benhabib, J., R. Rogerson, and R. Wright. 1991. Homework in macroeconomics: Household production and aggregate fluctuations. *Journal of Political Economy* 99: 1166–1187.

Cooley, T. (ed.). 1995. *Frontiers of business cycle research*. Princeton: Princeton University Press.

Graham, J., and C. Green. 1984. Estimating the parameters of the household production function with joint products. *The Review of Economics and Statistics* 66: 277–282.

Greenwood, J., and Z. Hercowitz. 1991. The allocation of capital and time over the business cycle. *Journal of Political Economy* 99: 1188–1214.

Gronau, R. 1970. The effect of traveling time on the demand for passenger transportation. *Journal of Political Economy* 78: 377–394.

Gronau, R. 1973. The effect of children on the housewife's value of time. *Journal of Political Economy* 81(Part II): S168–S199.

Gronau, R. 1977. Leisure, home production, and work-the theory of the allocation of time revisited. *Journal of Political Economy* 85: 1099–1123.

Gronau, R. 1980. Home production: A forgotten industry. *The Review of Economics and Statistics* 62: 408–416.

Gronau, R. 1988. Consumption technology and the intrafamily distribution of resources: Adult equivalence scales reexamined. *Journal of Political Economy* 96: 1183–1205.

Gronau, R., and D. Hamermesh. 2001. *The demand for variety: A household production perspective*, Working paper, vol. 8509. Cambridge, MA: NBER.

Grossman, M. 1972. *The demand for health: A theoretical and empirical investigation*, Occasional paper, vol. 119. New York: NBER.

Hawrylyshyn, O. 1976. The value of household services: A survey of empirical estimates. *Review of Income and Wealth* 22: 101–131.

Kerkhofs, M., and P. Kooreman. 2003. Identification and estimation of a class of household production models. *Journal of Applied Econometrics* 18: 337–369.

Kuznets, S. 1944. *National income and its composition*, vol. 2. New York: NBER.

Michael, R. 1973. Education in non-market production. *Journal of Political Economy* 81(Part I): 306–327.

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

Mincer, J. 1963. Market prices, opportunity costs, and income effects. In *Measurement in economics*, ed. C. Christ. Stanford: Stanford University Press.

Pollak, R., and M. Wachter. 1975. The relevance of the household production function and its implications for the allocation of time. *Journal of Political Economy* 83: 255–278.

Rupert, P., R. Rogerson, and R. Wright. 1995. Estimating substitution elasticities in household production models. *Economic Theory* 6: 179–193.

Van Praag, B., and M. Warnaar. 1997. The cost of children and the use of demographic variables in consumer demand. In *Handbook of population and family economics*, vol. 1A, ed. M. Rosenzweig and O. Stark. Amsterdam: North-Holland.

Willis, R. 1973. A new approach to the theory of fertility behavior. *Journal of Political Economy* 81: S14–S64.

Household Surveys

Duncan Thomas

Abstract

Household surveys play a pivotal role in empirical economics. Cross-section and longitudinal surveys are regularly conducted worldwide. A description of survey design and sampling methods provides the foundation for discussing survey errors. These include errors associated with sampling, survey coverage and non-response (which includes attrition from panel surveys), and errors of observation or

measurement. In recent years, surveys have tended to become more complex and broader in scope with many reaching beyond measuring economic choices, constraints and outcomes. This trend will likely continue and exciting technological innovations in survey methods and implementation promise to revolutionize the field.

Keywords

Bootstrap; Clustering; Cohort survey; Consumer expenditure; Coverage error; Cross-section surveys; Demographic surveys; Engel's Law; Fertility surveys; Health surveys; Household production; Household surveys; Human capital; Jackknife; Longitudinal (panel) surveys; Non-observational and observational errors; Probability sampling; Sampling error; Synthetic panels

JEL Classifications

O1

Household surveys provide one of the pillars upon which some of the most important innovations in economics during the last half of the 20th century have been built. Enumeration of households dates back at least to the collection of budget data in the late 18th century. Eden (1797) compiled information on the diet, dress, fuel, and habitation spending as well as earnings of households from 86 households in England, while Davies (1795) reported detailed budgets of 127 households engaged in agriculture. Both studies sought to describe the lot of the poorest in England and so the budgets are not representative of the English population at the time. Ducpétiaux (1855) published the budgets of 199 Belgian households. Those data provided the empirical foundation for Engel's Law (Engel 1857) which posits an inverse relationship between income and the share of the budget spent on food.

Statistical Foundations

The development of practical methods of probability sampling and a theory to support estimation and inference based on those samples had a major impact on the design and implementation of household surveys. Work by Neyman (1934) on stratified designs and work on randomization in agricultural experiments by Fisher (1935) were especially influential, and their work, in combination with contributions by inter alia Bowley (1926), Deming (1950), Kaier (1895) and Yates (1935) provided a theoretical foundation for survey design.

The importance of scientific surveys was underscored by some spectacular failures. For example, in 1936 the *Literary Digest* mailed out ten million questionnaires in a poll about the election of the next US president. About two million respondents mailed back their questionnaires, and the *Digest* predicted a victory for the Republican candidate, Alfred Landon. The election was won by a landslide – not by Landon but by his opponent, Franklin Roosevelt. There were also very influential survey successes. For example, Mahalanobis (1940) highlighted the advantages of surveys in terms of cost and timeliness of results. Using a sample survey of jute producers in Bengal, he estimated the area under jute within three per cent of the official estimate based on a complete census. The cost of the sample survey was only about eight per cent of the cost of the census. His sample survey cost eight per cent of the census.

These advances laid the foundation for an explosion in the quantity and quality of household surveys during the second half of the 20th century. Many of the surveys have been designed and implemented by national statistical agencies. At a substantive level, there are at least three important classes of household surveys, each of which has specific goals.

First, household budget surveys collect detailed information on the spending patterns of households. They are used to calculate price indices and poverty lines and to estimate the incidence of poverty. These include the Indian National Sample Survey, the Family Expenditure Survey (FES) in the UK and the Consumer Expenditure Survey (CEX) in the United States. Nowadays, virtually every country in the world conducts household budget surveys periodically. In some cases, respondents are asked to maintain a diary of spending over a pre-specified time period. In other surveys respondents are interviewed and asked to recall spending on items, often with varying recall periods depending on the item. The diary method typically covers a relatively short time period, which complicates modelling low frequency purchases and interpreting reported spending as indicative of longer-run resource availability. The interview method is potentially affected by recall error. This includes forgetting (which increases with the recall period) and telescoping, which may be positive (if spending before the recall time frame is telescoped into the recall period) or negative (if spending during the recall time frame is telescoped out of the period). Whether the interview or diary method yields less measurement error remains an open question.

Second, labour force and income surveys are collected routinely to monitor inter alia labour force participation, unemployment and earnings. Labour force surveys tend to be administered frequently and samples are large enough to detect small changes in the labour market. In the United States, for example, the Current Population Survey (CPS) is a monthly survey of over 50,000 households that has been conducted for over 50 years. Some surveys focus on income and wealth. The Survey of Consumer Finances measures the financial health of the US population and includes a special over-sample of the most wealthy households.

> The third class of surveys measure non-economic domains of well-being. Fertility surveys provide information on marriage and living arrangements, reproductive health including pregnancies and births, and use of health services. These are important for documenting the dramatic changes in family formation, composition and size that has occurred over the 20th century. Health surveys monitor the health of the population. In some cases, such as the National Health and Nutrition Examination Survey (NHANES), an extensive physical examination is performed by trained medical personnel in conjunction with a detailed questionnaire about health status and health-related behaviours. Several surveys integrate demographic with health information including the Demographic and Health Surveys (DHS), which grew out of the World Fertility Surveys and have been collected in over 75 countries. Surveys of attitudes, like the General Social Survey, are routinely collected across the globe.

In practice, the distinction between these classes of surveys is not clear-cut since many of the economic surveys record demographic, health or attitudinal information, and vice versa. To be sure, these topic-specific surveys are extremely important for monitoring the prevalence of indicators of interest to researchers and policymakers. However, the surveys are often inadequate for testing hypotheses about behaviours of individuals and their families.

In the late 1960s, surveys were designed to address this limitation, explicitly drawing on the theoretical models of household behaviour suggested by Gary Becker, T.W. Schultz, and their collaborators and students. One class of surveys explicitly recognized the dual role of households in agricultural economies as both producers and consumers of food. See, for example, Evenson (1978) for a discussion of a series of innovative household surveys conducted by nutritionists and economists in Laguna Province, Philippines. These surveys collect detailed information on farm inputs and output, non-farm activities, consumption, health and demographic behaviour.

Another class of surveys relied on the economic model of household production to guide the collection of information on individual choices and constraints people face. For example, the RAND Malaysian Family Life Survey (MFLS) was designed to capture multiple domains of the lives of each individual respondent, their family and community to better understand the determinants of fertility and investment in children during early life (Butz and DaVanzo 1975). As a result of the scope of the questions, MFLS has been used to address a far broader array of questions in economics and demography than those for which it was originally conceived. The International Crops Research Institute for Semi-Arid Tropics (ICRISAT) village-level studies (VLS) followed a similar approach. The best known of these was conducted in six villages in three regions of semi-arid India and collected very detailed data on a very broad array of topics from 240 farm households surveyed annually for ten years (Walker and Ryan 1990).

The Living Standard Measurement Surveys (LSMS) conducted by the World Bank drew heavily on the experiences of the Laguna, MFLS and ICRISAT studies among others. Conceived as broad-purpose surveys to monitor poverty and material well-being in developing countries and also contribute to the design of social policy, the surveys collect a wide array of indicators of well-being and behaviours of households along with extensive community data. Initiated in the mid-1980s, a hallmark of the LSMS program is a framework that is broadly consistent across many countries. Having been implemented in many low-income and transition countries around the world, LSMS and DHS stand out as leaders in the development of comparable survey data collected from a wide spectrum of social and economic contexts.

Survey Design

A typical household survey selects a sample of households from a frame which is the population of interest for the research. In many cases, the frame is a census and the sample is representative of a geographic area, although this need not be the case. The simplest sampling strategy randomly selects households from the frame. In practice, most household surveys follow a two-stage (or multi-stage) sampling design in which clusters are selected and then households are selected from those clusters.

There are several advantages associated with geographically-defined clusters. Administration costs are lower for surveys that involve face-to-face interviews. Clusters may facilitate incorporating neighbourhood-or community-level data in the survey or, alternatively, models might highlight variation within communities and control community-level heterogeneity with a fixed effect, for example.

Clustering also carries disadvantages since two sampled units within a cluster tend to be more similar than two randomly selected units. The loss of independence across sampled units results in lower precision and thus larger standard errors of estimates. The magnitude of this effect for a particular indicator is often summarized by the design effect which is the ratio of the variance, with the cluster design taken into account, to the variance if households were randomly selected. An alternative summary statistic is provided by the intra-cluster correlation coefficient. The greater the covariance within clusters relative to differences across clusters, and the larger the number of households within a cluster, the greater is the design effect and the greater is the loss of precision due to clustering. It is standard practice to estimate standard errors by taking account of the clustering following the method of Huber (1967) or a re-sampling approach such as the jackknife or bootstrap (Efron 1982). In short, clustering buys more information per unit cost but less information per sampled unit.

Many surveys are designed to oversample specific sub-populations, in which case estimates are typically adjusted for the probability of a household being selected into the sample. An important principle underlying population-based sampling is that because the probability of selection of every eligible unit is known and greater than zero, with appropriate weights, it is possible to reconstruct the population, although in some instances the complexity of survey designs becomes overwhelming.

Survey Errors

There are at least two classes of error in any survey. ‘Non-observational’ errors occur when part of the target population is not measured. ‘Observational’ errors are the result of incorrect measurement.

Sampling error, the most familiar survey error, is a form of non-observational error. It reflects the fact that any sample is a subset of the underlying population and so an estimate based on the sample will not be identical to the population value.

Coverage error, another source of non-observational error, arises when the sampling frame excludes part of the target population. Many sample frames are based on a list of household dwellings; those samples exclude homeless people and so are not representative of the entire population. If a household listing is based on an old census, more mobile people are at risk of being under-represented. A sampling frame based on telephone numbers (or e-mail addresses) will exclude those who do not have a telephone (or e-mail address) and oversample those with multiple numbers (or addresses).

A third source of non-observation error arises from non-response, of which there are two categories. First, survey non-response occurs when a target respondent cannot be located. It will also arise if the respondent refuses to participate in the survey (or fails to answer the telephone, respond to an e-mail or return a mailed-out survey). Second, item non-response occurs when a respondent fails to answer one or more questions in the survey either because he or she refuses to answer or does not know the answer to the question(s). The incidence of the latter is reduced by probing, and unfolding brackets have proved to be particularly useful for economic quantities (Hurd et al. 1998).

Broadly speaking, non-response rates tend to rise with the value of time of the respondent, and there has been a secular trend of increased non-response in many developed countries. Survey non-response in developing country household surveys is typically substantially lower than in higher income countries.

If, conditional on observed characteristics, coverage and non-response error are random, appropriate weights can be computed so that survey statistics are representative of the underlying population. Complications arise when these errors are selected on unobserved characteristics. Several procedures have been suggested to deal with non-response error including hot deck or matching procedures (Rosenbaum and Rubin 1983) and modelling the selection process with a control function (Heckman 1978).

The most familiar source of observational error is respondent failure to answer a question correctly. This may be intentional (in order to misrepresent reality) or unintentional. Interviewers may make errors in the administration of the survey, and there may be interviewer-specific effects in the ways questions are asked. Survey instruments are also prone to error. In general, the extent of observational error likely depends on interactions among the sources of error and also on the mode of the survey. Respondents in telephone surveys tend to provide shorter answers than those in face-to-face interviews, and web-based surveys are more likely to be ended prematurely.

While the distinction between observational and non-observational error is conceptually useful, in practice the distinction is often blurred. For example, survey non-response is typically related to interviewer characteristics. Both item non-response and respondent error have been shown to be related to questionnaire design and interviewer characteristics. Groves (1989) provides an excellent discussion of these and related issues.

Typology of Surveys

Cross-section surveys provide a snap-shot of a target population at a point in time. They are the bread and butter of research based on household surveys. Many cross-section surveys are repeated regularly, with independent samples drawn from the same target population, so that it is possible to track the evolution over time of indicators such as unemployment, poverty or inequality as well as map changes for population sub-groups. Synthetic panels of individuals created using repeated cross-section follow the same population subgroup over time, such as a birth cohort. They are straightforward to interpret if there are no entrants into or exits from the target population via, for example, immigration, emigration or death. Synthetic panels of households are more complicated. Household composition changes due, for example, to marriage or divorce result in changes over time in the unit being followed. It is difficult to distinguish composition changes from true change. Similar issues arise with synthetic panels of communities.

Longitudinal or panel surveys follow the same respondent over time, which provides opportunities for exploration not feasible with cross-section surveys. First, tracing the dynamic evolution of choices and outcomes over the individual's life provides insights into, for example, early life experiences and later life outcomes, resilience and recovery from adversity as well as the characteristics of those who cycle in and out of some state (such as poverty, unemployment, public assistance or poor health).

Second, panel data provide expanded options for treating unobserved heterogeneity in models like

$$y_{it} = x_{it}\beta + \mu_i + \varepsilon_{it}$$

where μ_i is an unobserved individual-specific characteristic. If μ_i is correlated with x_{it}, OLS estimates of β are biased. With repeated observations on the same individual in a panel, μ_i can be estimated (or the model cast in first-differences) to consistently estimate β. The 'fixed effect' μ_i absorbs all time-invariant individual characteristics that enter the model in a linear and additive way.

The advantage of a longitudinal survey is that the same sampling unit is followed over time. This is also its Achilles heel. Attrition from longitudinal surveys is a particular form of non-response error. The nature and magnitude of attrition varies with the study design. For example, in face-to-face interviews in the home, individuals who move are followed to their new location and interviewed there. Those who move the furthest are often the hardest and most expensive to find. Attrition tends to be selected on traits associated with migration – younger, better-educated adults being the most likely to move. The selectivity of the sample is exacerbated in panel surveys that do not follow people who leave the location in which they were interviewed at baseline. Attrition in telephone and web-based surveys have less to do with tracking people to new geographical locations and more to do with retaining the cooperation of respondents – an issue that also confronts face-to-face interviews. In multi-wave panel surveys, it is important to attempt to re-contact respondents who have been skipped in prior waves so that attrition does not cumulate. There are many examples of well-designed panel surveys that have kept attrition low across multiple rounds.

Statistical adjustments for attrition are the same as other forms of non-response error. Re-weighting will be effective when attrition is selected on observed characteristics. When selection is on unobserved characteristics, a control function approach is more likely to be successful. In analytical models, the importance of adjusting for attrition will vary with the research question. The stronger the association is between attrition and observed or unobserved characteristics in the model, the more important the adjustment is for attrition.

An alternative approach to treating attrition is to replace a respondent who attrits from the survey with a new, similar respondent – frequently people living in the same housing structure as the respondents in the previous wave (or the person who is assigned the telephone number, e-mail address, and so forth). There are several problems with this approach. First, it assures the study population appears stable since no primary sampling units will lose population; the reality may be quite different. Second, housing structures can change, be torn down or difficult to relocate, resulting in a different type of attrition. Third, even if populations are stable in aggregate and housing structures do not change, it is assumed that the replacement and original respondents are 'exchangeable' or effectively identical. It is not clear that this will be true as in the case of a respondent who died. Fourth, the key advantage of a longitudinal survey – following the same person through the life course – is lost.

It follows that a panel survey of households has little conceptual appeal. Although a household survey is often the baseline for a panel of individuals, households change over time and it is individuals who will be followed – possibly all the original household members. These respondents will often be interviewed along with the people in their new household and so the panel is a series of household surveys embedded in which is a longitudinal survey of individuals. A small number of longitudinal surveys have sought to follow family members over time.

The Panel Survey of Income Dynamics (PSID) is a long-running panel and one of the most widely used surveys in economics. Initiated in 1968, with a nationally representative sample of 5,000 households, interviews spanning 40 years with household members, and children born to them, has provided unique insights into the dynamics of income, human capital, health and living arrangements over the life course, across cohorts and across generations (Duncan et al. 2004).

> A cohort survey is a special type of longitudinal survey which follows a specific cohort of respondents, often a birth cohort. The advantages of the design are that, because of shared environments, cohort members are less heterogeneous than the entire population and there are power benefits to comparing people making similar life course transitions at the same time. A disadvantage is that age and period effects cannot be disentangled. To address this, cohort studies often draw new cohorts. The British Cohort Studies, for example, have mounted four large-scale population-representative birth cohort studies since the 1930s. The Health and Retirement Survey (HRS) is an innovative cohort study that focuses on the health and economic well-being of older Americans. The HRS has been replicated in several countries across the globe.

Statistical innovation presaged the explosion in household surveys since the 1950s. Technological innovation is likely to provide the foundation for the next revolution in survey design. For example, electronic communication devices, geographical information systems and innovations in health measurement along with sophisticated analytical tools have already begun to profoundly affect the scope and quality of household surveys.

Bibliography

Bowley, A.L. 1926. Measurement of precision attained in sampling. *Bulletin de l'Institut International de Statistique* 22(Suppl. to Livre 1): 6–62.

Butz, W.P., and J. DaVanzo. 1975. *The Malaysian family life survey: Summary report*. Santa Monica: The Rand Corporation.

Davies, D. 1795. *The state of labourers in husbandry stated and considered*. Bath.

Deming, W.E. 1950. *Some theory of sampling*. New York: Wiley.

Ducpétiaux, E. 1855. *Budgets économiques des classes ouvrières en Belgique*. Bruxelles.

Duncan, G., S. Hofferth, and F. Stafford. 2004. Evolution and change in family in come and wealth: the Panel Survey of Income Dynamics 1968–2000 and beyond. In *A telescope on society: Survey research and social science at the University of Michigan and beyond*, ed. J. House, T. Juster, and R. Kahn. Ann Arbor: University of Michigan Press.

Eden, F.M. 1797. *The state of the poor*. London: J. Davis.

Efron, B. 1982. *The jackknife, the bootstrap, and other resampling plans*. Philadelphia: SIAM.

Engel, E. 1857. Die Productions-und Consumptions verhältnisse des Königreichs Sachsen. *Zeitschrift des Statistischen Büreaus des Königlich Sächsischen Ministerium des Innern 8(9), 22 November. Repr. in Bulletin de l'Institut International de la Statistique* 9(1985): 1–54.

Evenson, R.E. 1978. Time allocation in rural Philippine households. *American Journal of Agricultural Economics* 60: 322–330.

Fisher, R.A. 1935. *The design of experiments*. London: Oliver & Boyd.

Groves, R.M. 1989. *Survey errors and survey costs*. New York: Wiley.

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

Huber, P.J. 1967. The behavior of maximum likelihood estimates under non-standard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 4: 221–233.

Hurd, M.D., D. McFadden, H. Chand, L. Gan, A. Merrill, and M. Roberts. 1998. Consumption and saving balances of the elderly: experimental evidence on survey response bias. In *Frontiers in the economics of aging*, ed. D. Wise. Chicago: University of Chicago Press.

Kaier, A.N. 1895. Observations et expériences concernant des dénombrements représentatifs. *Bulletin of the International Statistical Institute* 9(Livre 2): 176–183.

Mahalanobis, P.C. 1940. Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society A* 109, 329–378.

Neyman, J. 1934. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society* 97: 558–625.

Rosenbaum, P.R., and D.B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70: 41–55.

Walker, T.S., and J.G. Ryan. 1990. *Village and household economies in India's semi-arid tropics*. Baltimore, MD: Johns Hopkins University Press.

Yates, F. 1935. Complex experiments. *Journal of the Royal Statistical Society* Suppl. 2, 181–247.

Housework

Clair Brown and Amelia Preece

Housework consists of childrearing and the satisfaction of basic human needs through the provision of meals, clothing and shelter within the home. The functioning of the home economy ensures reproduction while it maintains adults so that they can engage in paid work outside the home. Housework is an essential part of an economic and social system because it not only provides essential services but also helps to maintain the unequal class structure. Within the home, children learn their place in the social structure while they prepare for their place in the labour market.

Social rules and customs (i.e. institutions) govern housework – required amounts, the way it is done and who does it. Housework will therefore vary across culture, over time, and by class within each culture. Within a class, however, what housework is done and who does it will be socially determined. The gender division of labour that has occurred in most industrial societies, with men engaged in production in return for wages and women engaged in family reproduction in return for sharing income, relegated women to the private sphere of the home. Although all adults engage in some housework, wives are primarily responsible for housework and engage in 30–50 h of housework weekly in the US, depending primarily on the number and ages of their children (Walker and Woods 1976). Since women have primary responsibility for the home, they can engage in paid labour only after making sure that the socially required housework activities are done.

With very limited access to money-producing activities and with their services primarily rendered to their children and husband, women do not acquire the power that is associated with exchange in public life (Friedl 1975). Production of status through housework rather than commodities in the marketplace ensured women's inferior position (Benston 1969; Papanek 1979).

This institutional analysis of housework assumes that people's needs and desires are formed by the social structure. Alternatively, neoclassical models assume that people's preferences are idiosyncratic and that the marketplace responds to people's desires. In a neoclassical world, housework is abstracted from the social structure, and the household is analysed as a small firm that produces commodities with time and market inputs (Becker 1965). Systematic substitution between time and market goods occurs by choosing among different consumption bundles and by varying the production process. In this model, women's work decisions are made on efficiency grounds, so that the wife equalizes the marginal return on unpaid (i.e. homework) and paid (i.e. market) work. Specifically, the wife is viewed as having flexibility in deciding how to combine her time with market goods in producing family meals, a pleasant home, presentable clothing and well-behaved children. However, empirical studies of the home-making process and family budgets in the US have shown that very little substitution occurs between the homemaker's time and market goods in housework. Empirically, after standardizing by family income, the employed wife uses few market goods and services outside of childcare to substitute for her own time, and both employed and full-time home-markers use the same techniques in performing housework (Brown 1979; Strober and Weinberg 1980; Berk and Berk 1979). The main substitution tends to be between the wife's market time and her leisure time.

The lack of substitution between housework and purchased goods and services reflects the social norms governing activities that provide family life. In addition, the services provided by the home-maker and the goods and services purchased in the marketplace are generally not comparable. The home economy specializes in producing mothering and the nurturing of family members, along with personalized care in providing food, clothing and shelter. The marketplace produces sophisticated medical care, advanced education, the means of transportation and communication, urban housing and the ability to pool risks through insurance, as well as mass-produced food, clothing, cars and other consumer durables. The family's evaluation of these dissimilar homeproduced and market-produced goods and services will be a major determinant of whether the wife works exclusively in the home or also has a job. The family's evaluation will vary with social position and experiences over time.

Although a great deal of attention has been paid to estimating the market value of housework, the lack of comparability between housework and market substitutes makes these estimates problematic. The full-time homemaker's provision of round-the-clock care of family members' needs makes it impossible to equate the value of her time with her permanent replacement cost (i.e. the wage rate such services would command in the marketplace). The personalized and on-call nature of her work prevents us from evaluating the services of the housewife as a combination of so many hours of chauffeur, cook, baby-sitter and laundress per day. In the real world, the household cannot contract to buy these services – more impersonalized than the housewife's – in the small amounts of time and at the random hours that the housewife actually performs these duties. The purchased services usually are not equivalent to the service which the housewife provides because she knows intimately the family members she is serving and takes responsibility for the organizing and providing of care as it is needed. Even in societies where a servant class provides cheap domestic labour, the servant must be directed and supervised by someone, usually the home-maker, and this affects the experience of family life. Housework has evolved historically as the economy has developed and as social needs have changed (Reid 1934; Gilman 1910). Two distinct stages characterize the interaction between the home and the industrialized market. Early industrialization began the process of transferring some production processes (e.g. cloth-making, sewing, ready-made crackers) from the home to the marketplace. Although the home economy could still produce these goods, the processes were arduous and the market economy was more efficient. The more important second stage was evident in the early part of the twentieth century as the marketplace began producing goods and services that had never been produced by the home economy, and the home economy was unable to produce them (e.g. electricity and electrical appliances, the automobile, telephone, television, advanced education, sophisticated medical care). In the second stage, the question of whether the home economy was less efficient in producing these new goods and services was irrelevant; if the family were to enjoy these fruits of industrialization, they would have to be procured in the marketplace. The traditional ways of taking care of these needs in the home, such as nursing the sick, became socially unacceptable (and, in most serious cases, probably less successful). Just as the advent of the automobile made the use of the horse-drawn carriage illegal and then impractical, and the advent of television changed the radio from a major source of news and entertainment to background music, so most fruits of economic growth did not increase the flexibility for the home economy in producing these goods and services in modern capitalist economies. Growth brought with it new requirements, such as more mobility in urban areas and increased diversity in consumption goods, along with increased consumer reliance on the marketplace. In order to consume these goods and services, the family had to enter the marketplace as wage-earners and consumers.

Meanwhile, the primary housework activities of meal preparation and clothing care used a declining share of the family's budget. A housewife's efforts to decrease expenditures in these areas by direct work activities (e.g. baking from scratch) or more careful shopping had less impact on the family's budget. Purchases of food and clothing – 56% of the average wage-earner family's disposable income in 1918 and 40% in 1950 – accounted for only 26% in 1972 (Brown 1986). Thirty per cent of the family's budget that had previously been spent on food and clothing now became freed for other kinds of expenditures, primarily transportation, insurance and retirement, and home ownership. Although the output of the home economy has declined in relative importance as a determinant of the family's total consumption standard, privatized housework is still an essential part of the prevailing social and economic structure.

The demand by the women's movement for economic independence and the equalization of sex roles has brought into clear relief the contrast and contradictions between housework and paid market work. The differences between the two economies have helped perpetuate sexual inequality. Because women have been prepared to run the home economy when they assume their roles as wives and mothers, their sense of identity and personal power is grounded in this economy. The market economy and home economy have their own value structures, work structures and reward structures, which can be contrasted by five major characteristics:

1. Supervision. The housewife is her own supervisor, while most workers have a formal supervisor, who decides what work needs to be done in what manner.
2. Pay. No systematic relationship exists between the output of the housewife's effort and the family income, while a paid worker has a rate of pay for a job performed, with rules governing behaviour on the job, sick leave, vacation days and hours of work.
3. Mobility. 'Changing jobs' for the housewife usually means continuing to work (i.e. caring for the children) without a guarantee of pay, while the worker can usually find another suitable job and has unemployment insurance during job search.
4. Measure of value. Housework is socially required and does not carry a money value since the market has not provided a permanent replacement, while workers have the exchange value (i.e. wage) as the measure of value for work performed.
5. Personal behaviour. The home economy is based on the concept of mutual aid and service to others, with cooperative rather than competitive behaviour, while the competitive market economy rewards the individual.

Since money and individual advancement are not part of the reward structure of the home economy, a woman who takes the cooperative and service values of the home economy with her into the market economy will be at a disadvantage in demanding equitable compensation for her work according the values of the market economy.

Besides these conflicts between home and market economies, conflict over production and redistribution issues occurs between family members within the household and between the household and public bodies, such as the state and the workplace. The division of labour by gender that creates the basis of the conflict within the home, especially around housework, also creates interdependence among family members (Hartmann 1981).

How one views the structure of housework and its role in the economy and society determines how one evaluates female unemployment and the income distribution. In a neoclassical world, wives' unemployment results in only a small loss for both the individual and society, because the wife can substitute her own time in place of the market goods and services purchased with her earnings. Her family's income falls only to the extent that the market is more efficient in providing these goods and services. From an institutional perspective, the wife's unemployment results in an economic loss for both the individual and society that approximates her pay cheque because she cannot use her time to produce the market goods and services purchased with her pay cheque without a major change in the family's life-style. The decline in money income determines how short of expected social standards the family falls. Since income and housework time are both required and are not interchangeable, they cannot be aggregated into a single measure of 'full income' as an indicator of economic well-being. Provided that the family's required housework is being done, measured income determines a family economic well-being.

The neoclassical and institutional models of housework also suggest different strategies for women to use in their struggle for equality in the workplace. Neoclassicists can ignore the burden imposed on employed women by their housework time, since it assumes that this time can be bought off as desired. Institutionalists recognize that for equality to prevail in the labour market, either both spouses must share equally the housework hours required to sustain family life, or childcare and meal preparation must be transferred outside the home to the community or to the marketplace. However, industrializing housework within a private market economy will not necessarily fulfil the basic human needs now served by personalized housework. Such changes in housework will require a fundamental social restructing that will radically alter family life as the norms governing everyday life are transformed.

See Also

- ▶ Gender
- ▶ Household Production

Bibliography

Becker, G.S. 1965. A theory of the allocation of time. *Economic Journal* 75, September: 493–517.

Benston, M. 1969. The political economy of women's liberation. *Monthly Review* 21(4), September: 13–27.

Berk, R.A., and S.F. Berk. 1979. *Labor and leisure at home: Content and organization of the household day*. Beverly Hills: Sage.

Brown, (Vickery) C. 1979. Women's economic contribution to the family. In *The subtle revolution: women at work*, ed. R. Smith, Washington, DC: Urban Institute.

Brown, C. 1986. *Consumption norms, work roles and economic growth in urban America, 1918–1980*. Washington, DC: Brookings Institution.

Friedl, E. 1975. *Women and men: An anthropologist's view*. New York: Holt, Rinehart & Winston.

Gilman, C.P. 1910. *The home: Its work and influence*. New York: Chariton Co.

Hartmann, H.I. 1981. The family as the locus of gender, class and political struggle: The example of housework. *Signs* 6(3), Spring: 366–394.

Papanek, H. 1979. Family status production: the 'work' and 'nonwork' of women. *Signs* 4(4), Summer: 775–781.

Reid, M.G. 1934. *Economics of household production*. New York: Wiley.

Strober, M.H. and Weinberg, C.B. 1980. Strategies used by working and nonworking wives to reduce time pressures. *Journal of Consumer Research* 6 March: 338–348.

Walker, K.E., and M.E. Woods. 1976. *Time use: A measure of household production of family goods and services*. Washington, DC: American Home Economics Association.

Housing and the Business Cycle

Morris A. Davis

Abstract

Recent events have led to a renewed effort to understand the nature of cyclical fluctuations in the price and quantity of new investment in housing. This paper provides a brief summary of the existing literature modelling housing and the business cycle.

Keywords

Cobb-Douglas preferences; Expenditure shares; Housing expenditures; Multi-city models

JEL Classifications

D61

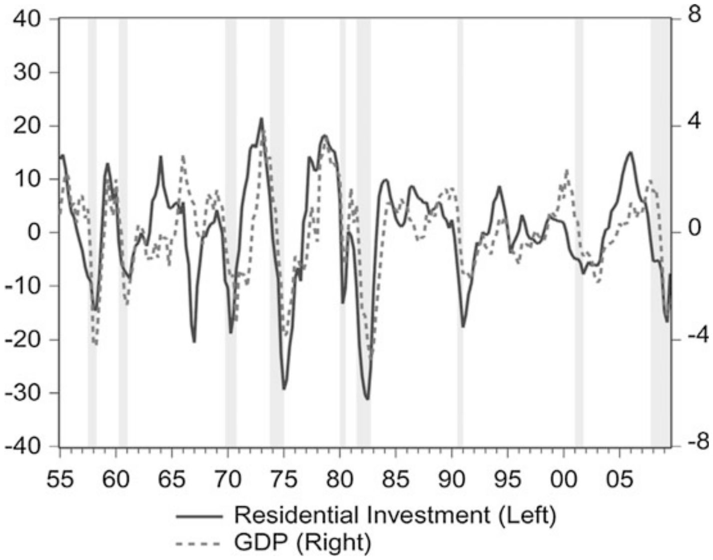
The boom and bust in residential investment and overall production during the first decade of the 21st century can be viewed as a continuation of patterns that are evident in post-Korean War US macroeconomic data. A few features of the data are worth highlighting. First, shown in Fig. 1, residential investment and real GDP are highly correlated at business cycle frequencies.[1] Second, residential investment is much more volatile than GDP and non-residential investment. Table 1 shows that the standard deviation of detrended residential investment is about twice as large as the standard deviation of detrended non-residential investment and more than six times greater than the standard deviation of detrended GDP. This last fact is also evident from the different scales of the axes of Fig. 1. Third, residential investment leads GDP by about one quarter, whereas investment in business capital lags GDP by about one quarter.

Finally, house prices are contemporaneously correlated with GDP and are volatile. An older literature studied the responsiveness of housing prices and quantities to changes in incomes, construction costs and interest rates. A few examples include Alberts (1962), Fair (1972), Poterba (1984), Topel and Rosen (1988).[2] These papers uniformly assume interest rates are fixed, or are set outside of the model, in the sense that interest rates – the price of current consumption relative to future consumption – are not linked to changes in the marginal utility of consumption. As emphasized by Prescott (1986b), interest rates are a key price in any macroeconomic model. So, while the

[1] All data have been logged and HP-Filtered with smoothing parameters $\lambda = 1{,}600$.

[2] See McCarthy and Peach (2002) for a recent example.

Housing and the Business Cycle, Fig. 1 Plot of real detrended residential investment and GDP, 1955:1–2009:3



Housing and the Business Cycle, Table 1 Properties of selected detrended US macroeconomic data, 1955:1–2009:3

variable *X*	Std.	Relative Std.	Correlation of variable X_s and GDP_t						
	Dev	Dev	s = t − 3	t − 2	t − 1	t	t + 1	t + 2	t + 3
	(1)	(2)	(3)	(4)	n	(6)	(7)	(8)	(9)
(a) GDP	1.57	1.00	0.38	0.62	0.85	1.00	0.85	0.62	0.38
(b) Consumption	0.85	0.54	0.50	0.67	0.80	0.83	0.71	0.53	0.33
(c) Res. invest	9.84	6.28	0.64	0.72	0.74	0.64	0.41	0.15	−0.09
(d) Non-res. invest	5.16	3.29	0.08	0.32	0.58	0.80	0.85	0.78	0.63
(e) House prices[a]	3.83	2.44	0.36	0.44	0.48	0.47	0.41	0.33	0.25
(f) Durables quant.	4.47	2.85	0.50	0.66	0.78	0.81	0.62	0.39	0.16
(g) Durables prices	0.96	0.61	0.16	0.06	−0.05	−0.16	−0.23	−0.27	−0.29

Notes: Data are quarterly. All data except the house price data are from the National Income and Product Accounts (NIPA) as produced by the Bureau of Economic Analysis (BEA). The house price data combine data from the Federal Home Finance Agency House Price Index (1975–1986) and the Case–Shiller–Weiss index as made available by Macromarkets, LLC (1987–2009). All variables have been logged and HP-Filtered with smoothing parameter λ = 1,600. Real house and durable prices are computed as the nominal price index divided by price index for consumption of nondurable goods and services

[a]House price data begin in 1975:1

discussion about housing, mortgages, and so-called 'Regulation Q' in these older papers is interesting, they do not fit into the modern literature of business cycles.

The first business cycle models (Kydland and Prescott 1982) did not distinguish residential investment or housing from other forms of capital.[3] The goal of these papers was to understand the fraction of the variability of post-war output that could be explained by a neoclassical growth model (Cass 1965; Brock and Mirman 1972) with stochastic stationary shocks to the level of multifactor productivity around a growing trend. Fairly early on, researchers learned that, while successful along some dimensions, the standard 'real' business cycle model underpredicted the volatility of hours worked. In the data, the standard deviations of HPfiltered log hours worked and log GDP are roughly the same, about 1.7% (Prescott

[3]See Cooley and Prescott (1995) for a review.

1986a). In the first set of real business cycle models, the standard deviation of simulated hours worked was roughly equal to half of the simulated standard deviation of output.

Soon after the study of Kydland and Prescott (1982), researchers worked on adapting the standard real business cycle model such that it would correctly predict that the standard deviation of hours worked and GDP are roughly the same order of magnitude. Early papers by Hansen (1985) and Rogerson (1988) modified the Kydland and Prescott model to allow for indivisible labour supply.[4] Soon after, researchers were augmenting the standard real business cycle model to allow for 'home production'. In a home production model, households receive utility from market consumption, denoted c_m, and home (or non-market consumption), c_n; they accumulate capital to be rented to the market for the purposes of producing market output, k_m, and accumulate capital for the purposes of home production, k_n; and they allocate their time between work in the market, h_m, work at home, h_n, and leisure l. Both the home and market production functions are subject to shocks to productivity.[5] For recent very good summaries of home production models, see Chang and Hornstein (2006) and Gangopadhyay and Hatchondo (2009).

The home production framework was considered an important extension of the original Kydland and Prescott (1982) model.[6] The available data suggest that households spend about as much time engaged in working at home as they do in the market (Juster and Stafford 1991). For this reason changes to the allocation of time across the home and market sectors may be of first-order importance in accounting for the cyclical volatility of market hours. For the purposes of studying the role of housing in the business cycle, the home production models were the first papers to explicitly specify a different purpose for residential investment than investment in market capital (such as spending on equipment and software and on non-residential structures).

Researchers have had a number of challenges in calibrating a basic home production real business cycle model, in part because the inputs into the home production process are not all observed. In sum, researchers have had to take a stand on (a) the elasticity of substitution between home and market consumption in utility; (b) the statistical process characterizing shocks to productivity in the home sector and the correlation of home and market productivity shocks; and (c) what (in the data) should be considered as home capital. Taking each of the points in order: Benhabib et al. (1991) use data on hours worked at home, hours worked in the market, and data on wages from the Panel Study of Income Dynamics to estimate the elasticity of substitution between home and market consumption. They find an elasticity of substitution greater than one, i.e. with preferences of the form

$$u(c_n, c_m, l) = \frac{\overline{c}^{1-\sigma}}{1-\sigma}\upsilon(l) \qquad \text{with}$$

$$\overline{c} = \left[\alpha c_m^{\rho} + (1-\alpha)c_n^{\rho}\right]^{1/\rho}, \tag{1}$$

they estimate $\rho = 0.8$. McGrattan et al. (1997) estimate the process for shocks to home and market productivity using a structural estimation approach that takes advantage of the set of first-order conditions of the model. The authors show that home shocks are 'relatively insignificant', in the sense that 'the result that home production matters does not depend critically on the presence of home technology shocks' (p. 282). Finally, and importantly, when matching model statistics to data, all papers in the home production literature define the stock of home capital in the data as the sum of the stock of housing structures and the stock of consumer durables.

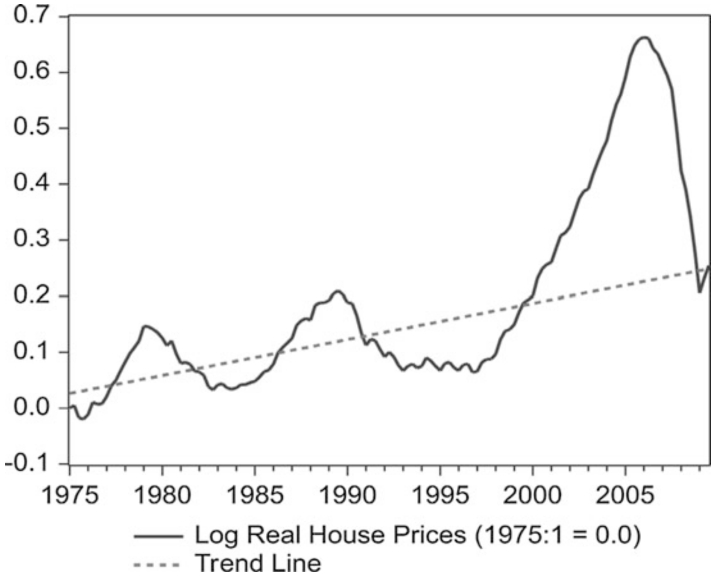
Generally speaking, the home production models have been challenged in matching two features of the data related to investment in the home sector. First, contrary to the data, the models

[4] Hansen (1985) shows that when the standard model is adjusted to allow for indivisible labour supply, the standard deviation of hours worked is equal to three-quarters of the standard deviation of GDP.

[5] See Benhabib et al. (1991) and Greenwood and Hercowitz (1991) for formal treatments.

[6] Gomme and Rupert (2007) argue that the home production model is now the benchmark real business cycle model. Recent and important examples include Fisher (1997), Gomme et al. (2001) and Fisher (2007).

Housing and the Business Cycle, Fig. 2 Plot of real log house prices and trend line, 1975:1–2009:3



tend to predict that investment in business capital is more volatile than investment in home capital (Gomme et al. 2001). Second, without adjustment costs, the home production real business cycle model predicts that investment in market and home capital are negatively correlated (Fisher 1997). In response to a positive shock to market productivity, households add to market capital first, since market capital is required to make more of everything. Later on, households increase their stock of home capital. As mentioned earlier, the data suggest that investment in home capital leads investment in market capital by about two quarters. Both of these points are returned to below.

Davis and Heathcote (2005) argue that home production models are somewhat ill suited to studying the business cycle properties of housing specifically. They make two related points. First, in home production models it is assumed that home capital (the sum of housing and durable goods) is produced using the same technology as all other output.[7] This implies that the real price of housing is constant over time, except for fluctuations due to the presence of adjustment costs. This is clearly at odds with the data. As mentioned earlier, the detrended real price of housing is volatile. But, as shown in Fig. 2, the real price of housing also has an upward trend: After averaging through booms and busts, the trend rate of growth of real house prices has been about 0.5% per year since 1975.[8]

Second, when calibrating home production models, researchers treat the stock of housing and the stock of consumer durable goods (hereafter called ‘durable goods’) as equivalent. But housing and durable goods have quite different properties. To start, housing is a much longer-lived asset than durable goods. The depreciation rate on the housing stock is 1.6% per year whereas it is 21.4% per year for other durable goods (Davis and Heathcote 2005). Second (and related), investment in housing is much more volatile than investment in other durable goods: Table 1 shows that the the standard deviation of residential investment is about twice that of consumer

[7] A notable exception to this is Hornstein and Praschnik (1997), who study production of durable and non-durable goods.

[8] The trend is computed using data from 1975–2002. The trend rate of growth over the entire 1975–2009 period for which we have data is 1.3% per year. Note that 1975 is the starting date for the reliable data series on the price of existing homes – see the notes to Table 1.

durables. Third, residential investment leads GDP by one quarter but consumer durables do not: the highest correlation of detrended real expenditures on consumer durables and GDP is at period *t*, cell f6. Fourth, house prices are about four times more volatile than the price of durable goods (cells e1 and g1). Finally, house prices are positively correlated with GDP (and might even lead GDP; cells e5 and e6), whereas durable goods prices are negatively correlated with GDP.

Davis and Heathcote (2005; hereafter DH) specify and simulate a model that is viewed by some as the first paper that explicitly studies the business cycle properties of housing. The DH model is a frictionless, representative agent, neoclassical growth model that is a relatively straightforward extension of an otherwise standard home production model. The key extension is that DH specify that housing is produced using a different technology from other goods, allowing it to have a nontrivial relative price. The point of the DH paper is to quantify the extent to which a wellcalibrated model can match the fluctuations in residential investment and house prices observed in the data. Any significant model failures in matching the data could then point to a meaningful role for frictions and/or incomplete markets.

The household side of the DH model borrows heavily from the home production literature. DH assume that households receive flow utility of

$$U(c_m, h, l) = \frac{\left(\bar{c}l^{\mu}\right)^{1-\sigma}}{1-\sigma} \text{ with } \bar{c} = c_m^{\alpha} h^{1-\alpha}, \qquad (2)$$

where c_m and l are market consumption and leisure, as before, and h is the stock of housing, not the quantity of home production as in equation (1). As shown by Greenwood et al. (1995), equation (1) reduces to equation (2) when (a) households have log-separable preferences over leisure, market consumption and home consumption, (b) the home produced good is produced using a Cobb-Douglas technology from home capital and labour, and (c) $\rho = 1$. DH argue, contrary to the results of McGrattan et al. (1997), that available data support the assumption of a unitary elasticity of substitution between consumption and housing.[9] DH calibrate utility function parameters to match the average share of time that households spend working and the average ratio of the value of the stock of residential structures relative to GDP.

As noted earlier, the production side of the DH model represents the most significant departure from the home production literature, and many recent macroeconomic models that generate nontrivial house prices borrow aspects of this production structure.[10] DH specify three types of firm in the economy. The first set of firms use capital and labour to make one of three intermediate goods called ‘construction’, ‘manufacturing’ and ‘services’. Output of intermediate good *i* in period *t*, denoted x_{it}, for i equal to b (construction), m (manufacturing) and s (services), is specified as

$$x_{it} = k_{it}^{\theta_i} \left(z_{it} n_{it}\right)^{1-\theta_i}, \qquad (3)$$

where k_{it} and n_{it} are the capital and labour employed in the production of good i and z_{it} is a sector-specific productivity shock. θ_i is the capital share of producers of intermediate goods i, which can vary for $i = b, m, s$. In contrast to the home production function in the home production models, DH show that all aspects of this production technology are directly observable with available data. DH use the Gross Domestic Product by Industry Tables, produced by the Bureau of Economic Analysis (BEA), to identify the capital shares θ_i; and, given a value of θ_i, DH use data from the Gross Domestic Product by Industry tables and the Fixed Asset tables, also produced by the BEA, to uncover time series data for k_{it}, n_{it}, and z_{it}.[11]

A second set of firms uses a Cobb-Douglas technology to combine the intermediate goods into two ‘final’ goods. The first final good can be costlessly split into consumption and investment in business capital; the second final good is

[9] Additional evidence supporting this claim is in Davis and Ortalo-Magné (2009).

[10] See Dorofeenko et al. (2009), Iacoviello and Neri (2010), Kahn (2009) and Kiyotaki et al. (2008), to name just a few recent examples.

[11] See the Data Sources Appendix of DH for more details.

residential investment. DH specify output of final good j in period t as y_{jt} for $j = c$ (consumption and business investment) and $j = d$ (residential investment) to equal

$$y_{jt} = b_{jt}^{B_j} m_{jt}^{M_j} s_{jt}^{S_j}, \tag{4}$$

where B_j, M_j and S_j are the value-added shares of construction, manufacturing and services in the production of final good j. DH show that these shares are identifiable using data from the Input–Output tables, also produced by the BEA.[12] DH show that residential investment is much more construction intensive than the other final good, which turns out to be important in explaining the relative volatility of residential investment.

A final set of firms in the DH model combine new residential investment with new land (made available by the government each period) to create new housing units. The specific production function for the quantity of new housing built in period t, y_{ht}, is

$$y_{ht} = x_{lt}^{\varphi} x_{dt}^{1-\varphi}, \tag{5}$$

where x_{lt} is the amount of newly developable land and x_{dt} is residential investment (produced according to equation 4). DH identify the parameter φ based on results about the share of the value of new housing attributable to raw land costs from an internal memo of the US Census Bureau.

Thus the DH model has three ingredients that allow for potentially interesting time-series variation in house prices. First, the statistical process (mean growth rate, variance, and autocorrelation) for z_{it} is allowed to vary across the construction, manufacturing and services sectors. Second, firms that produce residential investment use different combinations of these three intermediate goods than do firms that produce the other final good. The price of housing has a long-term upward trend according to the DH model for these two reasons: DH show that z_{bt} has zero trend growth, and construction accounts for about 50% of the value-added in residential investment (compared to 3% of the value-added of the other final good). Finally, new housing requires both new land and new residential investment, and new land is in fixed supply. The scarcity of land affects both the trend and the variance of house prices in the model.

Some key second moments from the data and from simulations of the DH model are reported in Table 2. The information in this table is copied directly from Table 10 of DH.[13] Rows (a) and (b) of Table 2 show that the DH model underpredicts the volatility of consumption and of hours worked. In this regard, the results of DH are similar to previous models. However, the DH model has great success in replicating key facts about residential investment, namely that residential investment is about twice as volatile as business investment (rows c and d) and that residential and business investment are positively contemporaneously correlated (row f). DH show that the low depreciation rate on structures and the relatively high labour share of the construction sector are largely responsible for replicating the relative volatilities of residential and business investment. With a low depreciation rate, it is possible for households to ‘concentrate residential investment in periods of high productivity’ (p. 774); and, with a high labour share of the construction sector, ‘it is easier to expand output rapidly the more important is labour in production, since holding capital constant, the marginal product of labour declines more slowly’ (p. 774). The positive correlation of residential and business investment is attributable to the fact that new housing needs new land as an input in production, and new land is in fixed supply. In this regard, land in the DH model acts analogously to adjustment costs in the home production models.

Although the DH model replicates some key features of housing investment, it does not match some key features of the housing data. The DH model cannot generate that residential investment

[12] DH calibrate these shares using data from 1992. The DH specification is inconsistent with the sectoral decline in manufacturing over the post-war period.

[13] See the notes to Table 2 for details.

Housing and the Business Cycle, Table 2 Business cycle properties of the Davis and Heathcote (2005) model[a]

Standard deviations relative to GDP			
	Variable	Data	DH
(a)	Consumption	0.78	0.48
(b)	Hours worked	1.01	0.41
(c)	Res. invest	5.04	6.12
(d)	Non-res. invest	2.30	3.21
(e)	House prices	1.37	0.40
Period *t* Correlations			
	Variables	Data	DH
(f)	Res. and non-res. invest.	0.25	0.15
(g)	Res. invest. and house prices	0.34	−0.20

Notes: All results and data in this table are taken from Table 10 of Davis and Heathcote (2005). Davis and Heathcote use annual data over the 1948–2001 range; they HP-Filter the data with smoothing parameters $\lambda = 100$. The use of annual data and the different sample range explain some of the discrepancies between this table and the data reported in Table 1.

leads GDP and business investment lags GDP (not shown).[14] Second, the DH model cannot replicate two important features of house prices. Shown in row (e) of Table 2, the DH model under-predicts the volatility of house prices by about a factor of three. The DH model also predicts that residential investment and house prices are negatively contemporaneously correlated, whereas in the data they are positively correlated (row g). Future researchers are actively focusing on reconciling these issues.

See Also

▶ Household Production and Public Goods
▶ Housing Supply
▶ Housing Wealth
▶ Urban Housing Demand

Bibliography

Alberts, W.W. 1962. Business cycles, residential construction cycles, and the mortgage market. *Journal of Political Economy* 70(3): 263–281.

Benhabib, J., R. Rogerson, and R. Wright. 1991. Homework in macroeconomics: Household production and aggregate fluctuations. *Journal of Political Economy* 99(6): 1166–1187.

Brock, W.A., and L.J. Mirman. 1972. Optimal economic growth and uncertainty: The discounted case. *Journal of Economic Theory* 4(3): 479–513.

Cass, D. 1965. Optimum growth in an aggregative model of capital accumulation. *Review of Economic Studies* 32: 233–240.

Chang, Y., and A. Hornstein. 2006. Home production. Federal Reserve Bank of Richmond Work Paper 06-04.

Cooley, T.F., and E.C. Prescott. 1995. Economic growth and business cycles. In *Frontiers of business cycle research*, ed. T.F. Cooley. Princeton: Princeton University Press.

Davis, M.A., and F. Ortalo-Magné. 2009, forthcoming. Household expenditures, wages, rents. *Review of Economic Dynamics*.

Davis, M.A., and J. Heathcote. 2005. Housing and the business cycle. *International Economic Review* 46(3): 751–784.

Dorofeenko, V., G.S. Lee, and K.D. Salyer. 2009. Risk shocks and housing markets. Working Paper.

Fair, R.C. 1972. Disequilibrium in housing models. *Journal of Finance* 27(2): 207–221.

Fisher, J.D.M. 1997. Relative prices, complementarities, and comovement among components of aggregate expenditures. *Journal of Monetary Economics* 39: 449–474.

Fisher, J.D.M. 2007. Why does household investment lead business investment over the business cycle? *Journal of Political Economy* 115: 141–168.

Gangopadhyay, K., and J.C. Hatchondo. 2009. The behavior of household and business investment over the business cycle. *Federal Reserve Bank of Richmond Economic Quarterly* 95(3): 269–288.

Gomme, P., and P. Rupert. 2007. Theory, measurement and calibration of macroeconomic models. *Journal of Monetary Economics* 54(2): 460–497.

Gomme, P., F. Kydland, and P. Rupert. 2001. Home production meets time to build. *Journal of Political Economy* 109: 1115–1131.

Greenwood, J., and Z. Hercowitz. 1991. The allocation of capital and time over the business cycle. *Journal of Political Economy* 99(6): 1188–1214.

[14] Fisher (2007) has made some headway on this issue, but his approach of including home capital as a direct input to market production is not without controversy.

Greenwood, J., R. Rogerson, and R. Wright. 1995. Household production in real business cycle theory. In *Frontiers of business cycle research*, ed. T.F. Cooley. Princeton: Princeton University Press.

Hansen, G.D. 1985. Indivisible labor and the business cycle. *Journal of Monetary Economics* 16(3): 309–327.

Hornstein, A., and J. Praschnik. 1997. Intermediate inputs and sectoral comovement in the business cycle. *Journal of Monetary Economics* 40(3): 573–595.

Iacoviello, M., and S. Neri. 2010, forthcoming. Housing market spillovers: Evidence from an estimated DSGE model. *American Economic Journal Macro*.

Juster, F.T., and F.P. Stafford. 1991. The allocation of time: Empirical findings, behavioral models, and problems of measurement. *Journal of Economic Literature* 29(2): 471–522.

Kahn, J. 2009. What drives housing prices. Working paper.

Kiyotaki, N., A. Michaelides, and K. Nikolov. 2008. Winners and losers in housing markets. *London School of Economics*. Unpublished manuscript.

Kydland, F.E., and E.C. Prescott. 1982. Time to build and aggregate fluctuations. *Econometrica* 50(6): 1345–1370.

McCarthy, J., and R.W. Peach. 2002. Monetary policy transmission to residential investment. *FRBNY Economic Policy Review* 139–158.

McGrattan, E.R., R. Rogerson, and R. Wright. 1997. An equilibrium model of the business cycle with household production and fiscal policy. *International Economic Review* 38(2): 267–290.

Poterba, J.M. 1984. Tax subsidies to owner-occupied housing: An asset-market approach. *Quarterly Journal of Economics* 99(4): 729–752.

Prescott, E.C. 1986a. Theory ahead of business cycle measurement. *The Federal Reserve Bank of Minneapolis Quarterly Review* 10(4).

Prescott, E.C. 1986b. Response to a skeptic. *The Federal Reserve Bank of Minneapolis Quarterly Review* 10(4).

Rogerson, R. 1988. Indivisible labor, lotteries and equilibrium. *Journal of Monetary Economics* 21(1): 3–16.

Topel, R.H., and S. Rosen. 1988. Housing investment in the United States. *Journal of Political Economy* 96(4): 718–740.

Housing Markets

John M. Quigley

The principal features that distinguish housing from other goods in the economy are its relatively high cost of supply, its durability, its heterogeneity, and its locational fixity. Of course, many other commodities exhibit one of these features. However, the interaction of these distinguishing characteristics complicates theoretical and empirical analyses of the housing market.

Durability, heterogeneity and fixity together indicate that the housing market is really a collection of loosely related but segmented markets for particular packages of underlying commodities differentiated by size, physical arrangement, quality and location. These sub-markets are connected in a predictable way. At neighbouring locations, differences in prices between submarkets cannot exceed the cost of converting a housing unit from one sub-market to another. At different sites, variations in prices within any sub-market cannot exceed the transport cost differentials for the marginal consumer. However, a price-inelastic demand for some of the attributes jointly purchased, combined with inelastic supply in the short-run, can make the pattern of housing prices rather complex, even in a market in temporary equilibrium.

Analyses of the supply and demand for housing are complicated by these somewhat peculiar characteristics. Consider the demand side of the market; take the case of renters. Presumably, quantity demanded depends upon price and income. The ‘quantity’ in this case consists of a vector of attributes. This quantity can, of course, be summarized by its market rent, but the rent of a dwelling unit is neither a price nor a quantity. Rent is measured in the units of price-times-quantity, and it is a formidable task to disentangle the two for statistical purposes.

The third variable included in the demand relationship, income, is equally difficult to measure in the housing market. Given the high costs of transforming residential capital and the high costs of moving, it follows that housing decisions are based upon some long-run or ‘permanent’ notion of income (Friedman 1957), a concept which has proved difficult to specify without ambiguity.

The relevant notion of the price of housing for the decisions of owner occupants is even more elusive. By observing transactions in the market, the value (V) of an owner-occupied dwelling can be ascertained. Under familiar but quite restrictive competitive conditions (infinite durability, no depreciation or maintenance, capital gains or taxes), the annual rent (R) for this dwelling is:

$$R = iV \tag{1}$$

where i is the rate of interest, assumed equal to the mortgage interest rates. Under more realistic conditions the annual cost of a dollar of residential capital, the so-called 'user cost' of capital (Jorgenson 1971), can be estimated. It varies with four broad classes of circumstance: (a) the expected rate of increase in housing equity, in alternative investments (including rents), and the fraction of the purchase financed by borrowing; (b) the type of mortgage and the holding period; (c) the rate of depreciation (δ) gross of maintenance expenditures, and the fixed cost of buying and selling the residence; and (d) the marginal tax rates for income (T_y), property (T_p), and capital gains (T_g), and the rules for tax liability. Assume four classes of simplified market conditions; (a) the rate of increase of rents equals the rate of increase in housing values (γ); (b) the net mortgage rate $i(1 - T_y)$ equals the net rate of return on alternative investment, for a fixed rate mortagage with an infinite holding period; (c) the buying and selling costs are zero; and (d) interest and property taxes are deductible from taxable income and the imputed return from living in a dwelling is not taxed. Under these simplified conditions (Rosen 1985), the annual cost of housing capital may be represented as:

$$R = \left[(1 - T_y)i - (1 - T_g)\gamma + \delta + (1 - T_y)T_p\right]V \tag{2}$$

Equation (2) emphasizes the importance of taxes and capital gains, as well as interest rates, in defining the effective price of housing services to an owner-occupant. For example, the expectation of capital gains (γ) decreases the effective cost of housing. This may be partly offset by capital gains taxation (T_g), but in many countries housing transactions are essentially free of this tax. As long as capital gains tax rates are less than marginal income tax rates, general inflation (i.e. increases in interest rates and capital gains) reduces the cost of home ownership more for higher-income households. Tax provisions, especially the tax-free nature of imputed rent, reduce the relative cost of home ownership at higher-income levels.

Together, these price and income concepts have been used to estimate the parameters governing housing demand and tenure choice. There seems to be some general agreement that: the elasticity of demand for the composite housing good is low for annual income, but much higher, approaching one, for average (one 'permanent') income; and that housing demand is price-inelastic. Evidence also suggests that tenure choice is rather insensitive to the relative prices of owning and renting dwellings.

The spatial pattern of housing and households defines the economic geography of urban life and the development of metropolitan regions. Modern economic theory which explains these spatial patterns (e.g. Muth 1969) owes much to the German economic geographers of the 19th century. In particular, the seminal work of von Thünen (1826) considers the question of agricultural production on an isolated plain relative to a central market place. The modern treatment considers the residential locations of workers employed at a central worksite. Workers (or farmers) are willing to pay a premium for central locations to reduce transport costs, so housing (or agricultural land) must become cheaper at more distant locations. To illustrate, assume consumers derive utility $U(h, x)$ from housing (h) and other goods (x). They confront a budget constraint which requires them to allocate exogenous income Y between housing consumption, whose price $P(t)$ varies with distance t, other goods (at a price of one), and transportation costs $k(t, y)$, which vary with distance and income; i.e., $y = x + P(t)h + k(t, y)$. Maximizing utility subject to this constraint yields:

$$hP' = -\partial k/\partial t. \tag{3}$$

The consumer chooses to locate at that point where the marginal savings from cheaper housing exactly offset the marginal costs of additional commuting. Clearly, the location chosen depends upon the household's preferred amount of housing. It can be shown (by differentiating (3) with respect to income) that higher-income households will choose less accessible ('suburban') locations under reasonable conditions (i.e., as long as the income elasticity of housing demand exceeds the income elasticity of marginal transport costs). The theory thus provides an explanation for the central location of the poor, an explanation which is quite distinct from competing theories based upon the prior location of the oldest housing stock (e.g. Burgess 1925).

The concentrations of low-income households and the existence of slum housing raise several questions about the operation of housing markets, the 'filtering' of dwellings and the role of externalities in housing. The concept of filtering arises from the observation that 'most households live in second-hand housing, even the Queen of England' (Grigsby 1963). If a middle-income household is induced to move to a newly built dwelling, it sets off a chain of moves, as the rent which can be charged for a vacated dwelling declines, making each one available to households of lower income. Under what circumstances does the filtering process make lower-income households better off? If the quality of housing were truly exogenous (for example, if it were only related to the vintage of the dwelling), then low-income households could benefit directly from the filtering process as higher-quality housing became available. On the other hand, if housing quality is sufficiently responsive to landlord maintenance decisions, then demand price declines may be matched by quality declines. It thus requires a very special view of housing to conclude that the 'filtering' process will lead to improved housing for the poor, even under static conditions.

Externalities in the housing market may arise from physical, 'social', or pecuniary conditions. The propinquity of dwellings does suggest that the maintenance decisions of landlords may be subject to a kind of 'prisoners' dilemma' in low-income neighbourhoods. Because the rent of a unit reflects the quality of adjacent dwellings, owners of neighbouring properties may maximize returns if 'the other guy' invests. Thus, housing or rehabilitation investment which is jointly profitable may not be undertaken at all.

The policy prescription for economic efficiency in this case is joint ownership or decision-making (or public renewal). But suppose the externality is of a social or demographic character. For example, suppose members of each of two races can only tolerate neighbourhoods in which they constitute at least x per cent of all households. Under such circumstances, Thomas Schelling (1978) has shown that integrated neighbourhoods will result, at least for some distributions of x. But he has also shown that this result may be highly unstable; the integrated outcome could easily unravel in response to an exogenous movement of a few households. Either segregated or integrated solutions may be 'efficient' in some very narrow sense. Suppose instead that members of one group have a uniform aversion to living with members of another group (or 'a taste for discrimination', in the terminology of Becker 1957). In this case a segregated pattern of occupancy may satisfy narrow allocative efficiency principles, since those who discriminate will be required to 'pay for their prejudices'. These different economic models of discrimination are disturbing, but they have only limited application to the housing market, since most empirical evidence suggests a different pattern of prices. Minority households pay higher prices for otherwise comparable housing, at least in North American markets (e.g. Kain and Quigley 1975).

As noted, the level of new construction is subject to great fluctuation: long term, in response to immigration and population readjustment (Kuznets 1952), as well as short term, in response to interest rates and credit availability. To some extent the organization of the industry may be a reflection of this cyclicity. The industry is still dominated by small firms, often undercapitalized, producing low levels of output. The relatively low rate of productivity growth in housebuilding may thus reflect an adjustment to cyclicity in demand as well as inherent technological considerations.

See Also

- ▶ Monocentric Models in Urban Economics
- ▶ Property Taxation
- ▶ Tiebout Hypothesis
- ▶ Urban Economics
- ▶ Urban Housing

Bibliography

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

Burgess, E.W. 1925. The growth of the city. In *The city*, ed. R.E. Park, E.W. Burgess, and R.D. McKenzie. Chicago: University of Chicago Press.

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

Grigsby, W. 1963. *Housing markets and public policy*. Philadelphia: University of Pennsylvania Press.

Jorgenson, D.W. 1971. Econometric studies of investment behavior: A survey. *Journal of Economic Literature* 9: 1111–1147.

Kain, J.F., and J.M. Quigley. 1975. *Housing markets and racial discrimination: A microeconomic analysis*. New York: Columbia University Press.

Kuznets, S. 1952. Long term changes in national income of the United States since 1870. In *Income and wealth*, Series II, ed. S. Kuznets. London: Cambridge University Press.

Muth, R.F. 1969. *Cities and housing*. Chicago: University of Chicago Press.

Quigley, J.M. 1979. What have we learned about housing markets. In *Current issues in urban economics*, ed. P. Mieszkowski and M. Straszheim. Baltimore: Johns Hopkins Press.

Rosen, H.S. 1985. Housing subsidies: Effects on housing decisions, efficiency, and equity. In *Handbook of public economics*, vol. I, ed. A.J. Auerbach and M. Feldstein. Amsterdam: North-Holland.

Schelling, T. 1978. *Micromotives and macrobehavior*. New York: W.W. Norton.

von Thünen, J.H. 1826. *Der isolierte Staat in Beziehung auf Landwirtschaft und Nationalökonomie*. Hamburg.

Housing Policy in the United States

John M. Quigley

Abstract

The most significant and most expensive housing policy in the United States is the treatment of owner-occupied housing for tax purposes. This treatment of housing under the tax code is analogous to that in many other countries (for example, Sweden), but certainly not in all developed countries (for example, Canada). Federal subsidies to US renter households are much smaller. Policy has evolved from programmes in which the government built, owned, and managed dwellings to programmes emphasizing housing demand through vouchers and rent certificates awarded to eligible households.

Keywords

Great Depression; Homeownership; Housing; Housing costs; Housing expenditures; Housing externalities; Housing finance; Housing policy in the United States; Housing subsidies; Housing tax credit; Income tax; Inflation; Interest rates; Internal Revenue Service (IRS); Land-use regulation; Rent control; Social capital; Tax expenditures; Taxation of capital income

JEL Classifications

R31

Public concern over housing arises from three sources. First, housing is the single largest expenditure item in the budgets of families and individuals in most modern economies. The average household in western Europe and the United States devotes more than one quarter of its income to housing expenditures. Thus, increased efficiency in the provision of housing services or reduced occupancy costs can have a large impact on non-housing consumption and household wellbeing. Second, consumers' housing and location choices condition many other aspects of the quality of urban life. For example, the transport, schooling, and neighbourhood opportunities of urban households are themselves greatly affected by the housing opportunities available to them. Third, it is widely presumed that there are significant externalities in housing consumption. These external effects range all the way from the consequences of the social and physical isolation of those living in low-income residential neighbourhoods to the presumed benefits of the 'social capital' and the increased political participation of households who own their homes.

In the United States, important policies providing subsidies to housing consumers are made by the central ('federal') government. Other policies governing housing – the regulation of housebuilding, service provision, and occupancy – are determined by local governments. At the national level, subsidies provided to selected housing consumers and producers are implemented by two government agencies: the Internal Revenue Service (IRS) and the Department of Housing and Urban Development (HUD). The policies administered by the IRS are clearly more important quantitatively, and they have large welfare effects.

The Federal Tax Code

The IRS administers two housing subsidy programmes: the tax expenditures to owner-occupants for housing consumption specified in the personal income tax code, and the tax expenditures for builders of rental housing under the Low Income Housing Tax Credit programme specified in the Tax Reform Act of 1986. This latter programme is small, having originated in the Tax Reform Act of 1986. The former programme is large, and has existed in its current form since the personal income tax was established in 1915. Indeed, the benefits to homeowners under these tax policies are among the most generous in the developed world. (But the form of these subsidies is certainly not unique to the United States. See Englund 2003, for a comparative discussion.)

Consider an individual who chooses between an investment in owner-occupied housing and an equivalent investment in some other asset – common stocks, say. The investment in owner-occupied housing offers three distinct tax advantages. First, under the US Internal Revenue Code, the returns on the investment in owner-occupied housing are untaxed (these returns are in the form of the housing services consumed in any year). In contrast, the dividends yielded by common stock are reported as income and are taxed in the year accrued. Second, capital gains arising from the housing investment can be deferred indefinitely. Moreover, a large capital gains exclusion is available to those over the age of 55. In contrast, capital gains in the stock market are taxed in the year they are realized. Third, some of the expenses associated with homeownership, notably property taxes and mortgage interest payments, can be itemized as deductions in computing federal tax liability under the personal income tax. No other interest payments are deductible as personal expenses under the Internal Revenue Code. This favourable treatment also extends to personal income taxation under the laws of all of the 50 states.

The net effect of these provisions of the US tax law is to reduce the price of homeownership, relative to renting, by a sizeable amount. Moreover, as a result of these policies, the relative price of homeownership varies by income level and the level of inflation.

It is useful to think of the price of homeownership as the cost of using the stock of residential capital. The rent R for using a unit of capital V is merely

$$R = iV, \tag{1}$$

where i is the real interest rate. i is simply the price of using a unit of capital V for a year. Housing is subject to local property tax at effective rate t. Annual expenditures of $100d$ per cent are required to maintain the property and to offset depreciation. The owner can expect real capital gains at a rate g. Let π be the rate of inflation. For housing, the user cost relationship is thus

$$R_1 = [(i + \pi) - t - d - (g + \pi)]V, \tag{2}$$

where the term in square brackets is the user cost of residential capital. Note that, in the absence of tax considerations, the user cost is insensitive to the level of inflation π. Now suppose nominal capital gains are untaxed and that mortgage interest payments and property taxes are deductible from gross income. Suppose net income is taxed at the rate of T per cent. Under these circumstances the user cost relationship is

$$R_2 = ([i + \pi][I - T] + t[I - T] + d - [g + \pi])V, \tag{3}$$

or

$$R_2 = R_1 - T(i + \pi + t)V. \tag{4}$$

The system of taxes leads to a reduction in the net price of housing capital by the amount of the second term. Note that the after-tax cost of homeownership declines with the value of the house, the real interest rate, the property tax rate, and the marginal income tax rate.

If federal tax rates increase with income or if higher-income households live in jurisdictions with higher property tax rates, the cost of homeownership declines with income. More important, as long as housing is a normal good with a positive income elasticity, the net cost of homeownership declines with income. Furthermore, a given level of inflation in the economy reduces the user cost more for higher-income than for lower-income homeowners.

More generally, the analysis shows that the costs of homeownership are sensitive to macroeconomic stabilization policies and to the structure of income tax rates. The marginal tax rates of the highest-income US households fell from 70 per cent to 30 per cent and then rose to 40 per cent during the 1980s and 1990s, before falling again in 2001. At the same time, the inflation rate plummeted from 15 per cent to less than three per cent. These changes have meant that the implicit policy toward housing and homeownership varied substantially.

For example, at reasonable values of the variables in Eq. 4 (say, $i = g = 3\%$, $t = d = 2\%$, $T = 30\%$), then as inflation declines from six per cent to 1 per cent, the after-tax user cost of residential capital roughly doubles. Similarly, at reasonable values of the variables (for example, $\pi = 3\%$ and, as before, $i = g = 3\%$, $t = d = 2\%$), then, as income tax rates decrease from 40 per cent to 20 per cent, the after-tax cost of owner occupancy increases by more than one-third. These are substantial price changes induced entirely by taxation and macroeconomic considerations which may be completely unrelated to any objective of housing policy.

These reductions in the user cost of housing capital may be expected to increase housing consumption; reductions in the price of owning relative to renting may be expected to increase homeownership. But econometric research suggests that the demand for housing is moderately price-inelastic. It also appears, at least for the United States, that the elasticity of homeownership with respect of the relative price of homeownership is quite small. Thus, the effects of these large subsidies on housing outcomes are quite small.

In contrast, the magnitude of the implicit subsidy arising from the personal income tax code is large and extremely regressive. The subsidy is available only to owners, who are typically more affluent than renters, and only to those who find it advantageous to itemize their deductions in computing their tax liabilities. (Under US tax law, households may claim a ‘standard’ deduction for expenses or they may list deductions separately. The propensity to itemize deductions separately increases with income.) Finally, as noted above, for those owners who do itemize deductions, the magnitude of the subsidy increases with income.

The second programme administered by the IRS, the low-income housing tax credit, was established in 1986 and expanded in 2001. Under this programme, tax credits are remitted to each state in proportion to population. These credits are awarded by states to developers who propose new construction of housing reserved for low-income tenants who pay 30 per cent of their incomes in rent. The credits, in turn, are sold to firms and high-income individuals, and the proceeds are invested in the designated projects.

The IRS monitors the compliance of these projects with the tax law requiring occupancy by low-income tenants for a 15-year period after construction.

The revenues forgone by the federal treasury as a result of these programmes are routinely estimated by the Joint Committee on Taxation of the Congress. The revenue costs of these subsidies are large. In 2005, for example, it is estimated that tax expenditures for owner-occupied housing totalled about \$147 billion – \$69 billion for the mortgage interest deduction, \$33 billion for the capital gains exclusion on home sales, \$28.6 billion for the exclusion of imputed rent, and \$16.6 billion for the property tax deduction. It is estimated that more than half of the benefits of the tax expenditures for homeowners accrue to the top 15 per cent of the income distribution.

In contrast, in 2005 the tax expenditures arising from the low-income housing tax credit were about $4.8 billion (in present value terms). Presumably much of this benefit accrues to low-income renters.

A more relevant benchmark for the costs of these tax expenditures may be a comparison with the housing programmes managed by HUD, whose principal beneficiaries are low-income households. Direct expenditures under these programmes are currently $41 billion, or about 28 per cent of the tax expenditures on behalf of owner occupants.

Subsidies for Renters

Federal housing policies for renters administered by HUD provide subsidies to about a third of low-income households. These programmes have evolved from those providing housing owned and managed by government to those providing direct cash assistance for deserving renters. The Public Housing Program was established in 1937 to subsidize local governments in building housing for those temporarily unemployed and also in providing construction jobs for unemployed urban labour during the Great Depression. Until the end of the 1970s, the programme subsidized virtually all of the capital costs of designated public housing dwellings and none of the operating costs. Since rent rolls were fixed at 25–30 per cent of tenant income, project managers who chose to serve households with the lowest incomes faced severe budgetary problems. Changes in the subsidy formulas helped local managers avoid this Hobson's choice, but the legacy of the original subsidy formula, the overcapitalization of projects to economize on maintenance expenses, is still manifest in the long-lived capital produced by the Public Housing Program.

The private sector was first induced to build, manage and provide rental dwellings for low-income tenants in the 1960s, through generous depreciation allowances provided to limited dividend corporations (under programmes such as Section 235 of the Housing Act of 1968). But it was not until 1974 that the subsidy provided to deserving tenants was divorced from the cost of supplying newly constructed housing.

The innovation in Section 8 of the Housing Act of 1974 was a programme of project-based housing assistance based upon long-term contracts in which the federal government guaranteed that participating landlords would receive the average rent in the local housing market (rather than the cost of building new housing). Low-income households pay 30 per cent of their incomes to a participating landlord and the difference, up to the 'fair market rent' in the housing market, is supplied under federal contract.

The radical departure to subsidize directly the demanders of low-income housing rather than the builders and suppliers of that housing was thoroughly tested by the Housing Allowance Experiments of the 1970s and 1980s, the most expensive social experiment in history, and the results were incorporated over time into the current Housing Choice Voucher Program which allocates vouchers or certificates to local authorities for distribution to low-income households. Under this programme, a qualifying household receives a voucher which pays the difference between 30 per cent of tenant income and the 'fair market rent'. This programme is administered by Local Housing Authorities, who screen applicants and certify eligibility. Under current practice, households with incomes below 80 per cent of the area median income are eligible for vouchers, but three-quarters of the vouchers are reserved for very low-income households, those whose incomes are below 30 per cent of the area median income. In principle, the voucher is completely portable. It can be used anywhere by a recipient to enter into a rental contract within 90 days of issue.

Vouchers offer several clear advantages over the alternative supply oriented housing subsidy programmes. First, they are considerably cheaper per household served than programmes linking subsidies to construction costs, including the Public Housing Program, but also the Low Income Housing Tax Credit Program. Second, they remove questions about the location of dwellings occupied by low-income subsidized households from the local political process.

Third, they preserve the anonymity of the low-income recipients of these subsidies. Fourth, they foster the spatial decentralization of the low-income population, reducing the concentration of disadvantaged households in particular neighbourhoods. Fifth, they better facilitate the operation of the labour market by encouraging recipients to live closer to actual or potential worksites.

Although new commitments by HUD for subsidies to low-income renters are concentrated in the voucher programme, the legacy of past programmes will remain for a considerable period. For example, in the last year for which complete data are available (1998), 1.3 million units of government-owned public housing were used to provide housing subsidies, as were 1.0 million units of Section 8 project-based housing and 750,000 units of housing produced by other supply-oriented programmes. In contrast, 1.4 million households were subsidized by tenant-based voucher programmes.

Local housing regulations impose a potentially serious impediment to the efficiency of vouchers as a vehicle for housing subsidies. With local property taxes as the basis for local service provision, it is often in the fiscal interests of individual governments to limit the construction of new housing and to restrict the construction of high-density housing. The land-use regulations of individual jurisdictions are not well coordinated regionally in the United States, and the resulting regulatory pattern may make the housing supply relatively inelastic. This may lead to higher housing prices in response to increases in demand throughout the market, and it may mean that housing may be less available to voucher recipients in some metropolitan areas.

Despite these real concerns, the most important factor keeping the rent-to-income ratio of the poor high is the limited availability of housing subsidies. In 2001, it was estimated that almost 14.5 million renter households paid more than 30 per cent of their incomes on rent, and more than 7 million paid more than half of their incomes on rent. In contrast, only about 5 million renter households received subsidies from all federal government housing programmes.

See Also

- ▶ Housing Supply
- ▶ Local Public Finance
- ▶ Markets
- ▶ Rent Control
- ▶ Residential Real Estate and Finance
- ▶ Urban Economics
- ▶ Urban Housing Demand

Bibliography

Englund, P. 2003. Taxing residential housing capital. *Urban Studies* 40: 937–952.

Gabriel, S. 1996. Urban housing policy in the 1990s. *Housing Policy Debate* 7: 673–693.

Quigley, J. 2000. A decent home: Housing policy in perspective. *Brookings–Wharton Papers on Urban Affairs* 1(1): 53–100.

Quigley, J., and S. Raphael. 2004. Is housing unaffordable? Why isn't it more affordable? *Journal of Economic Perspectives* 18(1): 191–214.

Housing Supply

Raven E. Saks

Abstract

This article reviews the key factors that influence the elasticity of housing supply in the United States. When housing demand increases, the response of the housing stock is determined by physical construction costs (materials, labour and land) and government regulation. During the past several decades, a widespread adoption of restrictive land-use policies has substantially reduced the elasticity of housing supply in many parts of the United States. As the housing stock has become more inelastic, housing supply conditions have become progressively more important for understanding the dynamics of house prices and the form of urban growth and decline.

Keywords

Construction costs; House prices; Housing supply; Income distribution; Land-use regulation; Public housing; Real estate; Urban economics; Urban growth; Zoning

JEL Classification

R31

The supply of housing has exerted a growing influence on the dynamics of US housing markets since the 1970s. An increase in aggregate housing demand is ultimately met by an expansion of the housing stock somewhere in the United States, but the response of the local housing supply to a change in demand varies substantially across geographic locations. Some metropolitan areas, like Charlotte, NC, have grown rapidly with only moderate increases in house prices, suggesting that the supply of housing is elastic in these locations. By contrast, in locations like New York City, large increases in house prices and low levels of construction activity indicate a considerably more inelastic supply. Places experiencing persistent declines in housing demand, like Detroit, illustrate yet another aspect of the housing supply. The durability of housing prevents sharp contractions of the housing stock when housing demand falls, limiting population outflows from these locations and contributing to the persistence of urban decline. The heterogeneity of supply responses across local housing markets has become a topic of great interest among urban economists, particularly as the supply of housing has become more inelastic in a growing number of areas in the United States.

Increases in Supply

The response of the housing stock to an increase in demand is governed by the need for three elements: a physical structure, land, and government approval to put the structure on the land. The costs associated with each of these elements determine the extent to which increases in demand are accompanied by an expansion of the housing stock or by higher house prices. A combination of rising prices and declines in construction activity in many parts of the United States suggests that there has been a secular decline in the elasticity of housing supply since the 1970s. Low barriers to entry and exit and the absence of significant returns to scale combine to make the homebuilding industry fairly competitive, so that changes in the elasticity of housing supply mainly reflect the costs of the three component elements.

Structure Construction Costs

The technology of homebuilding has not changed dramatically since the first half of the twentieth century, so the costs of building a housing structure are largely determined by the input prices of construction materials and labour. Although these costs account for the majority of new construction outlays, their importance has declined over time, and they have accounted for no more than 65% of the total market value of residential real estate since the mid-1980s (Davis and Heathcote 2005). Typically, labour makes up about two-thirds of these physical costs, and geographic variation in construction worker wages is the primary source of differences in construction costs across locations. The response of the housing supply to changes in physical structure construction costs is relatively elastic (Somerville 1999b; Gyourko and Saiz 2006), but increases in these costs cannot account for the entire decline in residential construction activity that has occurred during the past several decades (Glaeser et al. 2005a).

Land Availability

The housing supply is also a function of the amount of land available for new residential construction. Topography, the existence of bodies of water, and the geologic composition of the land can all contribute to the difficulty of building new houses, reducing the elasticity of housing supply. In a sample of 45 large cities, Rose (1989) estimates that about 30% of the variation in land prices across locations can be explained by natural restrictions on the supply of land. The availability of land is clearly important in explaining why some cities grow more quickly than others, but it is unlikely to be able to account for an inelastic housing supply in areas like Austin, TX. Moreover, places with a limited supply of land could expand the stock of housing by building taller structures. Even in places with little vacant land like Manhattan, many residential buildings are shorter than can be explained by the cost of an additional story (Glaeser et al. 2005b).

Government Regulation

The third factor influencing the elasticity of housing supply involves the permission to build. Even when the costs of materials, labour and land are low enough to generate an incentive to expand the housing stock, government restrictions often prevent developers from building as many residential units as they would like. Local governments have regulated the placement of residential structures ever since the 1920s, when zoning laws began to separate residential land uses from commercial and industrial development. While these regulations altered the geographic distribution of residential structures within cities, initially they did not have a notable impact on the aggregate supply of urban housing (Fischel 2004). It was not until the 1970s that municipalities began to enact growth controls and other exclusionary zoning practices designed to limit the absolute number of residential units in their jurisdiction. The popularity of these types of regulations has grown over the past several decades, and local governments now employ a wide range of regulatory practices including height and lot size restrictions, development moratoria, historic preservation rules and urban growth boundaries.

In contrast to these restrictive regulations, some government policies attempt to increase the supply of housing by providing tax incentives or subsidies to build units that will be affordable to low-income households. However, these policies do not have a notable impact on the aggregate stock of housing, as they mostly substitute for unsubsidized housing units (Malpezzi and Vandell 2002). Federally owned housing appears to be less substitutable for private units, but there has been virtually no new construction of public housing units since the early 1980s (Green and Malpezzi 2003).

Because land-use regulations are enacted by local governments and are frequently customized to meet the needs of individual neighbourhoods, these laws vary substantially across locations in both form and severity. This heterogeneity makes the degree of regulation difficult to classify in a manner that lends itself well to systematic empirical analysis. Despite this complexity, most empirical research has found a strong correlation of land-use regulation with higher house prices and less residential construction (Malpezzi 1996; Mayer and Somerville 2000; Saks, Saks 2005). Thus, these regulations appear to reduce the elasticity of housing supply in the areas in which they are enacted.

As the number of municipalities with restrictive residential land-use policies has expanded, researchers have become progressively more interested in trying to understand the political economy of these regulations. Recent decades contrast sharply with the regulatory environment during the 1950s and 1960s, when builders were generally able to influence the decisions of local zoning boards (Molotch 1976). Since that time, homeowners have become more successful at restricting residential construction in their neighbourhoods. The incentive of homeowners to constrain development has been linked to several motivations including the reduction of congestion costs, the preservation of local amenities (Hilber and Robert-Nicoud 2006), insurance against shocks to household wealth (Ortalo-Magne and Prat 2007), the reduction of free-riding on the provision of public goods (Fischel 2001), and the growing likelihood that homeowners work in a different jurisdiction from their place of residence (Fischel 2001). In addition to changes in homeowners' incentive to limit new construction, the rise of regulation may also be a function of their improved ability to influence the political process (Glaeser et al. 2005a).

While theories explaining the existence of supply restrictions have multiplied, empirical evidence on the determinants of zoning remains thin. Richer towns with more educated populations exhibit a higher propensity to restrict residential development (Evenson and Wheaton 2003; Glaeser et al. 2005a), and cities are more likely to enact land-use regulations when the policies of neighbouring municipalities are also restrictive (Brueckner 1998). However, these studies are based on cross-sectional evidence, making it difficult to distinguish causal mechanisms from location-specific characteristics and geographic differences in housing demand.

Decreases in Supply

The durability of housing structures means that the elasticity of housing supply is asymmetric in response to increases versus decreases in demand. Because housing depreciates slowly, the housing stock does not contract immediately in response to a decline in housing demand. Instead, places experiencing persistent declines in housing demand have low house prices relative to construction costs. The availability of cheap housing encourages households to remain in declining cities rather than moving to a location with growing labour demand. Thus, urban decline is slow and highly persistent (Glaeser and Gyourko 2005). The durability of housing may also influence urban growth through its impact on local land-use planning decisions (Turnbull 2006).

Broader Consequences of the Elasticity of Housing Supply

The effects of the housing supply extend far beyond changes in the relative distribution of house prices and city sizes across the United States. For example, by restricting the number of households in a location, the housing supply can limit the supply of workers, altering the dynamics of local wage and employment growth (Case 1991; Saks 2005). Aggregate economic activity may also be reduced as workers are prevented from living in the location where they would be most productive. The housing supply also affects the distribution of income across and within cities. By altering relative house-price differentials, supply restrictions will cause high-income households to sort into metropolitan areas with highly valued amenities (Gyourko et al. 2006). Moreover, the composition of the population within metropolitan areas will also depend on the elasticity of housing supply, as demographic groups with a higher propensity to move relocate in response to rising house prices.

While this article has focused on the United States, the underlying forces that shape housing supply conditions are similar around the world. Housing investment as a share of GDP in the United States has been around the median of other OECD countries since the late 1990s. In some countries, construction activity is lower than in the United States due to a greater scarcity of land and more restrictive land-use regulations. By contrast, some other developed countries have higher rates of housing investment due to a more active government role in subsidizing residential construction (Ball 2003). Given the widespread reductions in the elasticity of housing supply in many parts of the United States during the past few decades, further investigations into the determinants and implications of housing supply conditions promise to be an important direction of future research in both urban economics and macroeconomics.

See Also

- ▶ Housing Policy in the United States
- ▶ Low-Income Housing Policy
- ▶ Residential Real Estate and Finance
- ▶ Urban Economics
- ▶ Urban Growth
- ▶ Urban Housing Demand

Bibliography

Ball, M. 2003. Markets and the structure of the housebuilding industry: An international perspective. *Urban Studies* 40: 897–916.

Brueckner, J.K. 1998. Testing for strategic interaction among local governments: The case of growth controls. *Journal of Urban Economics* 44: 438–67.

Case, K.E. 1991. The real estate cycle and the economy: Consequences of the Massachusetts boom of 1984–87. *New England Economic Review* (September): 37–46.

Davis, M., and J. Heathcote. 2005. *The price and quantity of residential land in the United States*. Discussion Paper No. 5333, Center for Economic Policy Research.

Evenson, B., and W.C. Wheaton. 2003. Local variation in land use restrictions. *Brookings-Wharton Papers on Urban Affairs* 2003: 221–50.

Fischel, W. 2001. *The homevoter hypothesis: How home values influence local government taxation, school finance, and land use policies*. Cambridge, MA: Harvard University Press.

Fischel, W. 2004. An economic history of zoning and a cure for its exclusionary effects. *Urban Studies* 41: 317–40.

Glaeser, E.L., and J. Gyourko. 2005. Urban decline and durable housing. *Journal of Political Economy* 113: 345–75.

Glaeser, E.L., J. Gyourko, and R.E. Saks. 2005a. Why have house prices gone up? *American Economic Review Papers and Proceedings* 95: 329–33.

Glaeser, E.L., J. Gyourko, and R.E. Saks. 2005b. Why is Manhattan so expensive? Regulation and the rise in house prices. *Journal of Law and Economics* 48: 331–69.

Green, R.K., and S. Malpezzi. 2003. *A primer on U.S. Housing markets and housing policy*. Washington, DC: Urban Institute Press.

Gyourko, J., and A. Saiz. 2006. Construction costs and the supply of housing structure. *Journal of Regional Science* 46: 661–80.

Gyourko, J.,C. Mayer, and T. Sinai. 2006. *Superstar cities*. Working Paper No. 12355. Cambridge, MA: NBER.

Hilber, C., and F. Robert-Nicoud. 2006. *Owners of developed land versus owners of undeveloped land: Why land use is more constrained in the bay area than in Pittsburgh*. Discussion Paper No. 870, Centre for Economic Policy Research.

Malpezzi, S. 1996. Housing prices, externalities, and regulation in U.S. metropolitan areas. *Journal of Housing Research* 7: 209–41.

Malpezzi, S., and K. Vandell. 2002. Does the low-income housing tax credit increase the supply of housing? *Journal of Housing Economics* 11: 360–80.

Mayer, C.J., and C.T. Somerville. 2000. Land use regulation and new construction. *Regional Science and Urban Economics* 30: 639–62.

Molotch, H. 1976. The city as a growth machine. *American Journal of Sociology* 82: 309–30.

Ortalo-Magne, F., and A. Prat. 2007. *The political economy of housing supply: Homeowners, workers and voters*. Discussion Paper No. TE/2007/514, Suntory- Toyota International Centers for Economics and Related Disciplines.

Rose, L. 1989. Urban land supply: Natural and contrived restrictions. *Journal of Urban Economics* 25: 325–45.

Saks, R.E. 2005. *Job creation and housing construction: Constraints on metropolitan area employment growth*. Finance and Economics Discussion Series 49, Board of Governors of the Federal Reserve System (U.S.)

Somerville, C.T. 1999a. The industrial organization of housing supply: Market activity, land supply and the size of homebuilder firms. *Real Estate Economics* 27: 669–94.

Somerville, C.T. 1999b. Residential construction costs and the supply of new housing: Endogeneity and bias in construction cost indexes. *Journal of Real Estate Finance and Economics* 18: 43–62.

Turnbull, G.K. 2006. The investment incentive effects of land use regulations. *Journal of Real Estate Finance* 31: 357–95.

Housing Wealth

Pedro Silos

Abstract

This entry describes housing wealth and the role it plays in household finance.

Keywords

Borrowing constraint; Households portfolios; Real estate finance; Wealth effect

JEL Classifications

E21; G11; G12

Housing wealth is the combination of land and structures for the purpose of providing shelter or housing services. Housing plays a dual role as a durable good that provides shelter and as an asset that complements other sources of wealth in the portfolios of households.

In principle, if housing was perfectly divisible, no transaction costs were associated with it, capital markets were perfect, and financing frictions were absent, individuals would optimally choose housing services (through a rental market) independent of the amount of housing wealth in their portfolios. In reality, most households, at least in the USA, choose to enjoy housing services through ownership of their residences. The result is that housing becomes a major component of the portfolios of households in many developed countries. At the turn of the 21st century, the median household in the USA – according to income – tied about twothirds of its wealth to residential real estate. In aggregate, the share of housing is slightly smaller than that of other assets, but, due to the extreme concentration of non-housing wealth, wealth lies largely in housing for a large percentage of the population. Its weight in households' portfolios has a clear life-cycle pattern: young homeowners leverage themselves to purchase homes that result in large housing-tonet-worth ratios. As people age, their earnings increase, resulting in a larger accumulation of financial assets and decreasing the housing-to-wealth ratio. At retirement, financial assets are depleted at a faster rate than housing wealth is decreased, resulting in a minor increase in the ratio.

The interpretation of housing as a bundle of land and structures allows a calculation of changes in housing prices into changes in the price of land and changes in the price of structures. Davis and Heathcote (2007) report that changes in the price of residential land account for most of the low and business-cycle frequency changes in house prices and that the price of residential structures moves quite differently from that of land. Land prices increased significantly in real terms over the second half of the 20th century, as a result causing an increase in the importance of housing in the aggregate wealth portfolio according to Skinner (1994). Much economics research has focused on understanding the interplay between changes in housing wealth and the consumption, savings and portfolio decisions of households. For example, two reasons exist to explain why housing wealth plays a relevant role in decisions made by households. First, existing frictions in housing markets – for instance, in financing a house or in search of buying and selling properties – cause most households to make optimal housing consumption and investment decisions jointly through the purchase of a single property. Imposing this constraint on optimal portfolios has problematic implications for the mix of financial assets held by households. Second, changes in housing wealth have profound effects on households' consumption and savings decisions. These effects are the result of a direct change in perceived wealth or the result of changing the tightness of borrowing constraints if housing wealth is used as collateral.

As an asset, residential real estate is risky, with fairly volatile prices. Using data from the Panel Study of Income Dynamics (PSID) on self-reported property values and accounting for taxation and maintenance costs, Flavin and Yamashita (2002) calculated statistical returns to homeownership in the USA. The mean return over their sample period (1968–1992) was 6.6% per year (by comparison, the mean return for stocks was 8.2%), the standard deviation was 14% (relative to 24% for stocks), and housing was essentially uncorrelated with either stocks or bonds. As a result, from a portfolio perspective, it is optimal to hold some residential real estate because it helps diversify the risk present in financial assets. However, the solution to this portfolio problem is complicated for households that, due to frictions, choose one single house, which determines the consumption of housing services and the quantity of housing in the portfolio. Flavin and Yamashita (2002) show that within a mean-variance frontier framework, housing and its financing change the risk and return trade-off that households face. Households with positive housing-to-wealth ratios see a drop in the weight of riskless assets relative to risky bonds and stocks. In fact, the non-negativity constraint in the riskless asset positions is binding for households with large housing-to-wealth ratios. This fact helps explain cross-sectional data on the composition of financial assets over the life cycle of individuals. Cocco (2005) studies portfolio choice in the presence of housing, also finding important implications for the weight of financial assets in households' portfolios and explaining the observed positive correlation between stockholding and leverage.

The literature has also focused on the response of household consumption to changes in housing prices. The increase in consumption observed during periods of rising housing prices can be the consequence of a larger wealth effect, a relaxation of borrowing constraints for constrained homeowners, or simply because both variables depend on unobserved rises in expected income. Campbell and Cocco (2007) use UK household level data to estimate how consumption responds to changes in house prices. They find that the elasticity of consumption to house price for individuals changes by age. The consumption of young households does not react to house price changes, while the elasticity for older homeowners is positive and significantly different from zero. Li and Yao (2007) also find similar distributional effects of house price changes for different age groups using a structural model calibrated to US data. They find that in the face of rising house prices, the consumption of young individuals should respond negatively because it takes more savings to achieve a required down payment. As retirees downgrade the size of their housing holdings (to some extent), the positive capital gains allow them to increase their consumption. Middle-aged individuals see their welfare roughly unchanged.

Campbell and Cocco also find evidence that increases in house prices relax borrowing constraints. The introduction of home equity lines of credit (HELOC) has facilitated the use of housing wealth for smoothing consumption over the life cycle. Introduced at the beginning of the 1980s, HELOCs are loans that use equity holdings in real estate as collateral, and they have become increasingly important, particularly in periods of rising house prices. These instruments provide flexibility in transforming illiquid real estate wealth into liquid assets. As housing wealth is widely used as collateral, changing house prices can affect the ability of households to share risk in the face of idiosyncratic labour market risk. The amount of housing wealth relative to non-housing wealth in an economy becomes a candidate to explain some empirical failures of equilibrium asset pricing models. Lustig and Van Nieuwerburgh (2005) use this explanation to construct an economy in which a decrease in the amount of collateralisable housing wealth leaves households more exposed to labour market risk as borrowing constraints are more likely to be binding. The model helps explain why some empirical regularities are inconsistent with the standard consumption-based, asset-pricing model. For instance, the ratio of housing wealth to non-housing wealth helps predict stock returns at low frequencies.

See Also

- ▶ Collective Models of the Household
- ▶ Household Portfolios
- ▶ Housing Supply

Bibliography

Campbell, J.Y., and J.F. Cocco. 2007. How do house prices affect consumption? Evidence from micro data. *Journal of Monetary Economics* 54(3): 591–621.

Cocco, J. 2005. Portfolio choice in the presence of housing. *Review of Financial Studies* 18: 535–567.

Davis, M., and J. Heathcote. 2007. The price and quantity of residential land in the United States. *Journal of Monetary Economics* 54(8): 2595–2620.

Flavin, M., and T. Yamashita. 2002. Owner-occupied housing and the composition of the household portfolio. *American Economic Review* 92(1): 345–362.

Li, W., and R. Yao. 2007. The life-cycle effect of house price changes. *Journal of Money, Credit, and Banking* 39(6): 1375–1409.

Lustig, H., and S. Van Nieuwerburgh. 2005. Housing collateral, consumption insurance, and risk premia: An empirical perspective. *Journal of Finance* 60(3): 1167–1219.

Skinner, J. 1994. Housing and saving in the United States. In *Housing markets in the U.S. and Japan*, ed. Y. Noguchi and J. Poterba, 191–214. Cambridge, MA: National Bureau of Economic Research Inc.

Human Capital

Sherwin Rosen

JEL Classifications
J24

Human capital refers to the productive capacities of human beings as income producing agents in the economy. The concept is an ancient one, but the use of the term in professional discourse has gained currency only in the past twenty-five years. During that period much progress has been made in extending the principles of capital theory to human agents of production. Capital is a stock which has value as a source of current and future flows of output and income. Human capital is the stock of skills and productive knowledge embodied in people. The yield or return on human capital investments lies in enhancing a person's skills and earning power, and in increasing the efficiency of economic decision-making both within and without the market economy. This account sketches the main ideas, and the bibliography is necessarily restrictive. For additional detail and alternative interpretations, the reader should consult the surveys by Blaug, Rosen, Sahota and Willis, which also present complete bibliographies.

Differences in form between human and nonhuman capital are of less import for analysis than are differences in the nature of property rights between them. Ownership of human capital in a free society is restricted to the person in whom it is embodied. By and large a person cannot, even voluntarily, sell a legally binding claim on future earning power. For this reason the exchange of human capital services is best analysed as a rental market transaction. Quantitative analysis is restricted to the income and output flows that result from human capital investments: wage payments and earnings flows are viewed as the equivalent of rentals of human capital value, because a person cannot sell asset claims in himself. Even the longterm commitments found in enduring employment relationships are best viewed as a sequence of short-term, renewable rental contracts. By contrast, the legal system places many fewer restrictions on the sale and voluntary transfer of title to nonhuman capital. In fact, substantial activity on non-human capital asset markets is a hallmark of an enterprise system of organization.

Flexibility must be maintained, however, in these distinctions, which are not always hard and fast. The institution of slavery was the primary example of a transferable property right in human capital. To be sure, the involuntary elements of slavery are essential, but even voluntary systems have not been unknown. Similarly, indentured servitude was an example of a legally enforceable long-term contractual claim on the human capital services of others. And in many societies today there are severe legal restrictions on transfer of title to non-human capital: the chief example is collective and state ownership of non-human capital in planned economies.

Background

Classical economics maintained a tripartite distinction among the factors of production, Land, Labour and Capital; whereas modern economics is much less rigid in these divisions. Viewed from the perspective of supply, factors of production, whatever their form, can be increased and improved at some cost. To the extent that these improvements involve weighing future benefits against current costs, the principles of capital theory are applicable.

William Petty, the early actuary and national income accountant, is generally credited with the first serious application of the concept of human capital, when in 1676 he compared the loss of armaments, machinery and other instruments of warfare with the loss of human life. Elements of such comparisons survive to the present day. However, Adam Smith set the subject on its main course. *The Wealth of Nations* identified the improvement of workers' skills as a fundamental source of economic progress and increasing economic welfare. It also contained the first demonstration of how investments in human capital and labour market skills affect personal incomes and the structure of wages. Alfred Marshall stressed the long-term nature of human capital investments, and the role of the family in undertaking them. He also pointed out that nonmonetary considerations would play a unique role in these decisions because of the dual nature of workers as factors of production and as consumers of their work environments. The distinguished actuary and scientist Alfred Lotka provided the first quantitative application of human capital in collaboration with Dublin, calculating the present value of a person's earnings to serve as guidelines for the rational purchase of life insurance. J.R. Walsh made the first cost imputation of human capital value. Frank Knight focused upon the role of improvements in society's stock of productive knowledge in overcoming the law of diminishing returns in a growing economy.

These early contributions stand as landmarks. However, the impetus for rapid progress in this area came from the quantitative revolution in economics after World War II, when extensive data sources revealed certain systematic regularities. The first of these stems from economists' interest in understanding the nature and sources of economic growth and development in the 1950s and 1960s. Detailed calculations by national income accountants showed that conventional aggregate output measures grow at a more rapid pace than aggregate measures of factor inputs. A fundamental conservation law in economics would be violated unless the unexplained 'residual' was identified with (unexplained) technical change. Research associated with T.W. Schultz and Edward Denison attributed much of the measured residual to improvements in factor inputs. Schultz adopted an all-inclusive concept of human capital. At its heart lay secular improvements in workers' skills based on education, training and literacy; but he also pointed to sources of progress in improved health and longevity, the reduction in child mortality and greater resources devoted to children in the home, and the capacity of a more educated population to make more intelligent and efficient economic calculations. John Kendrick systematically pursued the empirical implications of these ideas and demonstrated that the rate of return on these inclusive human capital investments is of comparable magnitude to yields on non-human capital. This line of research as a whole proves that an investment framework is of substantial practical value in accounting for many of the sources of secular economic growth.

Another parallel strand of development arose from professional interest in the nature and determinants of the personal distribution of income and earnings. This problem was propelled, in addition, by substantial public interest in the problem of poverty and prospects for redistributing resources to the poor. Empirical bases for this inquiry were, and continue to be, supported by extensive personal survey instruments (such as Census and allied records) that have become widely available in the post-war period. Much of this work has focused on the role of education and training as important determinants of personal wealth and income. Herman Miller's updating and elaboration of Dublin and Lotka's calculation found a strong and systematic relationship between education and personal economic success, a finding that has been replicated many times in virtually every country where data are available to make the calculations.

The fundamental conceptual framework of analysis for virtually all subsequent work in this area was provided by Gary Becker, who not only organized the emerging empirical observations but also provided a systematic method for seeking new results and implications of the theory. Practically every idea in his book has been pursued at length in the research of the past two decades. Following Schultz's lead, Becker organized his theoretical development around the rate of return on investment, as calculated by comparing the earnings streams in discounted present value on alternative courses of actions. Rational agents pursue investments up to the point where the marginal rate of return equals the opportunity cost of funds. Hence, conditional on the sources of financing investments through the market and family resources, there is a tendency for rates of return to be equated at the margin. This theory of *supply* of human capital implies empirically refutable restrictions on intertemporal and interpersonal differences in the patterns of earnings and other aspects of productivity. In focusing on the development of a person's skills and earning capacity over the life cycle, human capital theory has evolved as a theory of 'permanent income' and wealth.

Becker also made a distinction between human capital that is specific to its current employment in a firm, and that which has more general value over a broader set of employments. The concept of firm-specific capital is closely allied with organizational capital, a person's contribution to a specific organization, the value of which is lost and must be reproduced by costly investment when the employment relationship is terminated. General human capital represents skills that are not specifically tied to a single firm and whose employment can be transferred from one firm to another without significant loss of value. This distinction has proved valuable for analysing the determinants of turnover and firm-worker attachments and its ramifications are still being pursued. For example, the concept of firm-specific capital underlies the transactions cost basis for recent research on labour market and other contracts.

The Rate of Return

The connection between the rate of return on investment in human capital and observable earnings is illustrated by Smith's discussion of the relative earnings of physicians and other professional workers. A person who contemplates entering one of these fields must look forward to a long period of training and costly personal investment before any income is forthcoming. Furthermore, the long training period cuts into the period of actual practice and reduces the period of positive earnings. Consequently earnings must *compensate* for the cost and effort required to practice the trade: if they did not, fewer people would find it attractive to enter.

The compensatory nature of earnings on prior investments, equivalent to a rate of return, is the fundamental insight of human capital theory. First, it points to the opportunities foregone by an action as a fundamental cost of undertaking it. Thus the direct tuition and other costs of education are only one component of the true cost. The fact that the person defers entering the market and gives up a current source of earnings is also properly counted as a cost. Second, the focus on the intertemporal and life-cycle nature of these decisions leads to a much different concept of income and inequality than simply examining current earnings. Human capital theory suggests that the distributions of *lifetime earnings* and human capital *wealth* are the keys to analysing the distribution of economic welfare, because earnings are the result of prior investments.

Two methods are widely used to calculate the return on human capital investments. Consider one alternative, call it the null alternative, which yields an earnings flow of $x_0(t)$. Consider another alternative, call it the investment alternative, which yields an earnings flow of $x_1(t)$. For example, in the leading case $x_0(t)$ is the expected flow of earnings in year t if one terminates education after high school graduation and $x_1(t)$ is the earnings that can be expected if one continues on to college. The time index t commences as of high school graduation, so $x_1(t)$ will typically show a phase (during the period of college attendance) of much smaller values than does $x_0(t)$. However, in later life $x_I(t)$ is generally larger than $x_0(t)$. This is precisely the investment content of the decision to continue school: there is a current cost in terms of income foregone, but a deferred benefit in terms of greater earnings prospects in the future. Write the difference $z(t) = x_1(t) - x_0(t)$. Then $z(t)$ shows a systematic pattern of negative values when t is small and positive values when t is large; $z(t)$ is increasing from negative to positive in between. Observed earnings in the two choices allows calculation of the internal rate of return, defined as the rate of interest which equates the present discounted value of the two earnings streams. If i is the internal rate, then $\sum z(t)/(1 + t)' = 0$.

Of course, it is not possible to observe earnings in the path not taken. A person either stops school or continues on to the next level. In practice, the calculation is made by using observed average earnings of college graduates at different ages as an estimate of $x_1(t)$ and using the observed average earnings of high school graduates as an estimate of $x_0(t)$. The typical calculation produces an estimate of i in the neighbourhood of ten per cent, comparable to the rate of return on investment in physical capital. Hanoch presents the most complete treatment of this problem. Remarkably, rates of return on education in the vicinity of ten per cent are found in a wide variety of countries and economic institutions.

Another method of calculation, first presented by Jacob Mincer, brings out the economic aspects of these estimates more clearly. Suppose a person contemplates a level income in amount $y(s)$ over the life work-life cycle if s years of schooling are undertaken. If schooling is productive we must have that $y'(s) = \mathrm{d}y/\mathrm{d}s$ is positive, that is, anticipated earnings must be increasing in years of schooling. The present discounted value of wealth associated with some choice s, from the point of view of the present time, is simply

$$W(s) = y(s)\int_s^n \mathrm{e}^{-rt}\mathrm{d}t,$$

where the index of integration runs from *s*, the time the person completes school and enters the market, to *n*, the time the person retires. Since *n* is large, we may take the approximation

$$W(s) = y(s)\int_s^\infty \mathrm{e}^{-rt}\mathrm{d}t = y(s)\mathrm{e}^{-rs}/r.$$

Assume that the schooling decision is made to maximize human capital wealth *W(s)*. Then differentiating with respect to *s*, the first order condition is $[y'(s) - ry(s)]\ \mathrm{e}^{-rs} = 0$, or $y'(s)/y(s) = r \cdot y'/y$ is nothing other than the marginal internal rate of return on investment in schooling, so schooling is chosen such that its marginal internal rate equals the rate of interest. This rule, similar to the economic problem of when to cut a tree or uncork the wine, is one that maximizes lifetime consumption prospects for the person.

Now extend this argument to many people. In an economy with many similar individuals making schooling choices, all would choose the same value of *s*, satisfying $d \log y(s)/d \log s = r$. Since there would be no differences in schooling choices among them occupations and jobs that required either more or less education would go unfilled, and the labour market would not clear. Yet, if we observe that in the market equilibrium different people choose different amounts of schooling, with some actually choosing more education and some actually choosing less, then the market earnings on jobs with different schooling requirements must adjust so that the marginal condition is an identity for all possible values of *s*. That is, people must be indifferent as to how much education they choose. Viewing the marginal condition as a differential equation in *y* and *s* and integrating yields the *restriction* $y(s) = y_0\mathrm{e}^{rs}$, where y_0 is the earnings of a person without any schooling. Substituting this back into the definition of *W*(*s*), we have

$$W(s) \quad = \quad y_0\mathrm{e}^{rs}\int_s^\infty \mathrm{e}^{-rt}\mathrm{d}t \quad = \quad y_0/r$$

is *independent of s.* Writing $W(s) = W$ to reflect this fact, we have $y(s) = (rW)\,e^{rs}$, and $\log y(s) = \log (rW) + rs$. Think of this last expression as a regression equation. Then after adjusting the income data for age and experience, a regression of the log of income on years of school yields an estimate of the marginal internal rate of return to education (*r*) as the regression coefficient on schooling. The constant term in the regression estimates 'earning capacity' log (*rW*).

The economic logic underlying this development clearly shows the compensatory nature of the returns to schooling and its relationship to the theory of supply. The equilibrium earnings–schooling function is an equalizing difference on the foregone opportunity and other costs of attending school. If people are alike, earnings must rise with schooling to cover the direct and interest costs. Otherwise no one would be inclined to undertake these investments. Notice that in this example, income differences are equalized on cost at every point and that the human wealth (*W*) is the same for all. Thus there is inequality of earnings, but complete equality of human capital wealth or life cycle earnings. Restricting attention to inequality in the observed distribution of earnings would give a highly misleading indication of inequality in the true distribution of economic welfare in this case.

This simple decision problem provides a convenient and powerful conceptual framework around which much of the research in this area has been organized. The value of this framework was first demonstrated by Becker, who expanded it to include interpersonal differences in abilities and talents and in family circumstances. Interpersonal differences in the rate of interest *r*, are identified with financial constraints on human capital investments associated with family background and related factors. A person confronting a higher rate of interest would be unable to finance human capital investments on favourable terms and would therefore rationally choose to invest less than a person who was able to borrow at lower rates. Similarly, there may be interpersonal differences in talents among people. Some may be more skilled in learning, which makes schooling effectively cheaper for them, or they may have natural talents which either complement or substitute for schooling in producing earning capacity.

Considerations such as these lead to an identification problem in the schooling–earnings relationship observed *across* different individuals (see Rosen 1977, for elaboration; also Willis). To begin, let us isolate the effects of family background and financial constraints by restricting attention to a subset of individuals with the same natural talents and abilities. Then differences in school choices within this group would be provoked by corresponding differences in family backgrounds and financial constraints. The reason for this goes back to the institutional feature of human capital assets noted above, that a person cannot sell an asset claim to future earning power. Thus human capital does not serve as collateral for investments in anywhere near the same way as title to physical capital does for non-human investment. A house, for example, serves as collateral for a mortgage. If the purchaser defaults on the mortgage then the creditor gains title to the house, which can then be sold to settle the debt. Non-transferable titles to human capital make this kind of arrangement impossible for personal investments. Relaxing these kinds of constraints is, of course, the fundamental economic logic behind the public provision of education in most countries throughout the world. But since direct tuition and related costs are only a part of the true costs of schooling, the importance of foregone earnings costs suggests that financial constraints would still remain a factor in educational decision-making. As Marshall noted, the social and economic status of the family play an important role in educational choices.

From the point of view of econometric estimation, observing a subset of the population where abilities are roughly constant, but where financial constraints dictate different schooling choices allows identification of the schooling–earnings relationship for that ability level. This in turn enables the analyst to calculate the social rate of return on investment, and to determine empirically the effect on personal and aggregate wealth of social policies that relax the financial constraints. Earnings of otherwise similar people who were less constrained serve as excellent estimates of the true earnings prospects for more constrained individuals.

Extensive empirical investigation of the connection between schooling, earnings, and family background shows a very strong and systematic relationship between parents' socioeconomic status and background and the school quality and completion levels of their children (e.g. Griliches 1970, 1977). This is prima facie evidence of financial constraints on educational choices, though it does not rule out other routes by which family background affects a person's economic success, such as complementary investments in the home in child care and quality. These studies also indicate a direct connection between family background and earnings given the schooling choices of children. The causal link between these direct effects of family background and earnings remain to be established. It could reflect common but unobserved variance components across generations within families, such as unobserved ability; and also unmeasured factors, such as school quality and the quality of parental inputs, that are correlated with family background. Whatever their source, these direct linkages are numerically small compared with the effect of schooling itself on earnings. Most of the effect of family background on economic success works through its effects on the educational decisions of children and through that to economic success as measured by income and earnings. The direct effect on income, while persistent and significant, is quantitatively small.

Some Applications

Perhaps the main policy area where these ideas on financial constraints are important is in public provision of training and 'manpower' development programmes for the poor. The logic of these policies rests on the proposition that a person's income in a market economy reflects the quantity of resources that the person controls and the value of these resources. People who are permanently poor have less skills and also less valuable skills then the non-poor. So an attractive policy to help eliminate poverty is to give them more and better resources through education and training. The rate of return has been widely used for programme evaluation. For if the social return to investment in subsidized training is less than the rate of return on other forms of social investment, then programmes emphasizing direct monetary and other transfers to the poor are better bets for society overall than devoting resources to skill enhancement. There now exists a voluminous literature on manpower programme evaluation along these lines, largely stemming out of the social programmes that were instituted in the 1960s and 1970s in the United States. The evidence is mixed. While many examples of successful programmes can be found, the prevailing assessment among experts is that the average programme has not been clearly successful (Ashenfelter 1978). This empirically based conclusion suggests that the underlying causes of poverty are more complicated than simple family constraints on resources which thwart human capital investments. Lack of motivation, discrimination, ability, low quality prior education and insufficient investments in children in the home, as well as constraints on financing are among many of the possibilities that present themselves as causal factors in reducing personal investments in human capital.

The changing role of women in the workplace and in the home has refocused current professional interest on the role of families in determining economic success of children. While these intergenerational connections between the wealth and economic status of parents and their children have long been recognized as a key element in the question of poverty and the size distribution of income, these aspects have only been linked to human capital theory in very recent years. Again, the impetus for this interest lies in the empirical findings summarized above, and also in some that have come from unexpected quarters, namely the economic success of immigrants and their children.

Recent work by Barry Chiswick (1978) has established a systematic empirical pattern for many immigrant groups into the United States. Chiswick finds that members of the first generation of immigrants earn less than comparable native born citizens in the first two decades of their life in the US. At that point their incomes reach parity with native born citizens and beyond it actually surpass the incomes of the native population. More remarkably, the sons of these immigrants – the members of the second generation – earn incomes which exceed those of the sons of native born workers. However, by the third generation there is parity, and the effects of foreign-born status wash out. While certain aspects of Chiswick's findings remain controversial and are being studied at length, they support the 'melting pot' view of economic life in the US. There is obviously substantial interest and importance in examining similar phenomena in other countries.

The chief theoretical work in the intergenerational transmission of wealth and economic status through families is contained in the research of Becker and Tomes. This work directly addresses intergenerational linkages through preferences and attitudes of parents toward their children, through natural hereditary transfers of ability and through discretionary transfers of resources through the generations. This work is the most complete theoretical description of the intergeneration distribution of wealth available so far. Inheritability of abilities is known from statistical theory to imply a regression-toward-the-mean phenomenon. Thus the fortunes of one generation are not only linked by direct transfers of non-human wealth and human capital investments, but also by inherited traits. These two forces interact in the intergenerational transmission mechanism. The economic fortunes of generations are more closely linked the greater the degree of inheritability of ability and the greater the propensity of parents to invest in their children's human capital. The effects of good fortune in one generation spills over to the next through the transfer mechanism. Interestingly, it may spill over to several subsequent generations. Thus regression toward the mean may occur only after several generations rather than after only one.

When borrowing constraints are imposed on this structure even more persistence is implied because low income families do not have sufficient resources to invest in their children, whose incomes as parents are smaller than they would otherwise be. These issues are important for understanding social and economic mobility, and only recently have data become available to study them empirically. In the end this may be one of the most important developments in human capital theory.

Ability Bias

The other major area where considerable research progress has been made is the role of ability in determining economic success. In terms of the decision model above, interpersonal differences in ability shift the earnings–schooling relationship. More able persons earn more at a given level of schooling than the less able, so the observed income-schooling relationship does not necessarily represent the returns available to a given person. Thus consider a group of individuals who have the same financial resources (the same value of r in the term discussion above). If ability is complementary with schooling then the rate of return to schooling will be larger for the more able and they will choose to invest more. A person observed choosing less education rationally does so because the personal return is relatively small under these circumstances. Comparing the earnings of persons who choose less education with those of persons choosing more education leads to a biased assessment of the returns due to differences in their abilities. This ‘ability bias’ issue has been examined in much detail.

The basic issue was originally posed by Becker, using the discounted earning stream comparisons presented above. If $x_0(t)$ is the earnings stream of people who stop school after high school completion and $x_1(t)$ is the earning stream of those who continue to college, then $x_1(t)$ is likely to be a biased estimate of the earnings prospects of high school graduates had they continued on to college. In so far as their average ability is lower than college graduates, their earnings had they chosen to continue on to college are likely to be smaller than $x_1(t)$. Similarly, the higher average abilities of college going persons makes it probable that $x_0(t)$ is a downward biased measure of what they would have earned had they stopped their education after high school graduation. Thus comparing $x_1(t)$ with $x_0(t)$ yields an upward biased estimate of the rate of return to education for either group.

In order to correct this bias it is necessary to purge the earnings data of the direct effects of ability. Several methods have been proposed, and most find that the effect of ability biases in rate of return calculations is positive but relatively small (Griliches). The fundamental reason for this is due to a finding of Welch, that while the direct effect of measured ability on earnings is positive (given schooling), its numerical effect is quite small. Even a person whose measured ability is one standard deviation above the mean receives, on average, an income that is only a few percentage points above average.

Most of the research in this area has concentrated on indexes of ability associated with IQ and other measures meant to predict school performance. However, predictors of school performance and grades are not necessarily good predictors of economic success. The most sophisticated studies employ factor analytic statistical models, in which measured abilities embodied in IQ scores and the like serve only as indicators of underlying and unobserved ‘true’ abilities. These studies show that ‘raw’ rate of return estimates unadjusted for ability differences overstate ‘true’ rate of return calculations by only a few percentage points. The rate of return to school remains substantial, and of comparable magnitude to that on other forms of investment even after ability adjustments have been made.

Most of this ability-bias research assumes that ability can be captured statistically as a single factor (in the statistical sense). However, some recent work is based on a multiple-factor view of ability in which there are different dimensions and components (Willis and Rosen 1978). This multi-factor framework is familiar from the theory of comparative advantage in economics.

A unidimensional specification of ability only allows for absolute advantage, where a person who is more able in one thing is necessarily more able in everything else. By contrast, a comparative advantage specification allows for both absolute and relative advantages. A person may be very talented in all things (absolute advantage), but may also be relatively more talented in some things than others. Furthermore, absolute advantage may not be so important. A great musician is not necessarily adept at non-musical activities such as accounting; and the typical accountant may well have no more than the average musical ability in the entire population. An extension of the model above shows that people would naturally select themselves into those occupations and educational categories that exploit their comparative advantage. Thus those who choose to specialize their human capital investments in musical activities would be likely to have more natural talent for it than the population at large. Similarly, those who learn the plumbing trade would be likely to have more mechanical ability than those who make some other choice. These types of selection problems gain research interest because educational and occupational choices are closely linked. While much important work remains to be done in this area, available evidence is at least consistent with the existence of comparative advantage and occupational selection. If so, the overall ability bias in simple rate of return calculations is likely to be relatively small.

The question of ability bias and selection comes up in a quite different manner in the literature on educational screening and signalling (Spence 1973). In its most extreme form, the signalling literature maintains the hypothesis that education has no direct effect on improving a person's skills, but rather serves as an informational device for identifying more and less talented people. This model rests on a unidimensional view of ability and also on the suppositions that direct observation of a person's ability and productivity is very costly and that a person knows much more about his own abilities than other persons do. In these circumstances, education serves as a signal of ability if the more able can purchase the educational signal on more favourable terms than the less able. For then education and ability are highly correlated, and the higher income earned by those with more schooling is supported in equilibrium by their higher ability-productivity.

Several points must be made in this connection. The first is, that taken on its own terms, the signalling and human capital models have very similar implications for the rational choice of schooling. In fact they appear to be econometrically indistinguishable on the basis of income and schooling data alone. The chief difference is a normative one, that schooling has a little social value when it serves as a signal, and has much social value when it produces real human capital. Second, the data reveal considerable 'noise' in the schooling–earnings relationship. An investigator does very well when a third of the total variance in earnings can be 'explained' in the analysis of variance sense by observable personal factors such as education, experience, ability measures, family background and other factors. The schooling–earnings relationship is very strong in the sense of population averages, but the error in prediction is very large for any given person. Large personal prediction errors dull the value of education as a signal. This fact also suggests that education is a personally risky investment. Third, when the signalling model is expanded, it does not necessarily imply that educational signals are socially unproductive. Education may have significant social value in identifying naturally talented people if there is social value in classification and sorting. For example, there may be significant interactions among workers in an organization. If so, then the organization must be structured to choose the optimal *distribution* of talent within it; for example, it may be socially beneficial for the most talented people to work together. In so far as the educational system serves to classify people for these purposes, it is producing a form of human capital (information in this case) which has both private and social value. Finally, the value of education in assisting persons to find their niche in the overall scheme of the economy, precisely because they do not know so much about themselves, has never been quantified.

Signalling and Information

A definitive empirical study capable of distinguishing signalling and human capital views of investment in education is yet to be produced in spite of many attempts to do so. Most work in this area has floundered on the fact that the two views imply very similar equilibrium implications about the observed relationship between earnings and schooling, so that if any real progress is to be made, future investigations will have to look elsewhere. A promising area is to examine the direct effects of education on productivity (and not on income alone). Much research has been done on educational production functions, which have an obvious bearing on these linkages and how a different form of education might affect them. For example, some evidence suggests that preschool training can overcome the adverse effects of a poor home environment in educational success. Hanushek (1977) reviews the literature on educational production.

Surprisingly few studies have attempted to examine the schooling–productivity linkage directly, probably because data on personal productivity measures are hard to find, but those few that have managed to do so have found some very impressive results. Griliches reviews the issues at the aggregate level. However, the sharpest results have arisen in agriculture, a sector which has shown an enormous and sustained growth in productivity for at least five decades. The rate of return to education among farmers is substantial. Since most of these persons are selfemployed and sell their produce in impersonal, competitive markets, it is difficult to make an *a priori* case that signalling plays any significant role in their educational decisions. Moreover, detailed study shows how these returns come about. More educated farmers control larger resources in the form of larger farms. It is possible that there is a common connection with family background and wealth. However, available evidence suggests that these farmers are also much more efficient in their techniques of production, and that their education is used primarily to keep them informed of recent technological changes in agricultural production, which they adopt with greater frequency and with quicker response. The case that education makes farmers more efficient processors of new information is very well made in the work of Welch (1976, 1979). Schultz indicates that similar findings would apply to much of agricultural production throughout the world, and broadens the argument to make it more generally applicable to all walks of life.

Non-Monetary Considerations

Another potential source of bias in rate of return calculations arises from the limitations of earnings data. Using expected discounted earnings as the choice criterion is a first order approximation to a more complete formulation. Discounted expected *utility* is the ideal choice index, because an employment relationship is a tie-in between the productive services rendered by human capital skills on the one hand, and the consumption of non-pecuniary aspects of the work environment on the other. The imputed monetary equivalent value of these job-consumption items should be added to earnings in a complete calculation. The same is true of the skills that are utilized outside of the market sector, such as in home production (see Michael 1982).

That individuals may differ in their tastes for employment of alternative forms of human capital leads to the existence of rents in human capital valuations. Furthermore, the evidence suggests that on-the-job consumption values increase with education and skill. Jobs which require more schooling are likely to be more desirable on *both* monetary and non-monetary grounds (this evidence is reviewed in Rosen 1986). Economic theory suggests that some portion of earning capacity would be ‘spent’ on more desirable and more amenable jobs. To the extent that the value of work amenities increase with schooling, observed earnings are a downward biased estimate of total earnings for the more educated, and measured rates of return are downward biased.

These issues are most sharply drawn in the treatment of hours worked in rate of return calculations. For example, if observed earnings alone are used in the calculations, groups such as physicians are found to exhibit large rates of return on their medical education, whereas groups such as teachers are found to earn much lower returns. But physicians work very long hours, perhaps as much as 40 per cent more than the typical worker, whereas teachers work far fewer hours than most other workers; they do not work in the summer, for instance. It is necessary to make judgements about the imputed value of leisure to deal adequately with these differences. If leisure is valued at the wage rate, the proper calculation refers to 'full' income at a common hours-worked standard. Similar considerations apply to growth accounting calculations: The secular increase in embodied skills and human capital has been accompanied by a secular decrease in working hours among the employed population. The imputed value of the quantity and quality of increased 'leisure' should be counted in a measure of welfare. Also, using only market transactions as a basis for calculation conceals the significant value of human capital in home production among those groups, especially women, whose activities have shifted between the non-market and market sectors.

Occupational Choice

The discussion so far has concentrated on the role of formal schooling in human capital production. A small but important literature has used these ideas to analyse occupational choice, especially among the professions. The first, and still significant work in this area is due to Friedman and Kuznets, who set the general framework in terms of wealth maximization and rate of return calculations on entry into law, medicine and dentistry. Subsequent literature, of which the work of Freeman is especially notable, has applied modern time-series statistical methods to these problems, concentrating especially on the role of income prospects in attracting or repelling new entrants into a profession.

The human capital perspective suggests that longer term income prospects should play an important role in occupational decisions of the young and that shortterm and transitory fluctuations should be of lesser consequence because they have small impact on expected lifetime wealth. Nevertheless, a central finding in this literature is that current market conditions have large effects on occupational choice, and that supply to a specific occupation is relatively elastic with respect to current wages. The effects of long-term prospects have been much more difficult to isolate empirically, depending as they do on specific formulations of expectations and the connections between future earnings expectations and current and past realizations. In so far as a person is 'locked in' to a profession after choosing it, economic theory suggests that long-term expectations should be the primary determinant of choice. The finding that current prospects are highly significant in these choices suggests considerable mobility and recalibration of choices after training. For example, many lawyers use their skills outside the formal practice of law and in complementary ways in the business sector more generally. However, the nature and extent of expost mobility possibilities remains to be thoroughly examined.

Learning From Experience

From the theoretical point of view, formal schooling decisions are only half the story in human capital accumulation and skill development. Investment does not cease after schooling: there is another sense in which it just begins. Formal schooling sets the stage for accumulation of specific skills and learning in concrete work situations, through on-the-job training. The human capital literature interprets the term 'on-the- job training' very broadly. Only a small part of the overall concept is included in formal training programmes, apprenticeships and the like. The greater part is associated with learning from experience. This broad and inclusive interpretation is supported by persistent empirical observations on the evolution of earnings over the life-cycle. The age structure of earnings shows remarkably systematic patterns. Earnings rise rapidly in the first several years of working life, but the rate of growth falls toward mid-career and tends to turn negative toward retirement. In panel data, wage rates rise throughout the life cycle, with the greatest rate of increase in the early years. An attractive interpretation of these observations is that the increase in earnings with work experience is due to increasing productivity and human capital accumulation over the entire life cycle.

A fruitful empirical approach for studying these patterns has been developed by Jacob Mincer (1974). The conception of the problem extends the education model above. A person is viewed as making human capital investment choices at each point in the life cycle. Workers who choose to invest more pay for their choice by accepting lower earnings when young and earn returns on their prior investments in the form of larger earnings when they are older. This is essentially a choice between a level experience–earnings pattern (if investments are small) and a 'tilted' one, starting at a lower point and rising to a higher one if investments are large. Mincer develops the concept of 'overtaking' to impute the total return to human capital. The basic idea extends the Smithian principle of compensation to on-the-job training investments. Suppose a person has a large variety of possible investment opportunities after completing school. If no further investments are made, the experience earnings profile is relatively flat. The slope of the earnings-experience profile is increasing and the intercept of the profile is decreasing with the magnitude of investment. Hence the investment level defines an entire family of age earnings profiles, which are spun out around a roughly common crossing point, labelled the 'overtaking' point, if in market equilibrium wealth is approximately independent of investment.

The model has a very sharp empirical prediction that in a cohort of individuals with the same schooling level and different post-school investments, the interpersonal variance of earnings should be decreasing with experience up to the overtaking point and increasing thereafter. These systematic variance patterns have been found by many investigators in a variety of data sources. The assumptions that on-the-job investments are completely equalizing and that human wealth is the same for all investment paths makes it possible to decompose total investments into formal education and on-the-job components. Mincer reports that the on-the-job components are substantial, of the order of a third or more of the total.

The complete education–experience human capital model has important implications for the analysis of poverty and income distributions. In a nutshell, human capital theory suggests that lifetime earnings is the appropriate construct for understanding inequality. To the extent that age-earnings patterns are the result of rational investments in human capital, it is misleading to use unadjusted crosssection annual earnings data for inequality analysis. For those young persons who are intensively engaged in investment activities and whose current income is therefore small at present may be classified erroneously as poor even though they are not poor in the lifetime sense. These life cycle issues have not been given sufficient attention in the extensive literature on the social welfare consequences of inequality, in spite of the fact that Paglin (1975) conclusively shows that they have large consequences for the measurement of inequality. Taking the life cycle view yields Gini coefficient estimates of real inequality that are smaller than when only current incomes are used in the calculations.

More detailed econometric work on the dynamic structure of individual earnings based on panel data helps resolve questions of the extent to which poverty status is permanent or transitory over the life cycle. The most sophisticated study so far (Lillard and Willis 1978) decomposes earnings into several components. One is measureable characteristics of persons, such as education and experience, which reflect human capital and other considerations. Another is a 'person effect' capturing unmeasured components of ability, health, and related factors which permanently affect a person's earning power relative to his cohort. Finally, the third component reflects more transient variations, reflecting such factors as luck and other random events which may persist for a time but which eventually die out. Each component explains about one-third of the total variance of earnings. Since the measurable factors are, by human capital theory, largely equalizing on prior investments and the transitory effects have only small effects on life cycle wealth, this leaves about one- third of the total variance of life cycle earnings as attributable to permanent differences among persons or to 'pure' inequality. Certainly this is quite a different picture than emerges from examining the cross-section distribution of current earnings.

Other approaches to understanding age–earnings profiles in the human capital framework have used a more formal capital–theoretic structure. Here human capital is associated with the latent stock of embodied skills and investment with skill acquisition and learning. A person must give up current income to learn more and increase the stock of skills available for rental at a later date. The optimal investment programme maximizes the present value of lifetime earnings. This basic set-up of the problem was first formulated in an important paper by Ben-Porath (1967), who structured the investment control as choice of the division of a person's time between working and investing. An extension by Rosen (1972) structures it as choice among a spectrum of jobs which offer different learning environments and opportunities. The wage on a job that offers more learning possibilities is lower and the programme is implemented by a 'stepping stone' progression of positions.

This capital theoretic formulation of the problem has virtues in demonstrating the conceptual commonalities between capital and growth theory and human capital theory. However, its generality comes at the cost of providing less robust predictions. Thus it seems fair to say that extensive work attempting to implement these rigorous ideas empirically has not met with overwhelming success in extracting information from observed age–experience trajectories. It appears that other important forces also affect these patterns. Several possibilities have been suggested. One relates to investments in information and search for enduring long job attachments. Job turnover is much larger among young workers than older ones. While this is a form of human capital accumulation and much recent work has been devoted to these issues, it has so far proven difficult to link this class of problems with the ideas reviewed here. Nor has human capital theory yet adequately come to terms with the fact that job patterns typically exhibit discrete jumps and 'promotions', where the character of human capital services rendered changes at each step. Competition for higher ranking positions is properly considered within the human capital framework, but little analysis is available so far.

Any review of human capital would be remiss in not calling attention to parallel developments and important applications in economic historians' interpretation of slavery. The work of Fogel and Engerman (1974) stands out as the primary example of the approach. Here the empirical work focuses on direct human capital valuations rather than on earnings. The principles of capital valuation are used to examine such issues as the long-term economic viability of slavery as an economic institution in the absence of intervention. In addition, some important and fascinating agency problems must be confronted because of an inherent conflict in the master-slave relationship. The conflict arises because the owner naturally desires more effort than the slave prefers to put forth. Various institutions, involving *both* punishments and rewards, were structured to help resolve these conflicts. Mention also should be made of research on indentured servitude by economic historians (Galenson 1981), which is analysed as a response to a capital market imperfection. A person voluntarily indentured himself for a period of years as payment for a loan to provide transportation and connections in the New World. Repayment was guaranteed by a legally binding claim on the person's services for the period of the contract.

Demographic Effects

Over the years there has been increasing recognition of the relationship between human capital and economic demography. This is inherent in the role of families as both producers and financiers of human capital investments. Two important recent developments strongly rest on these connections.

The first one is related to large demographic changes in the age structure of the population in the post-war period (the 'baby boom') in the United States. Rates of return on education had remained remarkably constant for a thirty-year period. This in spite of the fact that there had been an enormous increase in education over that period. However, Freeman identified a decline in the rate of return commencing in the late 1960s. The evidence currently available suggests that the rate fell by several percentage points for a 10–12 year period throughout the 1970s, but had gradually returned to its prior level. The leading explanation for this has been provided by Welch (1979) and relates to increased competition for jobs within cohorts as a function of their size.

A stable age distribution of the working population provides a naturally stable progression of work and job opportunities over a person's working life. Not only the level, but also the nature and productive role of human capital changes over the life cycle. Young workers perform different tasks and have different responsibilities than do older workers. Therefore competition and supply of human capital of various types in the labour market is strongly age related. Thus as the large birth cohorts of the 1950s began to enter the market in the late 1960s and 1970s, the increased supply of educated young workers lowered their wage rates and reduced the rate of return. These effects are diffused as the large cohort ages and works its way through the age distribution, and as the structure of work is altered to accommodate their large numbers. The weight of extensive research in this area has shown that returns and wage rates are affected by cohort size. The consequences of this research for the future development of human capital theory will be important, because it requires considering heterogeneous human capital investments and the evolution and development of different types of skills over working life. It may ultimately require analysing how work itself is organized and structured.

Human Capital and Discrimination

A final important recent development proceeds on somewhat more conventional theoretical grounds. It addresses the role of human capital in observed wage differences between men and women, and is ultimately related to questions of labour market discrimination. The work in this area is firmly based on empirical calculations. The main fact to be explained is that women earn less than men, even after adjusting for differences in occupational status and hours worked. Labour market discrimination against women is one possible interpretation. However, there may be more subtle forces at work. Mincer and Polachek (1974) build an alternative interpretation on the observation that earnings–experience profiles of women are flatter and exhibit much less life-cycle growth than that of men, and tied it to the well known fact that women traditionally have exhibited less stronger labour force attachments than men due to the sexual division of labour in the home and the bearing and raising of children.

The value of an investment increases with its rate of utilization. Compare two persons: one who expects to utilize an acquired skill very intensively and one who expects to utilize it less intensively. Suppose further that the costs of acquiring the skill are approximately independent of its subsequent utilization. Then the rate of return on investment is larger for the intensive user and that person will tend to invest more. The application to male–female wage differential is apparent upon connecting intensity of utilization with labour force attachments and hours worked. In so far as married women play dual roles in the market and in the household, there is a tendency to invest less in labour market skills and more in non-market skills. The opposite is true of men, given prevailing marriage institutions. These differential incentives can account for differences in age earnings patterns between men and women as well as the larger average wages of men. Research on female labour supply supports the point by showing overwhelming evidence that labour force activities of married women are severely constrained by the presence of children in the home. Mincer and Polachek provided direct empirical support by demonstrating that earnings of never-married women closely approximate those of men.

Considerable research is in progress on these ideas (see, for example, *Journal of Labor Economics,* 1985). At a minimum, the human capital perspective shows that these issues are more complicated than appears on the surface. Yet there are some unresolved puzzles. In spite of the vast increase in female labour force participation in the past two decades, the relative wages of men and women have not changed very much in the United States, though they have come closer to parity in a number of other countries. Part of this may be due to differences in the importance of the government sector as employers of women, as well as differences in compliance with equal pay legislation. A definitive answer is not yet on the horizon.

This essay started by noting the twin origins of developments of the theory of human capital in understanding the sources of economic growth on the one hand and the distribution of economic rewards on the other. Much progress has been made on both counts. However, these two branches have not yet been clearly joined. Future progress will have to come to terms with the issue of how private incentives to acquire human capital affect the available social stock of productive knowledge and how changes in social knowledge become embodied in the skills of subsequent generations.

See Also

- ▶ Cognitive Ability
- ▶ Family Economics
- ▶ Gender Differences (Experimental Evidence)
- ▶ Gender Roles and Division of Labour
- ▶ Household Production and Public Goods
- ▶ Measurement
- ▶ Signalling and Screening
- ▶ Value of Time

Bibliography

Ashenfelter, O. 1978. Estimating the effect of training programs on earnings. *Review of Economics and Statistics* 60 (1): 47–57.

Becker, G. 1964. *Human capital*. 2nd ed, 1975. New York: Columbia University Press.

Becker, G., and N. Tomes. 1978. An equilibrium theory of the distribution of income and intergenerational mobility. *Journal of Political Economy* 87 (6): 1153–1189.

Ben-Porath, Y. 1967. The production of human capital and the life cycle of earnings. *Journal of Political Economy* 75(4) Pt 1: 352–365.

Blaug, M. 1976. The empirical status of human capital theory: A slightly jaundiced survey. *Journal of Economic Literature* 14 (3): 827–855.

Chiswick, B. 1978. The effect of Americanization on the earnings of foreign-born men. *Journal of Political Economy* 86 (5): 897–921.

Denison, E. 1962. *The sources of economic growth in the United States and the alternatives before us*. New York: Committee for Economic Development.

Dublin, L., and A. Lotka. 1930. *The monetary value of a man*. New York: Ronald Press.

Fogel, R., and S. Engerman. 1974. *Time on the cross*. New York: Little, Brown.

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

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

Friedman, M., and S. Kuznets. 1954. *Income from independent professional practice*. Princeton: Princeton University Press.

Galenson, D. 1981. *White servitude in colonial America*. Cambridge: Cambridge University Press.

Griliches, Z. 1970. Notes on the role of education in production functions and growth accounting. In *Education, income and human capital*, ed. L. Hansen. New York: National Bureau of Economic Research.

Griliches, Z. 1977. Estimating the returns to schooling: Some economic problems. *Econometrica* 45 (1): 1–22.

Hanushek, E. 1977. *A reader's guide to educational production functions*. New Haven: Institution for Social Policy Studies, Yale University.

Kendrick, J. 1976. *The formation and stocks of total capital*. New York: National Bureau of Economic Research.

Knight, F. 1944. Diminishing returns from investment. *Journal of Political Economy* 52 (March): 26–47.

Lillard, L., and R.J. Willis. 1978. Dynamic aspects of earnings mobility. *Econometrica* 46 (5): 985–1012.

Marshall, A. 1920. *Principles of economics*. 8th ed, 1930. London: Macmillan.

Michael, R. 1982. Measuring non-monetary benefits of education: A survey. In *Financing education: Overcoming inefficiency and inequity*, ed. W. McMahon and T. Geske. Urbana: University of Illinois Press.

Miller, H. 1960. Annual and lifetime income in relation to education, 1929–1959. *American Economic Review* 50 (December): 962–986.

Mincer, J. 1958. Investment in human capital and personal income distribution. *Journal of Political Economy* 66 (August): 281–302.

Mincer, J. 1974. *Schooling, experience and earnings*. New York: Columbia University Press.

Mincer, J. and S. Polachek. 1974. Family investment in human capital: Earnings of women. *Journal of Political Economy* 82(2) Pt II: S76–S108.

Paglin, M. 1975. The measurement and trend of inequality: A basic revision. *American Economic Review* 65 (4): 589–609.

Petty, W. 1676. Political arithmetic. In *The economic writings of Sir William Petty*, vol. 1, ed. C. Hull. Cambridge: Cambridge University Press, 1899.

Rosen, S. 1977. Human capital: A survey of empirical research. In *Research in Labor Economics*, ed. R. Ehrenberg, vol. 1. Greenwich: JAI Press.

Rosen, S. 1985. The theory of equalizing differences. In *Handbook of labour economics*, ed. O. Ashenfelter and R. Layard. Amsterdam: North-Holland.

Rosen, S. 1986. The theory of equalizing differences. In *Handbook of labour economics*, ed. O. Ashenfelter and R. Layard. Amsterdam: North Holland.

Sahota, G. 1978. Theories of personal income distribution: A survey. *Journal of Economic Literature* 16 (1): 1–55.

Schultz, T. 1961. Investment in human capital. *American Economic Review* 51 (March): 1–17.

Schultz, T. 1975. The value of the ability to deal with disequilibria. *Journal of Economic Literature* 13 (3): 827–846.

Smith, A. 1776. *an inquiry into the nature and causes of the wealth of nations*, Modern library edition. New York: Random House, 1947.

Spence, M. 1973. Job market signaling. *Quarterly Journal of Economics* 87 (3): 355–374.

Walsh, J. 1935. Capital concept applied to man. *Quarterly Journal of Economics* 49 (February): 255–285.

Welch, F. 1970. Education in production. *Journal of Political Economy* 78 (1): 35–59.

Welch, F. 1976. *Ability tests and measures of differences between black and white Americans*. Rand Corporation.

Welch, F. 1979. Effects of cohort size on earnings: The baby boom babies' financial bust. *Journal of Political Economy* 87(5) Pt II: S65–97.

Willis, R. 1986. Wage determinants: A survey and reinterpretation of human capital earnings functions. In *Handbook of labour economics*, ed. O. Ashenfelter and R. Layard. Amsterdam: North-Holland.

Willis, R. and S. Rosen. 1978. Education and self-selection. *Journal of Political Economy* 87(5) Pt II: S65–S97.

Human Capital, Fertility and Growth

Oded Galor

Abstract

The worldwide demographic transition of the past 140 years has been identified as one of the prime forces in the transition from stagnation to growth. The unprecedented increase in population growth during the early stages of industrialization was ultimately reversed. The rise in the demand for human capital in the second phase of industrialization brought about a significant reduction in fertility rates and population growth in various regions of the world, enabling economies to convert a larger share of the fruits of factor accumulation and technological progress into growth of income per capita.

Keywords

Agricultural Revolution; Child labour; Child quality; Consumption smoothing; Credit constraints; Demographic transition; Economic growth; Educational attainment; Fertility; Gender gap; Growth take-offs; Human capital; Industrial Revolution; Industrialization; Infant mortality; Life expectancy; Mortality; Old-age security hypothesis; Population growth; Public education; Technological progress; Wage differentials

JEL Classifications

O15; J10; J20

The transition from stagnation to growth has been the subject of intensive research in recent years. The rise in the demand for human capital and the associated decline in population growth have been identified as the prime forces in the movement from an epoch of stagnation to a state of sustained economic growth. They have brought about a significant formation of human capital along with a reduction in fertility rates and population growth, enabling economies to convert a larger share of the fruits of factor accumulation and technological progress into growth of income per capita.

Historical Evidence

The evolution of economies throughout human history has been characterized by Malthusian stagnation. Technological progress and population growth were minuscule by modern standards, and the average growth rate of income per capita was even slower, due to the offsetting effect of population growth on the expansion of resources per capita. In the past two centuries, on the other hand, the pace of technological progress increased significantly, alongside the process of industrialization. Various regions of the world departed from the Malthusian trap and initially experienced a considerable rise in the growth rates of income per capita and population. In contrast to episodes of technological progress in the pre-Industrial Revolution era, which failed to generate sustained economic growth, the increasing role of human capital in the production process in the second phase of the Industrial Revolution ultimately prompted a demographic transition, liberating the gains in productivity from the counterbalancing effects of population growth. The decline in population growth and the associated advancement in technological progress and human capital formation paved the way for the emergence of the modern state of sustained economic growth.

The evolution of population growth in the world economy has been non-monotonic. The growth of world population was sluggish during the Malthusian epoch, creeping at an average annual rate of about 0.1 per cent over the years 0–1820 (Maddison 2001). The Western European take-off along with that of the Western Offshoots (that is, the United States, Canada, Australia and New Zealand) brought about a sharp increase in population growth in these regions. The world annual average rate of population growth increased gradually reaching 0.8 per cent in the years 1870–1913. The take-off of less developed regions and the significant increase in their income per capita generated a further increase in the world rate of population growth, despite the decline in population growth in Western Europe and the Western Offshoots, reaching a high level of 1.92 per cent per year in the period 1950–73. Ultimately, the onset of the demographic transition in less developed economies in the second half of the 20th century, reduced population growth to an average rate of 1.63 per cent per year in the period 1973–98.

The timing of the demographic transition differed significantly across regions. A reduction in population growth occurred in Western Europe, the Western Offshoots, and Eastern Europe towards the end of the 19th century and in the beginning of the 20th century, whereas Latin America and Asia experienced a decline in the rate of population growth only in the last decades of the 20th century.

The demographic transition in Western Europe occurred towards the turn of the 19th century. A sharp reduction in fertility took place simultaneously in several countries in the 1870s, and resulted in a more than 30 per cent decline in fertility rates within a 50-year period. Over the period 1875–1920, crude birth rates declined by 44 per cent in England, 37 per cent in Germany, and 32 per cent in Sweden and Finland. A decline in mortality rates preceded the decline in fertility rates in most of Western Europe. It began in England nearly 140 years prior to the decline in fertility, and in Sweden and Finland the corresponding figure was 100 years. The decline in fertility outpaced the decline in mortality rates and brought about a decline in the number of children who survived to their reproduction age.

A similar pattern characterizes mortality and fertility decline in less developed regions. The total fertility rate over the period 1960–99 plummeted from 6 to 2.7 in Latin America, from 6.14 to 3.14 in Asia, and declined moderately from 6.55 to 5 in Africa, along with a sharp decline in infant mortality rates.

Theories of the Demographic Transition

The Decline in Infant and Child Mortality

The decline in infant and child mortality rates has been a dominating explanation for the onset of the decline in fertility in many developed countries, with the notable exceptions of France and the United States. Nevertheless, this viewpoint appears inconsistent with historical evidence. While it is highly plausible that mortality rates were among the factors that affected the level of fertility throughout human history, historical evidence does not lend credence to the argument that the decline in mortality rates accounts for the reversal of the positive historical trend between income and fertility.

The mortality decline in Western Europe started nearly a century before the decline in fertility and was associated initially with increasing fertility rates in some countries and non-decreasing fertility rates in others. In particular, the decline in mortality started in England in the 1730s, and until 1820 was accompanied by a steady increase in fertility rates. The significant rise in income per capita in the post- Malthusian regime apparently increased the desirable number of surviving offspring and thus, despite the decline in mortality rates, fertility increased significantly so as to reach this higher desirable level. The decline in fertility during the demographic transition occurred in a period in which this pattern of increased income per capita (and its potential effect on fertility) was intensified, while the pattern of declining mortality (and its adverse effect on fertility) maintained the trend that existed in the 140 years preceding the demographic transition. The reversal in fertility patterns in England and in other Western European countries in the 1870s suggests therefore that the demographic transition was not prompted by a decline in infant and child mortality.

Furthermore, most relevant from an economic point of view is the cause of the reduction in net fertility (that is, the number of children reaching adulthood). The decline in the number of surviving offspring that was observed during the demographic transition is unlikely to have been a result of mortality decline. Mortality decline would have led to a reduction in the number of surviving offspring if the following implausible conditions had been met: (*a*) there existed a precautionary demand for children, that is, individuals were risk averse with respect to the number of surviving offspring; (b) risk aversion with respect to consumption was smaller than risk aversion with respect to fertility (evolutionary theory would suggest the opposite); (*c*) sequential fertility (that is, replacement of non-surviving children) was modest.

The Rise in the Level of Income Per Capita

The rise in income per capita prior to the demographic transition has led some researchers to argue that the demographic transition was triggered by the asymmetric effects of the rise in income per capita on household income and on the opportunity cost of bringing up children. Becker (1981) argues that the rise in income induced a fertility decline because the positive income effect on fertility was dominated by the negative substitution effect that was brought about by the rising opportunity cost of children. Similarly, he argues that the income elasticity with respect to child quality is greater than that with respect to child quantity, and hence a rise in income led to a decline in fertility along with a rise in the investment in each child.

This theory suggests that the timing of the demographic transition across countries in similar stages of development would reflect differences in income per capita. However, remarkably, the decline in fertility occurred in the same decade across Western European countries despite their differing significantly in their income per capita. In 1870, on the eve of the demographic transition, England was the richest country in the world, with a GDP per capita of 3191 dollars (measured in 1990 international dollars: Maddison 2001). In contrast, Germany, which experienced the decline in fertility in the same years as England, had in 1870 a GDP per capita of only 1821 dollars (that is, 57 per cent of that of England). Sweden’s GDP per capita of 1664 dollars in 1870 was 48 per cent of that of England, and Finland’s GDP per capita of 1140 dollars in 1870 was only 36 per cent of that of England, but their demographic transitions occurred in the same decade. The simultaneity of the demographic transition across Western European countries that differed significantly in their income per capita suggests that the high level of income reached by Western Europeans countries in the post-Malthusian regime had a very limited role in the demographic transition.

The Rise in the Demand for Human Capital

The gradual rise in the demand for human capital in the second phase of the Industrial Revolution (and in the process of industrialization of less developed economies) and its close association with the timing of the demographic transitions has led researchers to argue that the increasing role of human capital in the production process induced households to increase investment in the human capital of their offspring, ultimately leading to the onset of the demographic transition.

Galor and Weil (1999, 2000), argue that the acceleration in the rate of technological progress gradually increased the demand for human capital in the second phase of the Industrial Revolution, inducing parents to invest in the human capital of their offspring. The increase in the rate of technological progress and the associated increase in the demand for human capital brought about two effects on population growth. On the one hand, improved technology eased households' budget constraints and provided more resources for the quality as well as the quantity of children. On the other hand, it induced a reallocation of these increased resources towards child quality. In the early stages of the transition from the Malthusian regime, the effect of technological progress on parental income dominated, and the population growth rate as well as the average quality increased. Ultimately, further increases in the rate of technological progress, stimulated by human capital accumulation, induced a reduction in fertility rates, generating a demographic transition in which the rate of population growth declined along with an increase in the average level of education. Thus, consistent with historical evidence, the theory suggests that prior to the demographic transition, population growth increased along with investment in human capital, whereas the demographic transition brought about a decline in population growth along with a further increase in human capital formation.

Galor and Weil's theory suggests that a universal acceleration in technological progress raised the demand for human capital in the second phase of the Industrial Revolution and generated a simultaneous increase in educational attainment and demographic transition across Western European countries that differed significantly in their levels of income per capita. Consistent with the theory, the growth rates (as opposed to the levels) of income per capita among these Western European countries were rather similar during their demographic transition, ranging from 1.9 per cent per year over the period 1870–1913 in the UK, 2.12 per cent in Norway, 2.17 per cent in Sweden, to 2.87 per cent in Germany. Moreover, the demographic transition in England was associated with a significant increase in the investment in child quality as reflected by years of schooling. Moreover, international trade and its differential effects on the demand for human capital had an asymmetric effect of the timing of the demographic transition (Galor and Mountford 2006).

Evidence about the evolution of the return to human capital over this period is scarce and controversial, but it does not indicate that the skill premium increased markedly in Europe over the course of the 19th century, nor is it an indication of the absence of a significant increase in the demand for human capital. Technological progress in the second phase of the Industrial Revolution brought about an increase in the demand for human capital, and indeed, in the absence of a supply response, one would have expected an increase in the return to human capital. However, the significant increase in schooling in the 19th century, and in particular the introduction of publicly provided education, which lowered the cost of education, generated a significant increase in the supply of educated workers. Some of this supply response was a direct reaction to the increase in the demand for human capital, and thus may only operate to partially offset the increase in the return to human capital. However, the removal of the adverse effect of credit constraints on the acquisition of human capital (for example, Galor and Zeira 1993 and Galor and Moav 2006), as reflected by the introduction of publicly provided education, generated an additional force that increased the supply of educated labour and operated towards a reduction in the return to human capital.

The Decline in Child Labour

The effect of the rise in the demand for human capital on the reduction in the desirable number of surviving offspring was magnified via its adverse effect on child labour. It gradually increased the wage differential between parental labour and child labour, inducing parents to reduce the number of their children and to further invest in their quality (Hazan and Berdugo 2002). Moreover, the rise in the importance of human capital in the production process induced industrialists to support education reforms (Galor and Moav 2006) and thus laws that abolished child labour (Doepke 2004; Doepke and Zilibotti 2005), and thus fertility.

The Rise in Life Expectancy

The impact of the increase in the demand for human capital on the decline in the desirable number of surviving offspring was reinforced by improvements in health and life expectancy. Despite the gradual rise in life expectancy prior to the demographic transition, investment in human capital was insignificant as long as a technological demand for human capital had not emerged. The technologically based rise in the demand for human capital during the second phase of the Industrial Revolution and the rise in the expected length of productive life increased the potential rate of return to investments in children's human capital, reinforcing the inducement for investment in education and the associated reduction in fertility rates (Galor and Weil 1999; Moav 2005; Soares 2005).

Natural Selection and the Evolution of Preference for Offspring's Quality

The impact of the increase in the demand for human capital on the decline in the desirable number of surviving offspring may have been magnified by cultural or genetic evolution in the attitude of individuals towards child quality. Galor and Moav (2002) propose that during the epoch of Malthusian stagnation that characterized most of human existence, individuals with a higher valuation for offspring quality (in the context of the quantity-quality survival strategies) gained an evolutionary advantage and their representation in the population gradually increased. The Agricultural Revolution facilitated the division of labour and fostered trade relationships across individuals and communities, enhancing the complexity of human interaction and raising the return to human capital. Moreover, the evolution of the human brain in the transition to *Homo sapiens* and the complementarity between brain capacity and the reward for human capital has increased the evolutionary optimal investment in the quality of offspring. The distribution of valuation for quality lagged behind the evolutionary optimal level and individuals with traits of higher valuation for their offspring's quality generated higher income and, in the Malthusian epoch, a higher number of offspring. Thus, the trait of higher valuation for quality gained the evolutionary advantage. This evolutionary process was reinforced by its interaction with economic forces. As the fraction of individuals with high valuation for quality increased, technological progress intensified, raising the rate of return to human capital. The increase in the rate of return to human capital along with the increase in the bias towards quality in the population reinforced the substitution towards child quality, setting the stage for a more rapid decline in fertility along with a significant increase in investment in human capital and a transition to sustained economic growth.

The Decline in the Gender Gap

The rise in the demand for human capital and its impact on the decline in the gender gap in the last two centuries could have reinforced a demographic transition and human capital formation. Galor and Weil (1996, 1999) argue that technological progress and capital accumulation complemented mental-intensive tasks and substituted for physical-intensive tasks in industrial production. In light of the comparative physiological advantage of men in physical-intensive tasks and women in mental-intensive tasks, the demand for women's labour input gradually increased in the industrial sector, decreasing monotonically the wage differential between men and women. In early stages of industrialization, the wages of both men and women increased, but the rise in women's wages was not sufficient to induce a significant increase in the female labour force. Fertility, therefore, increased due to the income effect that was generated by the rise in men's absolute wages. Ultimately, however, the rise in women's relative wages was sufficient to induce a significant increase in labour force participation. It increased the cost of bringing up children proportionally more than household income, generating a decline in fertility and a shift from stagnation to growth.

The Old-Age Security Hypothesis

The old-age security hypothesis (Caldwell 1976) has been proposed as an additional mechanism for the onset of the demographic transition. It suggests that in the absence of capital markets that permit intertemporal lending and borrowing, children are assets that permit parents to smooth consumption over their lifetime. The process of development and the establishment of capital markets reduce this motivation for bringing up children, contributing to the demographic transition. The significance of the decline in the role of children as assets in the onset of the demographic transition is questionable. The rise in fertility rates prior to the demographic transition, in a period of improvements in the credit markets, raises doubts about the significance of the mechanism. Furthermore, cross-section evidence (Clark and Hamilton 2006) from the pre-demographic transition era indicates that wealthier individuals, who presumably had better access to credit markets, had a larger number of surviving offspring.

Concluding Remarks

The rise in the demand for human capital in the second phase of industrialization and its effect on decline in population growth have been among the prime forces in the transition of economies from an epoch of stagnation to a state of sustained economic growth. They brought about a significant formation of human capital along with a reduction in fertility rates and population growth, enabling economies to advance technologically and to convert a larger share of the fruits of factor accumulation and technological progress into growth of income per capita.

See Also

- ▶ Demographic Transition
- ▶ Economic Growth in the Very Long Run
- ▶ Growth Take-Offs

Bibliography

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

Caldwell, W.J. 1976. Toward a restatement of demographic transition theory. *Population and Development Review* 2: 321–366.

Clark, G., and G. Hamilton. 2006. Survival of the richest. *Journal of Economic History* 66: 707–736.

Doepke, M. 2004. Accounting for fertility decline during the transition to growth. *Journal of Economic Growth* 9: 347–383.

Doepke, M., and F. Zilibotti. 2005. The macroeconomics of child labor regulation. *American Economic Review* 95: 1492–1524.

Galor, O., and O. Moav. 2002. Natural selection and the origin of economic growth. *Quarterly Journal of Economics* 117: 1133–1192.

Galor, O., and O. Moav. 2006. Das human-kapital: A theory of the demise of the class structure. *Review of Economic Studies* 73: 85–117.

Galor, O., and A. Mountford. 2006. Trade and the great divergence: The family connection. *American Economic Review* 96: 299–303.

Galor, O., and D.N. Weil. 1996. The gender gap, fertility, and growth. *American Economic Review* 86: 374–387.

Galor, O., and D.N. Weil. 1999. From Malthusian stagnation to modern growth. *American Economic Review* 89: 150–154.

Galor, O., and D.N. Weil. 2000. Population, technology and growth: From the Malthusian regime to the demographic transition and beyond. *American Economic Review* 110: 806–828.

Galor, O., and J. Zeira. 1993. Income distribution and macroeconomics. *Review of Economic Studies* 60: 35–52.

Hazan, M., and B. Berdugo. 2002. Child labor, fertility and economic growth. *Economic Journal* 112: 810–828.

Lagelof, N. 2003. Gender equality and long-run growth. *Journal of Economic Growth* 8: 403–426.

Maddison, A. 2001. *The World economy: A millennia perspective*. Paris: OECD.

Moav, O. 2005. Cheap children and the persistence of poverty. *Economic Journal* 115: 88–110.

Soares, R.R. 2005. Mortality reductions, educational attainment, and fertility choice. *American Economic Review* 95: 580–601.

Human Development in the Middle East and North Africa

Djavad Salehi-Isfahani

Abstract

Recent uprisings in the Arab world raise important questions about the human development performance of the Middle East and North Africa (MENA) in recent decades. In this article I review the record of progress in key aspects of human development. According to the widely used Human Development Index, the average level of human development in MENA is commensurate with the region's general level of economic development. The assessment is less favourable when we examine how opportunities for human development are distributed across the population. Other aspects of wellbeing that are not included in the standard measures of human development, but which affect its assessment, such as the challenges faced by women and young people, further revise downward our evaluation of the region's progress in human development. I complement this quantitative review with a discussion of how key structural features of MENA economies – high oil income, late demographic transition and low productivity of education – have shaped human development in the MENA region.

Keywords

Human development; Human Development Index; Inequality; Gender; Middle East and North Africa; Youth

JEL Classifications

O15; O53; J13; J16; D63

Introduction

Widespread political unrest in the Arab world since 2011 might easily give the impression of a region in a very poor state of human development, but, at least according to international indicators, the record of human development in the Middle East and North Africa (MENA) in recent decades has been quite respectable. As recently as 2010, before the first uprisings in Tunisia shook the region, the United Nations Human Development Report (HDR) placed five Arab countries (including Tunisia, the first Arab country to revolt) on its list of the top ten countries that improved their human development in the preceding 40 years (UNDP 2010). The positive view of human development in MENA based on the Human Development Index (HDI) – a combination of measures of income per capita, education and health published annually by the United Nations Development Program (UNDP), has not gone unchallenged, however. For example, the 2002 Arab Human Development Report, the first in the series, while acknowledging the positive achievements in HDI and falling poverty in the previous three decades, highlighted a 'freedom deficit' in the Arab world, which manifests itself in lagging democratic institutions and lack of civic and political rights. Human development is closely bound with enhancing human capabilities that enable individuals 'to exercise freedom and human rights'. Thus without removing the freedom deficit the report is pessimistic about sustained human development in the Arab world.

In this article I survey the state of human development in MENA, emphasising those aspects of human development that are quantifiable and for which data is available across countries and time. I emphasise comparison with other countries because I believe that human development should be assessed not just in relation to the advanced regions of the world but also relative to the level of a nation's economic development. So the question is more about how MENA is doing relative to its own past and relative to other countries with similar levels of development than in absolute terms and relative to an ideal.

I begin with a detailed discussion of what we can learn from the standard HDI analysis and then proceed with other dimensions of human development that the HDI metric does not take into account. Inequality in human development is the most important limitation of the standard HDI reporting, which is often at the level of national average values. Since 2010, HDRs report inequality-adjusted HDI values that are adjusted for the inequality in the distribution of each sub-index. Because of the relatively low level of inequality in MENA (Bibi and Nabli 2009; Belhaj-Hassine 2015), the overall story of human development in the region as told by HDI does not change much when viewed with inequality-adjusted HDI values. A less favourable picture emerges when we examine recent research on inequality of opportunity, which provides evidence that opportunities for human development are unevenly available to young people from advantaged and disadvantaged backgrounds. Other important aspects of human development that affect individual welfare, such as civic participation and political liberties, which are difficult to quantify and compare over time and across countries, but are critical for human development, receive relatively less attention in this article. Finally, I look more closely at young people and women, two social groups that face specific challenges in the context of MENA, which are not well reflected in the HDI calculus (Salehi-Isfahani 2013).

Discussing human development for the MENA region as a whole can be misleading because of the considerable heterogeneity in economic development in the region. While MENA countries are fairly homogeneous in terms of religion, language and culture, they differ widely in income per capita, making comparisons between MENA and other regions meaningless. To deal with this issue, where appropriate, I discuss human development for three groups of MENA countries: low income, defined as those with less than $3,000 in GDP per capita (in 2005 Purchasing Power Parity US dollars); middle income (those between $3,000 and $15,000); and high income (those above $15,000). Table 1 presents the list of countries in these groups. While these cutoffs are rather arbitrary, they do well in separating MENA countries into groups with similar levels of development. The poor countries of Djibouti, Mauritania and Somalia are Arab countries, but because they are often not considered as Middle Eastern, I group them with Sub-Saharan African countries.

The rich countries of the Gulf Cooperation Council, the GCC, comprising Bahrain, Oman, Qatar, Saudi Arabia and United Arab Emirates, along with Libya, enjoy high levels of per capita income because of rents from hydrocarbon exports. This group accounts for only 9% of the region's 725 million population. A second group, with 78% of MENA population, consists of middle-income countries, including oil exporters Iran and Iraq. The third group, accounting for the remaining 13% of the population, consists of the three lowest income countries of the region – Sudan and Yemen. This group, with a (population weighted) average per capita GDP of only $2,150 in 2005 Purchasing Power Parity, is poorer than the average Sub-Saharan African country (Table 1). The middle-income group enjoys an average per capita GDP that is almost four times as high ($7,624) and is slightly below Latin America and the Caribbean. Average income in the richest group ($25,097) is 3.5 times higher than the middle group and in the range for advanced countries.

Human Development in the Middle East and North Africa, Table 1 The list of countries in MENA categories

MENA	Low income	Middle income	High income
Algeria	Sudan	Algeria	Bahrain
Bahrain	Yemen	Egypt	Kuwait
Egypt		Iran	Libya
Iran		Iraq	Oman
Iraq		Jordan	Qatar
Jordan		Lebanon	Saudi Arabia
Kuwait		Morocco	United Arab Emirates
Lebanon		Palestine	
Libya		Syria	
Morocco		Tunisia	
Oman		Turkey	
Palestine			
Qatar			
Saudi Arabia			
Sudan			
Syria			
Tunisia			
United Arab Emirates			
Yemen			

Notes: Low income group have GDP per capita less than $3,000 (2005 PPP), middle income $3,000–$15,000 and high income greater than $15,000

The next section will offer a broad assessment of the performance of the MENA countries according to HDI and its components. The following section discusses the role of poverty and inequality and the extent to which taking them into account revises our assessment, and the final two sections discuss human development issues specific to two social groups: women and young people. Gender inequality is an important aspect

Human Development in the Middle East and North Africa, Table 2 HDI by region over time

Region	1990	2000	2005	2010	2013	Growth 1990–2013
MENA-low	0.36	0.40	0.44	0.47	0.48	33.3
MENA-middle	0.55	0.63	0.66	0.70	0.71	29.1
MENA-high	0.67	0.75	0.78	0.81	0.83	23.9
Sub-Saharan Africa	0.41	0.40	0.43	0.47	0.49	19.5
Asia	0.50	0.56	0.60	0.65	0.66	32.0
LAC	0.62	0.67	0.70	0.73	0.73	17.7
Oceania	0.77	0.80	0.81	0.82	0.83	7.8
Europe	0.77	0.84	0.86	0.88	0.88	14.3
North America	0.86	0.88	0.90	0.91	0.91	5.8
World	0.55	0.61	0.63	0.67	0.68	23.6

Note: All average values are population weighted
Source: Author’s calculations using UNDP 2014 database, accessed 20 May 2015

of social and private life in most MENA countries, which affects human development of half of the population. Youth unemployment, especially of educated youth, challenges the value of their education, which is an important element of human development.

The Human Development Index

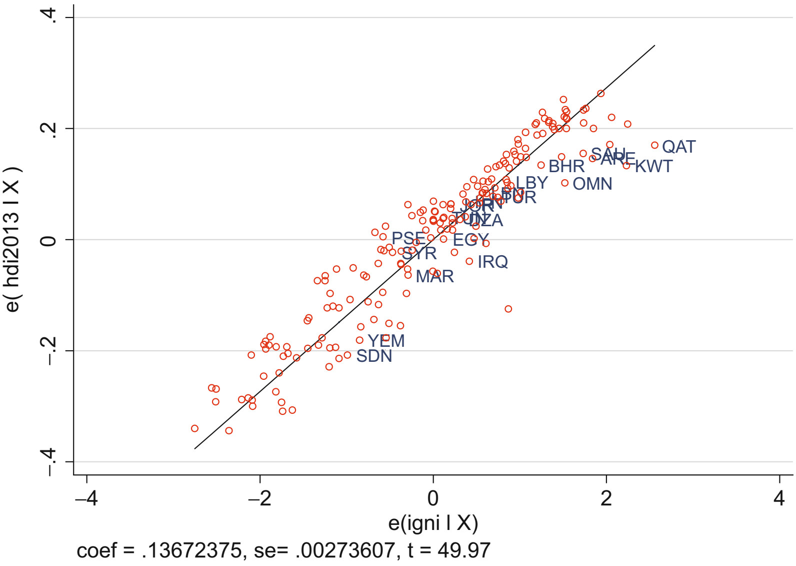
We begin with an assessment of human development in MENA countries using the Human Development Index. This index combines progress in three important dimensions of human welfare: income, health and education. Income is measured by gross national income (GNI) per capita, education by average years of schooling for adults aged 25 years and older and expected years of schooling for children of school entering age, and health by life expectancy at birth. Standardised indices for these dimensions are turned into a single index, the HDI. Since 2012, the HDI is a geometric mean of its three constituent indices.

According to HDI, as a group MENA countries have done relatively well in improving the welfare of their average citizen (see Table 2). Over the 1990–2013 period all three groups of MENA countries improved their HDI faster than the world average (23.6%). The low-income MENA group increased its HDI by 33.3%, followed by the middle-income group (29.1%) and the high-income group (23.9%). Only the Asian group of countries did as well or better, increasing their HDI score by 32.0% over the same period. HDI for Sub-Saharan Africa, which had the lowest overall score to begin with, rose by 19.5%, followed by Latin America with an increase of 17.7%, and the developed regions, which, because of their already high scores could not increase their HDI by as much as the less developed regions.

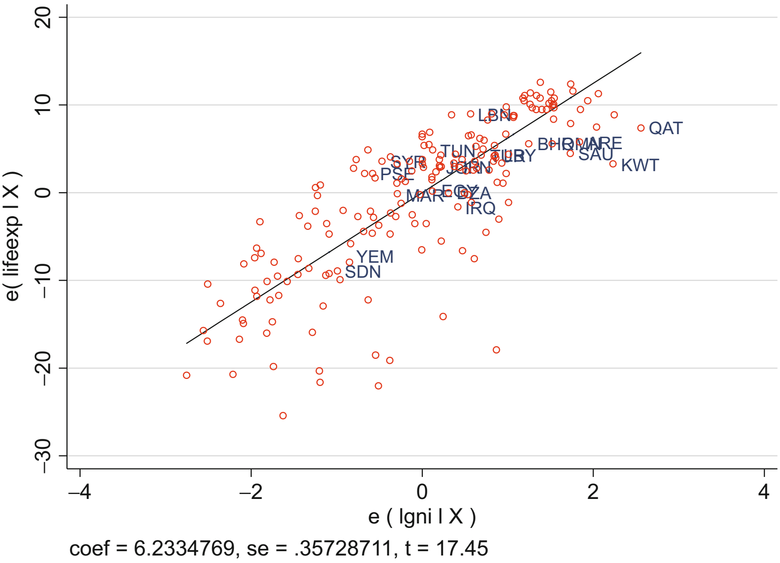
There is wide variation in the level of human development within the region, mostly caused by the considerable heterogeneity in per capita incomes, which is itself caused by differences in oil wealth. In 2013, the oil-rich group of MENA countries had a population-weighted average HDI of 0.83, which is close to those observed for the developed countries of Europe, North America and Oceania. The middle-income countries of the region do about as well as Latin America and the Caribbean, and the poorest group (Yemen and the Sudan) is slightly worse off than Sub-Saharan Africa.

A more meaningful comparison of HDI between MENA countries and the rest of the world should consider HDI relative to each country’s resources. To do this I regress the HDI values for all countries (MENA and non-MENA) on log per capita GNI, and use the predicted values forming a straight line in Fig. 1 as the basis for comparison. This comparison leaves MENA countries looking worse than in Table 2. Except for Jordan, Palestine and Syria, MENA countries have lower HDIs than is predicted by their income level.

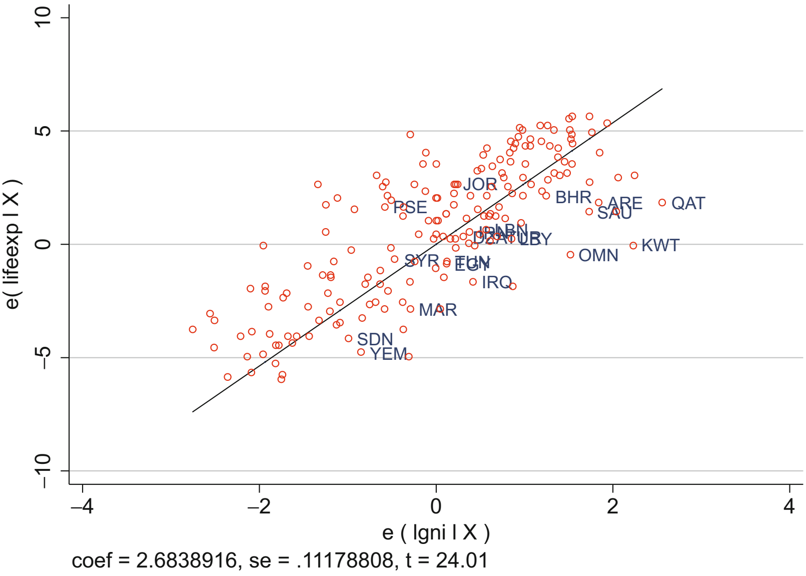
Human Development in the Middle East and North Africa, Fig. 1 HDI in MENA and the rest of the world, conditional on per capita income (Source: Author's calculations using UNDP data)



Human Development in the Middle East and North Africa, Fig. 2 MENA life expectancy compared to the world average, conditional on log of GNI per capita (Source: Author's calculations, UNDP data)



Since income is one of the three components of HDI, the underperformance of MENA countries conditional on income must be due to the other two components: health and education. Figures 2 and 3 depict the HDI values for health and education relative to the predicted line representing average performance conditional on income. Figure 2 shows the HDI measure for life expectancy at birth, a summary measure of health which is most sensitive to health outcomes at early ages, against log per capita GNI. On this count, several MENA countries perform better than the world average. Among them, Syria stands out. The position of Syria relative to the rest of MENA may surprise many in view of the destructive civil war since 2011, which is often attributed to poverty and deprivation. Syria happens to have had one of the lowest rates of infant mortality in MENA, even lower than the more developed countries of Tunisia and Turkey, thanks to an effective immunisation program and the government's focus on child health. The percentage of infants



Human Development in the Middle East and North Africa, Fig. 3 Years of schooling and GNI per capital (Source: Author's calculations using UNDP data files)

with low birth weight in 2003 was the lowest in the Arab world: 7% compared to 15% in UAE (UNDP 2005). In 2013, life expectancy in Syria was 74.6 years, which was 6.7 years higher than the average for median HDI countries with similar average income level (UNDP 2014). The case of Syria illustrates the importance of health policy in improving welfare in poorer countries as well as the need to look beyond income when assessing average wellbeing. The same view of education offers a less attractive view of MENA countries relative to the rest of the world. The HDI index for average years of schooling falls below the world average for all MENA countries except Palestine and Jordan (see Fig. 3).

Education quantity, as measured by years of schooling, is usually considered the better part of education in the MENA region (Salehi-Isfahani 2012). But attainment of years of schooling can be a gross exaggeration of a nation's level of human capital, as famously argued by Pritchett (2001). Schooling no doubt promotes human development by contributing to personal development and increasing civic participation, but it is a more effective measure of human development when it also increases individual productivity. As I argue below, the education that the majority of MENA youth receive – rote memorisation aimed at passing tests – is deficient on both grounds (UNDP 2003, p. 54). MENA countries have done well in increasing their average years of schooling, but not in education quality.

The low contribution of schooling to individual productivity can be regarded as an issue of education quality, which is, significantly, not part of the HDI, but it is the subject of much lament in the region. Two Arab Human Development Reports (2003 and 2004) that are deeply critical of the deficit in human capital in the Arab world offer insights into the problem of education quality, as well as several other empirical papers (Assaad 2014). The low productivity of education in MENA has been noted in cross-country studies that find that the growth of education in MENA does not explain the growth of output (Pritchett 1999; Makdisi et al. 2006). Evidence of low educational achievement in mathematics and science is available from international tests, such as TIMSS (Trends in Mathematics and Science Study) and PISA (Program for International Student Assessment), in which a number of MENA countries have participated, but none was able to achieve the global average (Salehi-Isfahani et al. 2014; Steer et al. 2014). The richest MENA countries in which average years of schooling is below the predicted line in Fig. 3 also have lower average scores in international tests. In TIMSS 2007, Qatar's average mathematics and science score was the second lowest in the sample of 54 countries that took the test. An obvious reason why students in oil-rich countries perform poorly despite abundance of resources for education is a lack of incentives. The ability to tap into rent income from oil with little productive skills dulls the incentive to acquire such skills. Consistent with this conjecture, boys in oil-rich MENA countries consistently score lower than girls, because they are favoured in jobs, businesses and inheritance over girls and therefore have less incentive to supply effort.

Another type of over-estimation of human development occurs when the measure of income used in HDI is not related to individual productivity. This is the case in the oil-rich group of MENA countries – the GCC, plus Algeria, Iran, Iraq and Libya – which, as we have seen, are the best performers in terms of the HDI. But not all types of income that increase the purchasing power of individuals imply the same level of increase in human capability and development. An increase in income that is the result of higher productivity as opposed to a higher price of oil is surely more congruent with a reasonable notion of human development (Salehi-Isfahani 2013). For the oil-rich MENA countries, we should keep in mind the inflated appraisal of human development that arises from the inclusion of rent income in their HDI.

Poverty and Inequality

The discussion of human development using national averages can be misleading if it is unequally distributed. Nowhere do all citizens enjoy the same average level of wellbeing, so it is important for comparison of human development across countries and over time to include measures of how average indicators are distributed across a country's population. A ready way to do this is to use the inequality-adjusted HDI (IHDI), which has been available since 2010 (Hicks 1997; Alkire and Foster 2010). Using IHDI does not substantially change the assessment of human development in the MENA region offered in the previous section. This is in part due to the fact that inequality of income in the MENA region is low compared to the average for developing regions, in particular Latin America (Bibi and Nabli 2009; Belhaj-Hassine 2015). The average Gini coefficient for 2001–2011 for MENA is 36.7, compared to 43.3 for middle-income countries to which the majority of MENA countries belong (Salehi-Isfahani 2013). More recent data, for 2014, indicates a large drop in the IHDI values for MENA countries (19 points), larger than for the average for middle-income countries (14 points). This difference appears to be caused by the drop in IHDI values for education, not health. The inequality-adjusted health index in 2014 is the same for MENA and middle-income countries (0.71), while the education index is significantly lower (0.34 compared to 0.52).

Poverty is about how the people at the bottom of the income distribution fare. In principle, it is more about the absolute level of wellbeing of those at the bottom and as such is distinct from inequality, which may be high but consistent with providing those at the bottom with a decent standard of living. MENA countries met their MDG goal of halving the proportion of the population in extreme poverty (less than $1.25 per person per day) in 2010, five years ahead of the 2015 deadline (World Bank 2015, p. 230). This achievement may not seem all that impressive in light of the high price of oil in 2010. It remains to be seen if during the coming decade, when oil prices are likely to remain low, poverty rates can continue to fall.

Inequality measured at a particular point in time (such as the Gini index calculated from cross-section data) is also relatively moderate in MENA. But as recent research has revealed, inequality of opportunity, especially in education, is alarmingly high. A study of inequality of opportunity in child health – measured by height and weight of children under 5 – by Assaad et al. (2012) found that a large proportion of the inequality in children's anthropomorphic measures is accounted for by the circumstances into which children are born, mainly the characteristics of their families and communities. Salehi-Isfahani and Vahidmanesh (2016) provide estimates of the Human Opportunity Index for 10 MENA countries. This index, measures access by children ages 10–17 to basic services, such as electricity, clean water, sanitation and schools, which are often supplied by governments. They show that children in MENA enjoy greater equality of access than other developing regions for which similar estimates are available. They attribute this to the relative success of post-independence populist and socialist governments in the Arab world whose public investment programs delivered these services more evenly to poor and rich areas.

In contrast, in areas of human development where investments by families complement public investment, as in education quality, several MENA countries exhibit high levels of inequality of opportunity. This is the case with inequality of opportunity in educational achievement estimated by Salehi-Isfahani et al. (2014) using TIMSS scores for 8th grade students, which show that in several MENA countries the amount of mathematics and science children know at age 14 depends greatly on circumstances beyond their control. In these MENA countries, parental education and the quality of the neighbourhood explained a larger share of total inequality in these countries than in Europe or even Latin America. Their estimates indicate that the chance of scoring in the top 10% of their class, a good predictor of success at the university entrance examinations, differs greatly for children from advantaged and disadvantaged families. Iran, Turkey, Tunisia and the oil-rich countries of the Gulf had higher than average inequality of opportunity within the region. Schooling attainment, measured by probability of entering school and reaching the secondary level, was also found to be opportunity unequal in that it depends greatly on circumstances beyond individual control (Assaad et al. 2014). Unlike in achievement, inequality of opportunity in attainment appears to improve with economic development; it is highest in Yemen and Iraq, the two least developed countries in a sample of seven countries, and much lower in Iran, Jordan and Tunisia.

Inequality of education opportunities is closely tied with credentialism. Some of the influence of circumstances on educational outcomes can be traced to the competitive nature of the education system in MENA countries. Having educated parents who are able to pay for private tutors and private schools gives a child a significant competitive edge to score highly in national exams and gain admission to a good public university. In Egypt, Assaad (2010) calculates that ‘an individual whose parents are both university educated, are from the highest wealth quintile and who live in the urban governorates ... has a 98.5 percent chance of accessing higher education as compared to a 5.5 percent chance for an individual whose parents are both illiterate, are from the lowest wealth quintile and live in rural Upper Egypt’. Inequality of opportunity in MENA countries appears to extend beyond education. Assaad et al. (2014) study the chances of success in employment for college graduates in Egypt and Jordan and find that the type of job and the wage graduates receive depend heavily on parental background after controlling for individual ability and education quality.

Progress in human development is closely bound with equality of opportunity and social mobility. Genuine human development requires that social and economic inequalities not persist from generation to generation. To the extent that MENA societies have failed to equalise opportunities in health, education and income, the progress they register in the inequality adjusted HDI does not fully describe how economic development of the region has benefited its citizens.

Gender Inequality

MENA societies are often characterised as patriarchal, with high levels of gender inequality in the economic, political and social spheres. The Gender Inequality Index (GII) computed by the Human Development Report 2014 quantifies gender disparities in reproductive health and access to education and paid work. The GII average for MENA is 0.49, indicating a high level of gender inequality, higher than Latin America and the Caribbean (0.43), but lower than the average values of this index for medium and low human development groups −0.51 and 0.59, respectively. The GII varies considerably within the region, which conflicts with the view that gender inequality is inherent to Islam or deeply embedded in the cultures of the region. In fact the index is smallest (0.30) for the most conservative, Islamic country in the region – Saudi Arabia. The poorest group has the highest level of gender inequality (0.67), followed by the middle-income group (0.48).

A related ranking of regions in gender inequality is available from the World Economic Forum (World Economic Forum 2015). This ranking places the Middle East sixth in the world, above only Sub-Saharan Africa, in overall gender disparity. The MENA region's ranked the last of seven regions according to the sub-index measuring labour market opportunities for women. The next worst performance was in the Political Empowerment subindex, in which MENA surpassed only Sub-Saharan Africa. Out of the 16 countries from the region, 13 were among the lowest performing countries in women's participation in the labour force and in reaching high political positions. Only in terms of gender disparity in health did the MENA region perform better, ranking fifth among the seven world regions.

The region's demographic history and its social norms have had a deep impact on the lives of its women. The recent literature on economic growth, summarised in Galor (2011), provides a useful framework for understanding the role of demographic transition in advancing the status of women and human capital. In the literature, demographic transition affects human development because it improves child health and education and empowers women. Having fewer children enables families and the state to increase their investments in health and education of children. It empowers women because it frees their time from procreation, allowing them to increase their participation in the economic and civic life of their communities. Empowerment of women further increases the allocation of family resources in the direction of investment in child health and education, and in favour of girls, all of which contribute to human development (Strauss and Thomas 1995; World Bank 2007).

Most countries of the MENA region have advanced far along the path of demographic transition, which partly explains their gender improvements in health and education. In Iran, Lebanon, Tunisia and the UAE, fertility has declined to below replacement level, while in several others the decline is well under way. Fertility transition still has some way to go among the poorest (Yemen) and the oil-rich countries (Iraq, Libya, Oman and Saudi Arabia).

The negative relationship between the level of economic development and fertility has found empirical validity globally, but the experience of a few oil-rich countries of the Middle East defies this correlation for reasons noted earlier: rising income in these countries represents increasing rent from oil rather than rising labour productivity, which is the reason for the increase in the value of parental time and lower demand for children.

Oil wealth may also help explain why rising income and female education in countries such as Oman and Saudi Arabia have failed to transform the status of women there. There is a debate about whether oil or conservative gender norms are responsible for the delayed decline in fertility and low participation of women in economic and civic life (Salehi-Isfahani 2007). Ross (2008, 2012) has argued that the incongruity between high income and gender equity is due less to traditional Islam than to the high share of oil income in the GDP.

While fertility decline requires deep social transformation, and is therefore less easily affected by the inflow of oil money, health outcomes respond more quickly to income regardless of its source: productivity or oil rent. MENA oil-rich countries have child mortality rates (CMR) at levels close to those in developed countries, 11 per thousand or below in 2005, except for Saudi Arabia (21). Low CMR in Oman and Saudi Arabia, relative to fertility, highlights the anomaly of human development in oil-rich countries. CMR has declined across the board in MENA countries, falling below the average in Asia (54) in 2005. Even in Yemen, with a CMR of 71 per thousand in 2005, child health appears better than in Sub-Saharan Africa (125).

The view that equates Islam with gender inequality ignores this variation and focuses instead on the unequal treatment of men and women under the sharia.

Naturally, the region's traditions and social norms, which consider women's role primarily to be mothers and homemakers, are closely bound with religion, but may also have their own independent origin and influence. This more complex view of gender relations in the region is corroborated by studies of history (Nashat and Tucker 1999), of modern social change (Moghadam 2004; UNDP 2005) and of other factors that affect women's role in society and economy, such as the oil rent (Ross 2008, 2012).

In several countries there has been considerable progress in changing the civil codes to better protect women's rights. Even where the legal situation has remained unchanged, women have made significant progress in health and education. Decline in fertility throughout the region has created the basis for more equal status within the Middle Eastern family. MENA countries compare favourably with the rest of the world in terms of the gender gap in school enrolment and average years of schooling (World Bank 2004b, p. 67). In Iran, Jordan, Kuwait and Tunisia women's life expectancy and average years of schooling equal or exceed those of men.

Despite these improvements, full gender equality remains elusive in the Middle East. Perhaps the most important manifestation of persisting gender inequality in MENA societies is lower participation of women in market work, a phenomenon that the World Bank (2004a) flagship report on gender refers to as the 'gender paradox'. The participation rate of women in MENA is 26%, compared to 74% for MENA men and 61% for women in middle-income developing countries. The paradox refers to the fact that the key correlates of women's labour force participation – fertility and women's education – have changed, but their labour force participation has not increased by much. In Muslim Malaysia women are three to four times as likely to engage in market work as women in MENA with similar education and fertility.

The relation between economic development and women's labour force participation is conditioned by changes in the overall structure of employment as a country develops. As Goldin (1995) has shown, participation initially falls when traditional jobs in agriculture disappear, but eventually rises as manufacturing and service sectors expand (Mammen and Paxson 2000). MENA economies have been slow to replace traditional women's jobs in agriculture with modern manufacturing and service jobs, a phenomenon that fits with the higher reservation wage due to oil income as well as with lagging social norms (Ross 2008, 2012; Salehi-Isfahani 2007).

Youth

Of the deficiencies in human development that afflict the MENA region, the failure to provide young people with a realistic transition from adolescence to adulthood is the most important, especially in explaining the region's political and social instability. The plight of the region's youth in transitions from school to work and from adolescence to adulthood is well documented in their high rates of unemployment, the long search for the first job for university graduates and the inability to get married and set up an independent family even late in their twenties (Dhillon and Yousef 2009; Egel and Salehi-Isfahani 2010).

Unlike in the rest of the world, education does not seem to improve the employment prospects of MENA youth. Widespread unemployment of university-educated youth is the biggest policy challenge in MENA, and is not limited to the resource-poor countries of the region. Region-wide unemployment of 15–24 year olds is four times the rate for adults (about twice the number in other developing countries (ILO 2012)) and rises with education (Assaad 2014; Salehi-Isfahani 2013). In 2011, average youth unemployment (aged 15–24) was 27% compared to 6.6% for adults. Young women fared worse than men, with an unemployment rate of 41%.

Several factors explain why MENA youth bear a disproportionate burden of unemployment in

Human Development in the Middle East and North Africa, Table 3 The ratio of youth (15–29) to adults (30–64)

Country	1980	1990	2000	2010	2020	2030
Algeria	1.26	1.13	1.06	0.85	0.51	0.53
Bahrain	1.18	0.76	0.65	0.63	0.47	0.36
Egypt	0.89	0.86	0.83	0.83	0.64	0.61
Iran	1.08	1.01	1.09	0.89	0.45	0.42
Iraq	1.09	1.21	1.20	1.01	0.91	0.79
Jordan	1.13	1.36	1.18	0.92	0.67	0.67
Lebanon	1.06	0.85	0.77	0.74	0.51	0.36
Libya	0.95	1.04	1.03	0.81	0.57	0.54
Morocco	1.24	1.10	0.93	0.76	0.56	0.52
Oman	0.95	0.82	1.06	1.12	0.51	0.31
Qatar	1.04	0.53	0.48	0.58	0.31	0.20
Saudi Arabia	0.96	1.02	0.97	0.66	0.48	0.47
Sudan	1.06	1.11	1.10	1.00	0.96	0.86
Syria	1.27	1.25	1.22	0.99	0.80	0.64
Tunisia	1.14	0.99	0.84	0.69	0.46	0.43
Turkey	0.98	0.87	0.82	0.66	0.54	0.47
UAE	1.04	0.62	0.67	1.09	0.29	0.26
Yemen	0.96	1.22	1.17	1.34	1.07	0.80
Total	1.04	0.99	0.97	0.84	0.60	0.55

Note: Ratios for labour importing countries include migrant workers
Source: UN Population Prospects 2012 revision

their countries. The most obvious is sluggish economic growth, especially since the Arab uprisings of the last few years. The few jobs that are created are often in the informal sector, as in Egypt, which young people who can afford to search longer shun (Assaad 2014). Another factor is their late demographic transition, which was noted in the preceding section. Thanks to high fertility in the recent past, the cohorts reaching adulthood are now facing an unfavourable job market. The ratio of youth (15–29 years old) to adults (30–64) in Table 3 shows that, until recently, youth outnumbered adults (youth to adults ratios greater than 1) even though youth ages span less than half the range for adults. Until 2000, the average youth–adult ratio for the region as a whole (weighted by adult population) was close to unity, compared to 0.4 in developed countries. Such an extreme age imbalance translates into an imbalance in labour market flows: for every one person who retires more than five enter the labour market, compared to 2 in Korea and 1.2 in the USA (Salehi-Isfahani 2012). Fortunately, this ratio is on the decline and for several countries that completed their demographic transition in the last decade it will reach less than half its value in the next 10 years (see Table 3), making absorption of youth into the labour force much easier.

Poor employment prospects for educated youth is also partly due to the low quality of human capital produced by the region's outdated education systems, which emphasise rote memorisation at the expense of skill formation. The Arab Human Development Report (UNDP 2004) blames failure of 'the knowledge systems' in the Arab countries on broader social and political factors, such as lack of democracy and inequality of wealth. A World Bank (2007) report puts the blame on the education system itself, specifically on public provision of education. But the blame can be more precisely put on the lack of incentives in the labour market for skill acquisition. The lure of academic credentials is the product of a long history of state employment of graduates (Assaad 2014; Salehi-Isfahani 2012). For decades, governments offered guarantees of public jobs to graduates, often implicitly but at times explicitly, as in Egypt and Morocco, which encouraged diploma-seeking behaviour instead of acquisition of skills. MENA public sectors still employ the largest proportion of the workforce in the world (Schiavo-Campo et al. 1997; Salehi-Isfahani 2007) and their ratio of the government wage bill to GDP (averaging 8.4% in for 2000–2008) is four times the share for all developing countries and twice that of high-income countries (Salehi-Isfahani 2013).

In recent decades, as state bureaucracies have filled up and structural adjustments rolled back the size of public sectors (Assaad and Barsoum 2009), graduates of universities have had to seek jobs in private sector, often only finding informal jobs there. But the value of credentials in the labour market as well as in the marriage market remains high, casting a long shadow on the region's educational institutions. The returns to education, which in developed countries are relatively constant across years of schooling – about 10% increase in wages per year of schooling (Card 1999) in MENA countries is heavily concentrated in the last stage – the university (Salehi-Isfahani et al. 2009). As a result, the share of workers with a university degree is incongruent with the demands of the labour market. For example, in Egypt the share of urban wage workers with tertiary degrees has risen to 29%, much higher than the more economically advanced Turkey, with 11%. The lure of university education means that a high share of the public budget is spent at the university level, which is much more expensive than basic education. It also leads educated youth to have excessive expectations in terms of the wage and the type of jobs the are likely to get, and this delays their integration into the labour market and causes longer queues and waiting times for public and formal private sector jobs.

In addition to the pains of unemployment after graduation, the lack of formal sector jobs also frustrates youth aspirations for marriage and forming independent families. The problems of the labour market thus easily spill over into the marriage market, complicating youth transitions (Assaad et al. 2010; Salehi-Isfahani and Dhillon 2008). As a result, age at marriage has increased sharply in MENA countries, and not all voluntarily. Voluntary delay in marriage, for example as a result of the decline in demand for children and increased demand for education, is consistent with human development and increased welfare. But when it is involuntary the opposite may be true, especially in view of social and legal taboos on sexual relations outside marriage in MENA societies. A contributing factor is the high cost of housing, which forces young people to delay marriage and family formation (Salehi-Isfahani and Dhillon 2008). Assaad and Ramadan (2008) argue that the recent decline in the age at first marriage of Egyptian men is in part due to increased access to housing resulting from a 1996 housing market reform that made rental housing more readily available to young men. Singerman (2007) blames the high cost of marriage due to social conventions regarding dowries and marriage ceremonies.

Conclusion

In the past three decades, the overall performance of the region in core components of human development, as defined by the Human Development Index, has been quite respectable. But this general picture has at least four caveats. First, in terms of income per capita the region is one of the most heterogeneous in the world, including the world's richest country, Qatar, and one of the poorest, Yemen. To take account of this heterogeneity, in discussions of the HDI I divided the region into three more homogeneous groups, and considered human development sub-indices in relation to per capita income. Accounting for income per capita, MENA countries performed at about the world average in terms of human development.

Second, the standard measures of human development do not apply well to MENA countries. I have singled out income and education for their lack of relevance to individual productivity and capability. Income from the oil rent affords a high level of purchasing power, at least when oil prices are high, but fails to imply the other aspects of human development we usually associate with economic development, such as improved gender equality and increased democratic rights. Likewise, the education that most MENA youth receive is but a caricature of the human capital that is needed to remain competitive in the global economy. Counting years of schooling may give MENA countries a high score in HDI calculus, but it does not imply the same level of human development.

Third, the progress of human development in the MENA region has been unequal. Evidence on inequality of opportunity shows that life chances in health and education for the children from poor and rich backgrounds are vastly different. Gender inequality remains the region's most glaring failure, and its youth do not share in the region's prosperity.

Fourth, progress in measured human development has not brought commensurate improvements in civic and political freedoms. Across the region the 'freedom deficit' remains large, both in countries that experienced the Arab Spring and those that did not. These less quantifiable aspects of human development, which I did not cover in this survey, may in the years ahead determine the extent to which past gains in human development are preserved and have the chance to expand.

See Also

- ▶ Economic Growth
- ▶ Growth and Inequality (Macro Perspectives)
- ▶ Inequality Between Nations
- ▶ Inequality (Global)
- ▶ Inequality (International Evidence)
- ▶ Inequality (Measurement)
- ▶ Labour Markets in the Arab World
- ▶ Oil and Politics in the Gulf: Kuwait and Qatar

Bibliography

Alkire, S., and J. Foster. 2010. Designing the inequality-adjusted Human Development Index (HDI). OPHI Working Papers, Queen Elizabeth House, University of Oxford.

Assaad, R. 2010. Equality for all? Egypt's free public higher education policy breeds inequality of opportunity. Policy Perspective No. 2, Economic Research Forum, Cairo.

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

Assaad, R., and G. Barsoum. 2009. Rising expectations and diminishing opportunities for Egypt's young. In *Generation in waiting. The unfulfilled promise of young people in the Middle East*, ed. N. Dhillon and T. Yousef, 67–94. Washington, DC: Brookings Institution Press.

Assaad, R., and M. Ramadan. 2008. Did housing policy reforms curb the delay in marriage among young men in Egypt. Middle East Policy Initiative Policy Outlook. The Wolfensohn Center for Development at Brookings and the Dubai School of Government.

Assaad, R., C. Binzel, and M. Ghadallah. 2010. Transitions to employment and marriage among young men in Egypt. *Middle East Development Journal* 2(1): 39–88.

Assaad, R., C. Kraft, N.B. Hassine, and D. Salehi-Isfahani. 2012. Inequality of opportunity in child health in the Arab world and Turkey. *Middle East Development Journal* 4(2): 3–40.

Assaad, R., C. Kraft, and D. Salehi-Isfahani. 2014. Inequality of opportunity in the labor market for higher education graduates in Egypt and Jordan. Paper presented at IEA, Jordan.

Belhaj-Hassine, N. 2015. Economic inequality in the Arab region. *World Development* 66: 532–556.

Bibi, S., and M.K. Nabli. 2009. Income inequality in the Arab region: Data and measurement, patterns and trends. *Middle East Development Journal* 1(2): 275–314.

Card, D. 1999. The causal effect of education on earnings. In *Handbook of labor economics*, ed. O. Ashenfelter and D. Card, vol. 3, 1801–1863. Amsterdam: Elsevier.

Dhillon, N., and T. Yousef, eds. 2009. *Generation in waiting: The unfulfilled promise of young people in the Middle East*. Washington, DC: Brookings Institution Press.

Egel, D., and D. Salehi-Isfahani. 2010. Youth transitions to employment and marriage in Iran: Evidence from the school to work transition survey. *Middle East Development Journal* 2(1): 89–120.

Galor, O. 2011. *Unified growth theory*. Princeton: Princeton University Press.

Goldin, C. 1995. The U-shaped female labor force function in economic development and history. In *Investment in Women's Human Capital*, ed. T.P. Schultz. Chicago: University of Chicago Press.

Hicks, D.A. 1997. The inequality-adjusted human development index: A constructive proposal. *World Development* 25(8): 1283–1298.

ILO. 2012. *Global employment trends 2012: Preventing a deeper jobs crisis*. Geneva: International Labor Organization.

Makdisi, S., Z. Fattah, and I. Limam. 2006. Determinants of growth in the MENA region. In *Explaining growth in Middle East and North Africa*, ed. J. Nugent and M.H. Pesaran. London: Elsevier.

Mammen, K., and C. Paxson. 2000. Women's work and economic development. *Journal of Economic Perspective* 14(4): 141–164.

Moghadam, V.M. 2004. Patriarchy in transition: Women and the changing family in the Middle East. *Journal of Comparative Family Studies* 35(2): 137–162.

Nashat, G., and J.E. Tucker. 1999. *Women in the Middle East and North Africa: Restoring women to history*. Bloomington: Indiana University Press.

Pritchett, L. 1999. Has education had a growth payoff in the MENA region? MENA Working Paper Series, No. 18. World Bank, Washington, DC.

Pritchett, L. 2001. Where has all the education gone? *World Bank Economic Review* 15(3): 367–391.

Ross, M.L. 2008. Oil, Islam and women. *American Political Science Review* 102(1): 107–123.

Ross, M.L. 2012. *The oil curse: How petroleum wealth shapes the development of nations*. Princeton: Princeton University Press.

Salehi-Isfahani, D. 2007. Microeconomics of growth in MENA: The role of households. In *Explaining growth in Middle East and North Africa*, ed. J. Nugent and M.H. Pesaran. London: Elsevier.

Salehi-Isfahani, D. 2012. Education, jobs, and equity in the Middle East and North Africa. *Comparative Economic Studies* 54(4): 843–861.

Salehi-Isfahani, D. 2013. Rethinking human development in the Middle East and North Africa: The missing dimensions. *Journal of Human Development and Capabilities* 14(3): 341–370.

Salehi-Isfahani, D., and N. Dhillon. 2008. Stalled youth transitions in the Middle East: A framework for policy reform. Wolfensohn Center for Development Working Paper, The Brookings Institution.

Salehi-Isfahani, D., and A. Vahidmanesh. 2016. Human opportunities in the Middle East and North Africa. ERF Working Paper.

Salehi-Isfahani, D., I. Tunali, and R. Assaad. 2009. A comparative study of returns to education in Egypt, Iran and Turkey. *Middle East Development Journal* 1(2): 145–187.

Salehi-Isfahani, D., N. Belhaj-Hassine, and R. Assaad. 2014. Inequality of opportunity in educational achievement in the Middle East and North Africa. *Journal of Economic Inequality* 12(4): 489–515.

Schiavo-Campo, S., G. De Tommaso, and A. Mukherjee 1997. Government employment and pay: A global and regional perspective. World Bank Policy Research Working Paper No. 1771.

Singerman, D. 2007. The economic imperatives of marriage: Emerging practices and identities among youth in the Middle East. Middle East Youth Initiative Working Paper 6, Brookings Institution, Washington, DC.

Steer, L., H. Ghanem, and M. Jalbout. 2014. *Arab youth: Missing educational foundations for a productive life?* Washington, DC: Center for Universal Education, The Brookings Institution.

Strauss, J., and T. Duncan. 1995. Human resources: Empirical modeling of household and family decisions. In *Handbook of development economics*, ed. H. Chenery and T.N. Srinivasan, 1883–2023. London: Elsevier.

UNDP. 2002. *The Arab human development report 2002*. New York: UNDP.

UNDP. 2003. *The Arab human development report 2003*. New York: UNDP.

UNDP. 2004. *The Arab human development report 2004*. New York: UNDP.

UNDP. 2005. *The Arab human development report 2005*. New York: UNDP.

UNDP. 2010. *The human development report 2010*. New York: UNDP.

UNDP. 2014. *The human development report 2014*. New York: UNDP.

World Bank. 2004a. *Gender and development in the Middle East and North Africa: Women in the public sphere*. Washington, DC: World Bank.

World Bank. 2004b. *Unlocking the employment potential in the Middle East and North Africa*. Washington, DC: World Bank.

World Bank. 2007. *The road not travelled: Education reform in the Middle East and North Africa*. Washington, DC: World Bank.

World Bank. 2015. *Development goals in an era of demographic change, global monitoring report 2015/2016*. Washington, DC: World Bank.

World Economic Forum. 2015. Global gender gap report 2015. Available from: http://www.weforum.org/reports/global-gender-gap-report-2015

Humbug Production Function

Anwar Shaikh

Neoclassical economics has always tried to portray wages and profits as mere technical variables. At an aggregate level, this is accomplished by connecting labour and capital to output through a 'well-behaved' aggregate production function, with the marginal products of labour and capital equal to the wage rate and profit rate, respectively. Thus in competitive equilibrium each social class is pictured as receiving the equivalent of the marginal product of the factor(s) it owns (Shaikh 1980).

The original optimism that aggregate production functions and their corresponding marginal productivity rules could be derived from more detailed general equilibrium models eventually gave way to the sobering realization that the conditions for any such a derivation were 'far too stringent to be believable' (Fisher 1971). Yet neoclassical economists continue to use aggregate production functions, apparently because they seem to fit the data well and their estimated marginal products closely approximate the observed wage and profit rates (so-called factor prices).

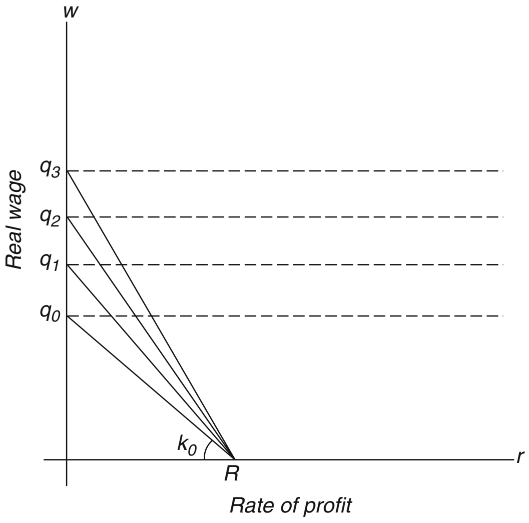
This apparent empirical strength of aggregate production functions is often interpreted as support for neoclassical theory. *But there is neither theoretical nor empirical basis for this conclusion*. We already know that such functions cannot be derived theoretically, except under conditions which neoclassical theory itself rejects (e.g. the simple labour theory of value) (Garegnani 1970). Moreover, Fisher (1971) discovered through simulation studies that the aggregate data generated by microeconomic production functions were not generally well fitted by aggregate production functions; that the functions which did best fit this data are not neoclassical in nature (this is a common finding, e.g. Walters 1963); and that in simulation runs where the wage share happened to be roughly constant and aggregate Cobb–Douglas production functions happened to work well, this goodness of fit was puzzling because it held even when the theoretical conditions for aggregate production functions were flagrantly violated.

Shaikh (1974, 1980) has shown that this last result is simply an artifact of the constancy of the wage share. To see this, let r_t represent the rate of profit, and q_t, w_t, k_t the per worker net output, wages and capital, respectively, all at time t. Then the national accounting identity $q_t = w_t + r_t k_t$ can be differentiated to yield percentage rates of change q', w', etc., weighted by the profit share $s_t = r_t k_t / q_t$ and the labour share $1 - s_t = w_t / q_t$:

$$q'_t = B'_t + s_t k'_t \quad \text{where} \quad B'_t = (1 - s_t) w'_t + s_t r'_t. \tag{1}$$

The preceding relation says nothing about the nature of the underlying economic processes, since it is derived from an identity. But if social forces happen to produce a stable profit (and hence wage) share, so that $s_t = s (a$ constant), we can immediately integrate both sides of (1) to get

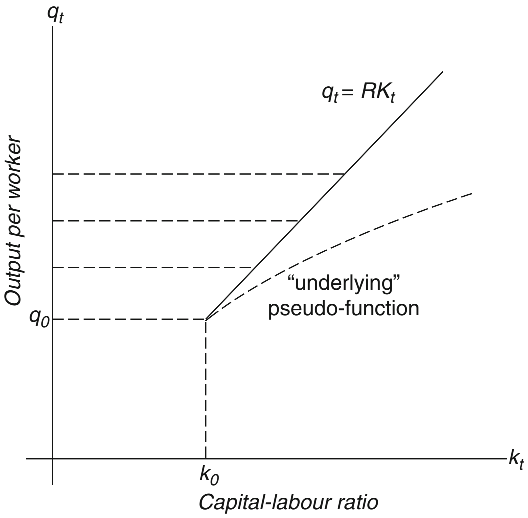
$$q_t = A_t k_t^s, \quad \text{where} \quad A_t = Ce^{\int B'_t \mathrm{d}t}, C = a \ \text{constants}. \tag{2}$$



Humbug Production Function, Fig. 1

Equation (2) looks like an aggregate Cobb–Douglas production function with constant returns to scale, marginal products equal to factor prices, and a technical change shift parameter A_t. It will even seemingly reflect neutral technical change if the rate of change B'_t can be expressed as a function of time. And yet *it is not a production function at all, but rather merely the algebraic expression of any social forces resulting in a constant share – even when the underlying processes are definitely not neoclassical in nature*. To illustrate this, we will now demonstrate that even a very simple 'anti-neoclassical' (Robinsonian) economy will fit such a function.

Consider an economy at time t_0, in which all possible techniques of production are dominated by a *single* linear technique (linear because capital–labour ratios are equal across all sectors). With one dominant technique, there is no neoclassical substitutability among techniques, and the linear wage–profit curve of the dominant technique is also the wage–profit frontier for the whole economy (the line q_0R in Fig. 1, for the given time period). Because q, k and R (net output/capital) are all *constant* along the wage–profit frontier, the marginal products of labour and capital therefore cannot even be defined. The determination of the so-called factor prices w and r cannot possibly be tied to some corresponding



Humbug Production Function, Fig. 2

marginal products. Lastly, because q and k are constant for any given frontier, a frontier such as q_0R in Fig. 1 contributes only a single point q_0k_0 to the q_t, k_t space in Fig. 2.

Now consider Harrod-neutral technical change, in which both output per worker q, $_t$ and the capital–labour ratio k, rise at the same rate, so that the output–capital ratio R remains constant:

$$q_t/q_0 = k_t/k_0 = \mathrm{e}^{at}, \quad \text{and since} \quad q_t/q_0 = R, q_t/k_t = R \tag{3}$$

This is depicted in Fig. 1 by the successive wage–profit frontiers and in Fig. 2 by the corresponding (solid) straight line q_t of slope R.

If we were simply concerned with the best relation between inputs and output, then the *true* relation $q_t = Rk_t$ would be the correct one. But within neoclassical theory, such a fitted function would imply a constant marginal product of capital, a zero marginal product of labour (Alien 1968, pp. 45–6), and no technical change (since the 'shift parameter' R is constant). A good neoclassical would therefore have to reject this best (and true) fitted function in favour of some more 'appropriate' functional form (Fisher 1971, pp. 312–13). How then might an aggregate production function fare in our anti-neoclassical world?

We have already assumed a constant profit share $r_tk_t/q_t = s$, and since the output–capital ratio $q_t/k_t = R$ is constant (Eq. (3), it follows that the rate of profit $r_t = sR$ is constant. Similarly, the assumption of a constant wage share $w_t/q_t = 1 - s$ and a steadily growing output per worker $q_t = q_0\ \mathrm{e}^{at}$ (Eq. 3), implies a steadily growing real wage $w_t = (1 - s)\ q_0\mathrm{e}^{at}$. All this allows us to solve explicitly for B'_t and A_t in Eqs. (1) and (2):

$$B'_t = (1 - s)w'_t + sr'_t = (1 - s)a \tag{4}$$

$$q_t = C\mathrm{e}^{(1-s)at}k_t^s, \quad \text{since} \quad A_t = C\mathrm{e}^{(1-s)at} \tag{5}$$

Thus when the wage share is constant, even a fixed proportion technology undergoing Harrod-neutral technical change is perfectly consistent with an aggregate pseudo-production function (Eq. 5). This is, however, a law of algebra, not a law of production. The above reasoning has been shown to have grave implications for production function studies (Shaikh 1980). For instance, Solow's (1957) so-called seminal technique for assessing technical change amounts to decomposing the true production relation into an 'underlying' pseudo-production function and a residual $\mathrm{A_t}$ whose rate of change is then taken to measure technical progress (Fig. 2). But this measures nothing more than distributional changes, since $\mathrm{B_t}$ is simply the weighted average of the rates of change of observed wage and profit rates (Eqs. 1 and 2). Similarly, Fisher's previously mentioned puzzle concerning the empirical strength of aggregate Cobb–Douglas production functions can be shown to be an artifact of the stability of the wage share over those particular simulation runs. Last, and perhaps most strikingly, it is interesting to note that even data points which spell out the word 'HUMBUG' can be well fitted by a Cobb–Douglas production function apparently undergoing neutral technical change and possessing marginal products equal to the corresponding 'factor prices'! Surely there is a message in this somewhere?

See Also

▶ Cobb–Douglas Functions

Bibliography

Allen, R.G.D. 1968. *Macro-economic theory: A mathematical treatment*. London: Macmillan.

Fisher, F. 1971. Aggregate production functions and the explanation of wages: A simulation experiment. *Review of Economics and Statistics* 53(4): 305–325.

Garegnani, P. 1970. Heterogeneous capital, the production function, and the theory of distribution. *Review of Economics and Studies* 37(3): 407–436.

Shaikh, A. 1974. Laws of algebra and laws of production: The humbug production function. *Review of Economics and Statistics* 51(1): 115–120.

Shaikh, A. 1980. Laws of algebra and laws of production: The humbug production function II. In *Growth, profits and property: Essays on the revival of political economy*, ed. E.J. Nell. Cambridge: Cambridge University Press.

Solow, R. 1957. Technical change and the aggregate production function. *Review of Economics and Statistics* 39: 312–320.

Walters, A.A. 1963. Production functions and cost functions: an econometric survey. *Econometrica* 31(1): 1–66.

Hume, David (1711–1776)

Eugene Rotwein

Keywords

Capital supply; Class; Economic growth; Economic individualism; Economic psychology; Excise taxes; Free trade; Gains from trade; Happiness; Historical economics; Human nature; Hume, D.; Institutional economics; International trade; Invisible hand; Laws of human behaviour; Liberty; Luxury; Multiplier; Natural history; Optimum taxation; Probability; Psychological hedonism; Public debt; Quantity theory of money; Rate of interest; Smith, A.; Specie-flow mechanism; Tax incidence; Utilitarianism; Utility of poverty doctrine

JEL Classifications

B31

David Hume's economic essays (which originally appeared in 1752 in a volume entitled *Political Discourses)* comprise a small portion of his writings. The scope of Hume's thought was vast. He wrote extensively in philosophy (the area in which his reputation primarily lies), explored several of the social sciences and the humanities, and was deeply interested in history. His multi-volume *History of England* (1754–1761) was a pathbreaking work in the field. Nonetheless, in the literature Hume's economic writings have typically been treated as an entirely self-contained aspect of his work. This is not surprising, since in his economic essays he does not allude to his other writings, and subsequent disciplinary specialization has not encouraged consideration of any interrelationships between the two. For their part, philosophers have often treated Hume's philosophical writings in isolation from his other work.

For Hume, however, there was no such sharp disjunction. In the Advertisement prefixed to his first and major philosophical work, *A Treatise of Human Nature* (1739), he states that he expects his philosophy to serve as the 'capital or centre' of all the 'moral' (that is, psychological and social) sciences and that he hopes to expand the *Treatise* to accommodate a study of these areas. Owing perhaps to the poor reception accorded his *Treatise*, Hume did not carry out his original intention. His treatment of the moral sciences was left mainly to his essays. But there are many links between Hume's philosophical thought and his essays, and this is true with respect to his economic essays. Indeed, in light of the importance of these links, Hume may be regarded as the outstanding philosopher economist of the 18th century.

Viewed in most general form, what is the nature of the relationship between Hume's economic and philosophical thought? Hume regarded the foundation of his entire philosophical system – its 'capital or centre' – as a body of 'principles of human nature', or elements and relations concerning human understanding and human passions that he believed to be irreducible and universal. These principles, which constitute the analytical phase of Hume's system of thought, are treated in Books I and II of the *Treatise.* In the second and synthetic phase Hume then relates various aspects of 'human nature' to environmental forces in seeking to frame laws of human behaviour, or generalizations indicating how man may be expected to behave under different specific conditions. These generalizations comprise the substance of the 'moral sciences' with which, as indicated, Hume dealt principally in the essays. An explicit and deep interest in psychology is thus a salient characteristic of Hume's treatment of the 'moral sciences' in general, and this is conspicuously evident in his economic analysis.

What were Hume's views concerning the prospects of developing reliable generalizations in the 'moral sciences?' That Hume should have distinct views on this issue is scarcely surprising in light of the depth of his interest, as a philosopher, in the epistemological basis of science. As he had argued, the contrary of any generalization concerning relations between matters of fact is always conceivable and hence always possible. Consequently, the only way of developing an understanding of these relations, he contended, is through empirical observation; and this can only yield probabilities, never certainty. With respect to his own principles of human nature, Hume believed that his propositions carried the highest order of probability because of the abundance of evidence on which they rested.

On the other hand, recognizing the complexity of the interrelationships between man's 'nature' and his environment, he stressed the difficulty in framing valid laws of human behaviour. He calls attention to the effect on human behaviour of imperceptible influences, emphasizes the extent to which it could be altered by changing conditions and notes the impracticality of conducting controlled experiments in the realm of psychological phenomena. He thus warns that in the social sciences 'all general maxims ... ought to be established with the greatest caution' and states that 'I am apt ... to entertain a suspicion that the world is still too young to fix many general truths in [the area of the social sciences] which will remain true to the latest posterity' (Hume 1875, vol. 3, pp. 156–7). Of all the social fields, however, he believed that a field such as economics lent itself especially well to scientific study, and here he was cautiously optimistic concerning the possibility of developing reliable generalizations through direct observation of man in the course of his day-to-day affairs. As he argued, behaviour here was governed by mass passions, which were 'gross' or 'stubborn', or were not as affected by imperceptible influences as passions governing the behaviour of small numbers of individuals. Uniformities in behaviour therefore could here be more readily discerned (1875, vol. 3, p. 176). It should be noted that, in accord with this view, Hume introduces his economic essays by contrasting the potential for scientific analysis in economics with the very limited prospects for such analysis in a field such as foreign diplomacy, where events are controlled by the behaviour of a small number of individuals (1955, pp. 3–4).

To return to the substance of Hume's economic thought, in addition to emphasizing psychological considerations Hume's analysis displays a deep interest in historical sequence. Hume's interest in history developed at a very early age, even before he undertook his *Treatise.* As it appears in his essays, however, his treatment of history differs from conventional historiography (with its concern with unique particulars) which predominates in his *History of England.* For, writing as a 'moral scientist', Hume sought to reduce historical sequence to generalizations which explain how transformations in human behaviour result from the impact of changing historical circumstance on 'human nature'. This type of study (which bore a relationship to the 'conjectural history' and the French *'histoire raisonée'* of the period) Hume termed 'natural history' – the term 'natural' here denoting the recurrent or probable, or the substance of laws of human behaviour. There are clusters of what Hume regards as historical laws of human behaviour in several of the essays. One essay bears the title 'The Natural History of Religion'. And in the economic essays the approach of 'natural history' is of fundamental importance.

This can be seen when Hume's economic essays are viewed on three different levels of analysis. The first is economic psychology, where Hume deals with economic motivation, or what he terms the ‘causes of labour’. This is the most basic level of his economic analysis in the sense that here one finds the links between his economic thought and his treatment of ‘human nature’ in the *Treatise.* On this level the analysis takes the form of a natural history of ‘the rise and progress of commerce’. In a word, Hume introduces the question of economic motivation in seeking to explain how changing environmental influences stimulated the economic growth of his general period through their impact on various human passions. Here Hume observes that there are four ‘causes of labour’ – the desire for consumption, the desire for action, the desire for liveliness and the desire for gain.

The first of these, which is commonly stressed by economists, simply denotes all the wants that may be gratified by consumption. The desire for action refers to a desire for challenging activity as such. However, its full effectuation, as Hume stressed, requires activity whose end or objective has independent value. Like hunting and gaming, economic pursuits (and especially the activities of the merchant and, more generally, the ‘industrious professions’) are seen as meeting these conditions. By the desire for liveliness Hume meant the desire for the experience of active passion as such (which he contrasts with a state of no passion, or in effect a state of waking sleep). This is not a completely independent cause of labour but is an important ingredient common to both consumption and interesting activity. The last cause of labour is the desire for monetary gain, which is a desire to accumulate the tokens of success in the economic ‘game’.

Hume argues that all these motives play a role in a nation’s economic growth – the initial stimulus to which he finds in the expansion of international trade. As compared with the treatments of economic motivation by economists (which commonly accord exclusive or over shadowing emphasis to the desire for consumption), a striking characteristic of Hume’s treatment lies in its multidimensionality. This multidimensionality is also found in Hume’s criticism of the doctrine of psychological hedonism. Here he argues that, in addition to seeking pleasure, man is driven by a variety of ‘instincts’ which lead him to do things for their own sake, and therefore will not automatically lead him to act in his own best interests. Hume’s position thus precludes any simple identification of wealth with welfare.

The second level of Hume’s economic analysis is his political economy, or his treatment of market relations. It is this which makes up the bulk of his economic essays. Here Hume considers several of the major economic issues of his own period, including monetary theory, interest theory, the question of free versus regulated trade, the shifting and incidence of taxes, and fiscal policy. In this context the natural history of ‘the rise and progress of commerce’ plays a dominant role. For repeatedly in his critical treatment of the economic doctrines of his period Hume seeks to show that their major deficiency lies in a failure to give proper attention to the importance of economic growth and to the underlying psychological and other factors associated with this growth process.

Let us consider first Hume’s quantity theory specie flow doctrine, which he presents (in the essay ‘Of the Balance of Trade’) in criticism of the mercantilist view that without restraints on international trade a nation would suffer losses in its money supply. Hume’s position, which has been recognized as an early anticipation of the classical view, is that, owing to the effects of specie flows on price levels in trading nations, the amount of specie in each automatically tends towards an equilibrium at which its exports and imports are in balance. Any attempt through restraints on trade to increase the amount of specie beyond this equilibrium level, as Hume argues, is destined to fail (on the assumption that the money circulates domestically) because the specie movement from abroad will raise the nation’s prices relative to those abroad, reduce exports and increase imports, and generate a return outflow of specie.

The relationship of this analysis to Hume’s historical perspective is evident in the purpose with which he introduces this doctrine. For in employing the quantity theory of money he is here arguing that the extent to which a specie inflow into a nation affects its prices depends on its total output. Consequently, as he is seeking to show, it is the level of a nation's economic development, or its productive capacity as determined by its population and the spirit of industry of its people, that controls the amount of specie a nation can attract and retain. As he states, 'I should as soon dread that all our springs and rivers should be exhausted as that money should abandon a kingdom where there are people and industry' (1955, p. 61).

To consider another of Hume's anticipations of the classical position – his interest theory presented in his essay 'Of Interest' – here he attacks the mercantilist view that the rate of interest is determined by the money supply. On quantity theory grounds he argues that an increased money supply will simply raise all prices and, necessitating an offsetting increased demand for loans to finance expenditures, will leave interest rates unaffected. It is therefore the supply of real capital that determines interest rates. The bulk of Hume's discussion, however, is concerned with the factors affecting the supply of real capital itself; and here he turns to a historical analysis in which he considers the effect of economic growth on the class structure of society and, through this, on economic incentives. In this context every 'cause of labour' considered in the natural history of 'the rise and progress of commerce' is brought into his treatment. In a feudal society, he points out, the supply of capital is low because there are only two classes – the peasants and the landed aristocracy. The peasants cannot save since they are poor. On the other hand, the landed aristocracy tend to be heavy borrowers. For, as they are idle and lack the sense of liveliness that interesting activity affords, they seek liveliness wholly through extravagant consumption expenditures. Capital is therefore scarce and interest rates are high. Economic development, however, spawns the growth of the merchant class and the industrious professions. These groups derive a sense of liveliness from economic activity. Consumption expenditure drops for this reason and also because the pursuit of profit nourishes a desire to accumulate gain as a token of success in the economic game. As the new industrious classes earn a substantial share of the growing national income, their disposition to save thus results in a significant increase in the capital supply and a decline in interest rates.

As noted, Hume employs the quantity theory of money in criticizing the mercantilist position. But Hume's monetary theory also exhibits a similarity to the mercantilist view. However, his treatment here too springs from an attempt to call attention to the importance of economic growth. Thus (in his essay 'Of Money') Hume – assuming a condition of less than full employment – grants that an increase in the quantity of money (as against a greater absolute quantity of money as such) need not simply raise prices but can stimulate economic activity. Here, in tracing the impact of the increased money supply as it courses through the economy, he presents a lucid description of the multiplier process. He denies, however, that the stimulating effect on industry – when resulting from a short-run increase in the money supply – can prove anything more than ephemeral. No justification for this view is given. But it serves to underscore the conclusion of his analysis. For he goes on to argue that, if the increase in the money supply is gradual and continues over a long period of time, its stimulating effects on output will prove enduring because it will nourish the 'spirit of industry' and therefore economic growth itself. Similarly, although Hume argued that an increase in the money supply does not affect interest rates, near the conclusion of his essay 'Of Interest' he points out that a long-run increase in the supply of money, by stimulating economic growth and inducing a change in spending and saving patterns, can increase the supply of capital and lower the interest rate.

Another noteworthy area of Hume's analysis is his treatment of the issue of free versus regulated markets. Since the relevant comments are not found in his economic essays but rather lie scattered through his *History of England,* the full extent to which Hume anticipated Adam Smith's 'invisible hand' argument has not been generally recognized. These comments make clear that Hume understood the role of a free price mechanism is governing the allocation of resources (1955, pp. 1xxviii–1xxx).

In applying the argument for free markets to the case of international trade, Hume emphasizes that free trade makes it possible for nations to enjoy the gains from an exchange of the products of their different resource endowments. However, in his most thorough treatment of the issue of international free trade (in his essay ‘Of the Jealousy of Trade’) it is not this static approach to the question that predominates. Rather, once again, it is economic growth considerations that receive primary emphasis. For here, where Hume seeks to meet the mercantilist argument that foreign economic development adversely affects home industry and employment, he takes the position that expansion abroad, on the contrary, commonly promotes economic development at home. By increasing foreign income, he argues, economic growth abroad not only leads to an expansion of foreign demand for domestic output but, through an emulation of foreign technological innovations, promotes the advance of technology at home. Hume goes on to argue that even when foreign expansion competes with domestic output, there is no need for concern provided the nation’s ‘spirit of industry’ – which is itself nourished by foreign trade – is preserved. For as long as a nation remains industrious it need not fear that other nations will encroach on the market for its staple and, even in the unlikely event that this does occur, an industrious nation can readily divert its resources to other uses. Moreover, in stimulating the spirit of industry, foreign trade also promotes the diversification of a nation’s resource use, and so reduces the impact of any shrinkage of demand that may occur from time to time in particular markets.

There are indications that Hume was more fully aware of the possible costs of free trade than one would gather from the main argument in the essay ‘Of the Jealousy of Trade’. Elsewhere he treats the interests of poor and rich countries as incompatible, and in one place he also justifies the use of a tariff in specific cases (1955, pp. 34–5, 76, 199–205). In the essay ‘Of the Jealousy of Trade’ itself he recognizes, in a modification of his main argument, that there are circumstances in which a nation facing a loss of markets to foreign countries may find resource diversion difficult (1955, p. 81). The character of this essay as a whole (which appeared six years after the other economic essays) suggests, however, that after much reflection and groping Hume had concluded that free trade would have a markedly favourable effect on long-term economic growth for all nations, and that, with this end in view, any associated costs – which would be of a shorter-term nature – would be well worth sustaining.

A further illustration of the role of natural history in Hume’s political economy is found in his treatment of the shifting and incidence of taxes (in his essay ‘Of Taxes’), where he considers the view that an expansion of taxes creates an expanded ability to pay the levies by increasing ‘proportionably the industry of the people’. This view was commonly held by the mercantilists and, in what came to be known as ‘the utility of poverty’ doctrine, was employed to justify the imposition of excises on goods consumed by the poor. Hume’s position here is twofold. He points out that history shows that natural burdens, such as relatively infertile soil, often stimulate industry, and he argues that artificial burdens such as taxes may have the same effect. This position springs from Hume’s view concerning the importance of a desire for interesting action as a ‘cause of labour’ since he here emphasizes that in order to prove interesting the activity must be difficult and challenging. On the other hand, he emphasizes that, since economic activity is also motivated by a desire for consumption, increasing difficulty beyond a certain level in achieving consumption ends will lead to despair. From the viewpoint of its stimulating effect on industry there is thus an optimum tax level, and Hume takes the view that taxes on the poor throughout Europe have already so substantially exceeded that optimum that they are threatening to ‘crush all art and industry’. Considered as a whole, Hume’s position represents an amalgam of both the mercantilist and the later classical view. He rejects the mercantilist ‘utility of poverty’ doctrine with its unqualified endorsement of higher taxes on goods consumed by the poor, but also would reject the view (which is based on the subsistence or accustomed standard of living theory of wages found in the writings of Smith and Ricardo) that any tax on labour would inevitably result in a reduction in its supply.

Hume's treatment of fiscal policy – the last major aspect of his political economy – does not reveal significant relationships to his natural history of the rise and progress of commerce. Owing to space limitations, his analysis – contained in the long essay 'Of Public Credit' – cannot here be considered in detail. It should be observed, however, that this essay, which deals specifically with the question of large and continually mounting public debt, constitutes in all essential respects a 'natural history of the rise and collapse of public credit'. Particularly noteworthy in this analysis are the extensive relationships Hume draws between economic and other social developments, especially of a political and sociological character. Of all aspects of his political economy, this essay most fully exhibits Hume's awareness, as a moral scientist, of significant interrelations between different realms of social experience.

The third and last level of Hume's economic thought in his economic philosophy, which is his appraisal, on ultimate moral grounds, of the desirability of a commercial and industrial society. In light of his general concern, as a philosopher, with moral questions, it is hardly surprising to find that the question of the moral aspects of commercial and industrial growth was of basic importance for Hume. Appearing in the second of the economic essays – 'Of Refinement in the Arts' – he considers this question before turning to an analysis of market problems. Although the essay is brief, its scope is broad; for Hume discusses the impact of the development of an advanced economy both on the individual and on society as a whole.

The standard for moral judgement Hume employs is drawn from the utilitarian ethic – a position which he himself had expounded and defended in his philosophical analysis. And here the role played by his natural history of the rise and progress of commerce is fundamental. As observed, in this natural history Hume dealt with various 'causes of labour'. In his economic philosophy three of these motives – the desires for consumption, for interesting activity and for liveliness – are now treated as ends which are regarded as major ingredients of the happiness of the individual. Here he argues that, by providing new consumption experiences, enlarging the scope for the enjoyment of economic activity as a form of interesting action and (through both the latter) enhancing a sense of liveliness, economic growth advances the fulfilment of all these ends. Economic growth, he contends, contributes to the fulfilment of a fourth end of importance to human welfare – a sense of peace and tranquillity or a state of no passion – which he argues is enjoyable only in 'recruiting the spirits' after intensive indulgence in lively experiences. It is noteworthy that Hume's treatment of these ingredients of human happiness bears a direct relationship to the principal conceptions of the good life as Hume construes these in an earlier series of essays entitled 'The Epicurean', 'The Stoic' and 'The Platonist'. Further, the pluralism reflected in his multidimensional prescription for human happiness springs from the position taken in a fourth essay on the good life entitled 'The Sceptic' (1955, pp. xcv–xcix).

Turning to a treatment of the effect of economic development on major aspects of social relations, Hume now expands the 'natural history' to encompass non-economic considerations. He argues that economic growth contributes to the growth of knowledge in the liberal as well as the mechanical arts, nurtures a sense of humanity and fellow-feeling, enhances a nation's spiritual as well as its economic ability to defend itself and, through its impact on the growth of knowledge and fellow-feeling, advances an understanding of the art of government and political harmony. A final political consideration, to which Hume gives special attention, is the charge (drawn from the experience of Rome) that luxury is corrupting and debasing and therefore is inimical to liberty. Hume argues that history shows that precisely the opposite is true. For the growth of commerce brings the expansion of the merchant class – the 'middling rank of men' who above all are interested in uniform laws protecting their property; and it is this development, he emphasizes, which has led to the growth of parliamentary government and the associated respect for individual liberty. Hume thus perceived the link between the growth of economic individualism and political liberty that has drawn so much attention since his time. Although Hume recognized that the development of commerce and industry could produce evils of its own, he argued that these were outweighed by its benefits. Owing apparently to an overzealous desire to counter the common religious objections to luxury, Hume overextends himself and leaves some of his arguments in support of economic growth open to criticism (1955, pp. cii–civ). His treatment nonetheless stands as an unusually broad and penetrating appraisal of a wealth-orientated individualistic society. In light of this it deserves recognition as an early classic.

Throughout our discussion, attention has been given to Hume's interest in the psychological and historical aspects of economic activity. A similar interest – pursued in varying degree – is found among other writings of Hume's own period. However, owing to his own searching analysis as a philosopher and historian, Hume's treatment was of a particularly high order; equally extraordinary was the extent to which he employed the method of 'natural history' in the treatment of a wide range of issues of economic theory and policy.

Comparing Hume with Adam Smith (his close friend), one is struck by the brevity of Hume's economic writings. Hume wrote a series of relatively short 'discourses' on selected topics. Smith's *Wealth of Nations* (1776) is a general economic treatise. In contrast to Smith, Hume moreover gives little systematic attention to price and distribution theory, which was to become the major concern of classical and neoclassical economics. In point of the general analysis of psychological and historical influences on economic activity, however, Hume's work is more comprehensive, more highly organized and more penetrating than Smith's. When dealing with the subjective aspects of human behaviour, Smith not infrequently regards them as universals (for example, his assertion that there is an innate disposition among men to 'truck and barter'), where Hume treats them as historical variables and himself seeks to explain the nature of the specific historical influences at work (1955, pp. cvii–cx). In this Hume did not foreshadow the mainstream of subsequent economic thought; it was Adam Smith's tendency in his economic theory to abstract from history that was to become the dominant characteristic of later economic analysis. In point of general perspective (though often not its conceptual framework) Hume's economic thought bears a relation to other subsequent lines of development – to the historical and institutional schools of economics, to the more current revived analytical interest in economic growth along with its associated cultural aspects, to the concern with psychological factors in dealing both with macroeconomics and the economics of non-competitive markets, and to the normative appraisals of economic systems in their fuller social settings.

In the standard histories of economic thought Hume has been accorded relatively little attention. He is often ignored altogether or treated cursorily as a predecessor of Adam Smith. Various studies of the technical aspects of economic analysis have called attention to several of Hume's contributions. These aspects of Hume's analysis are noteworthy in their own right. Their significance deepens and broadens when they are related to Hume's work as a philosopher and historian and are seen to take form within the context of 'natural history'.

See Also

- ► Bullionist Controversies (Empirical Evidence)
- ► Gold Standard
- ► Specie-Flow Mechanism

Selected Works

1752. *Political discourses*. Edinburgh: A. Kincaid & A. Donaldson.

1875. *The philosophical works of David Hume,* ed. with notes by T.H. Green and T.H. Grose, 4 vols. London: Longmans, Green & Co.

1955. *Writings on economics,* ed. E. Rotwein. London: Nelson.

Hunters, Gatherers, Cities and Evolution

Paul Seabright

Abstract

Human beings evolved in hunter-gatherer bands, and tended to flee from or to fight with strangers. They have subsequently learned to live in cities among a multitude of such strangers, at levels of violence far lower than those that characterized prehistory. The key to this development was the adoption of agriculture, which obliged humans to become sedentary to and to develop institutions to manage their encounters with strangers. We describe the evolution of the psychological preconditions for the agricultural revolution, and its consequences for social life.

Keywords

Agricultural revolution; City; Cooperation and its evolution; Division of labour; Equality; Evolution; Group selection; Hierarchy; Hunter-gatherers; Learning; Other- regarding preferences; Population growth; Reciprocity; Repeated games; Signalling; Slavery; Trust; Urbanization

JEL Classifications

R0

Modern human beings (*Homo sapiens*) evolved in Africa but now occupy all the continents of the globe. The environmental conditions in which they live are mostly quite different from those in the woodland savanna in which they evolved, since they occupy habitats that vary enormously in terms of temperature, humidity, terrain and vegetation, available foodstuffs and building materials, and dominant predators.

More surprisingly, the social conditions in which they live are dramatically different too. The latter change has happened much more recently and much more suddenly. For almost all of their existence, including during the time that they were fanning out from Africa to other continents, human beings have lived in bands numbering from a few dozen to a maximum of a few hundred individuals, and have survived through hunting and gathering. These individuals would have known each other fairly well, and many of them (in particular the men) would have had reasonably close genetic ties. At the beginning of the 21st century, however, around half of the world's population lives in urban areas, a proportion likely to rise to 60 per cent by 2030, and around 40 per cent of these live in agglomerations of more than one million inhabitants (United Nations 1999).

Many interesting questions are raised by this development, of which the most puzzling concern how human beings have managed to sustain a complex web of cooperation, such as that which underpins the sophisticated modern division of labour, between individuals who do not have ties of kinship. The goods that are consumed by the modern urban household are manufactured in many different stages by different people who have no relation to one another and may not even live in the same country. The theory of kin selection (Hamilton 1964) explains the evolution of cooperation among genetically related individuals, which is widespread in the animal kingdom (most famously among the social insects). But it predicts fierce rivalry among unrelated individuals, especially among males under conditions of strong sexual selection. With a few unimportant exceptions, significant cooperation among unrelated individuals has never evolved in any species other than man. Particularly puzzling is cooperation among strangers, which is the foundation of modern urban life. There is evidence that, while hunter-gatherer bands were mostly close knit and highly cooperative, encounters between strangers for much of human evolution have been accompanied by serious, often lethal violence. Human psychology has therefore been powerfully shaped by the fact that for much of our past we were one another's most dangerous predators (Sterelny 2003). But most human beings now encounter strangers in their thousands every day without giving the matter a second thought, and even in the world's more dysfunctional cities they run a risk of violence that is far lower than it was during the whole of human prehistory. Deaths by violence average a little over one per cent of all deaths in the world as a whole, whereas in prehistoric times they are estimated to have averaged between 10 and 40 per cent of all fatalities (WHO 2002; Keeley 1996). How has this remarkable transformation of human social existence come about?

A look at the dating of these changes makes clear that the answer does not lie in a change in the genetic basis of human social psychology, but rather in the flexible adaptation of an existing psychology to a new social environment. DNA and archaeological evidence both suggest that the basic genetic architecture of the mind of modern man was in place many tens and possibly some hundreds of millennia ago. Yet hunting and gathering in relatively small bands remained universal until around 10,000 years ago, and the first large-scale agricultural civilizations did not emerge until around 5000 years later. Modern human beings are navigating in their social lives with instruments that evolved to guide them in a quite different world.

A further puzzle about human social life is that the primate species from which we evolved – including the great apes, our cousins – live in bands governed by strong status hierarchies. Modern human social life is no less governed by strong inequalities of rank, status and access to economic resources as well as to intangible goods such as esteem. However, hunter-gatherer communities in the late Paleolithic, at a stage intermediate between hunting and gathering, appear to have been fairly egalitarian in the distribution of both resources and esteem, at least between individuals of the same sex (it is likely that relations between the sexes were more unfavourable to females than among our closest primate relatives, if only because human females were more dependent on males for both food and protection than is the case among chimpanzees and bonobos). How did human beings achieve such a degree of equality, and why did they lose it again?

The answer to these puzzles turns crucially on our understanding of the first agricultural revolution, which spread at a remarkable pace around the world, and which obliged human beings to become largely sedentary, encouraging them in the process to move into villages and towns for protection. It also enabled the production and storage of a surplus over subsistence that could be devoted to other economic ends, spurring the division of labour and the growth of complex and hierarchical civilizations. Three main questions stand out: first, which of our mental capacities that had evolved before the agricultural revolution was to prove most important in shaping how human beings responded to that dramatic development? Second, what caused the agricultural revolution, and why did it spread so fast? And third, what were its consequences for human social life?

The mental capacities that mark *homo sapiens* out from our ancestors and cousins in the hominid line have been the subject of much debate, many aspects of which remain unresolved. It seems likely, though, that they included most or all of the following elements: a capacity for symbolic thought, the ability to contemplate and refer to absent or invisible objects and events, an enhanced concern for the future and one's own place in that future, a 'theory of mind' that enabled greatly enhanced prediction of the behaviour of other people, a sophisticated ability to detect cheating on the part of those others, and a greatly enhanced capacity to imitate their behaviour in a flexible and creative way (Cosmides and Tooby 1992; Deacon 1997; Tomasello 1999). Mithen (1996) has even argued that only our own species has the capacity for consciousness in a proper sense, and has offered an intriguing theory of its evolution. For our purposes the crucial point is that these capacities would have been the very ones that, as described in the literature on cooperation in repeated games, enable human beings to cooperate even in the presence of conflicting interests.

The capacities to represent and care for the future, to predict how the behaviour of others may respond to our own, to respond appropriately to their trustworthiness or dishonesty, and to learn from the successes and mistakes of others – all these would no doubt have been highly useful for undertaking cooperatively the increasingly complex challenges of hunter-gatherer existence. Once in place they could then contribute to Renaissance statecraft, higher mathematics and running a 21st-century corporation, among other things. In addition, the ability to represent absent or invisible objects and events would have greatly enhanced the strategy space for would-be cooperators. If I know that by stealing a rival's food I risk not only his own retaliation (which might be restricted by his limited information or physical strength, or by the fact that since he is a stranger he is likely never to see me again), but also that of Mr Plod, Inspector Maigret and Judge Jeffreys, I shall have much more to lose. Other primates have sophisticated strategies of peacemaking (De Waal 1989) but only limited means with which to enforce them. Our own species' mental capacities enabled the invention of much more ingenious institutions of enforcement than had ever been available to hunter gatherers.

There is more controversy, however, over whether large-scale cooperation required evolutionary developments in the affective as well as the cognitive components of human psychology. It is argued by many that cooperation requires human beings to display 'other-regarding preferences', which depart from those that would maximize an individual's enlightened self-interest (interpreted as inclusive fitness according to the Hamiltonian model of kin selection). Specifically, such other- regarding preferences must include 'strong reciprocity' – a preference for repaying cooperation with cooperation, and cheating with revenge, even when this is not what the calculus of self-interest would require (see Henrich et al. 2004; Seabright 2004). These authors claim that purely self-interested behaviour, even of the sophisticated kind described in the repeated game literature, would not have permitted complex cooperation because of the problems of limited observability and consequent mistakes in the implementation of retaliation strategies. Conversely, a small amount of strong reciprocity, even among a subset of the relevant population, can go a long way in reinforcing cooperative behaviour (but see Binmore 2005; Gintis et al. 2006). How such preferences could have evolved remains an open question, with some favouring a version of group selection (Gintis 2000), and others preferring a form of signalling, in which the presence of other-regarding preferences made individuals more attractive as partners in cooperative activities (Frank 1988).

However this controversy is resolved, it seems indisputable that communities of human hunter-gatherers were governed by strong cooperative norms, held in place by some combination of kin altruism, mutual monitoring under repeated interaction, and other-regarding preferences. Boehm (1999) has argued that these communities were more egalitarian (among males) than any before or since. This was not because humans had lost the strong sense of rivalry, including status rivalry, displayed by other primate species, but because the strong competitiveness of individual motivation was held in check by social mechanisms that retaliated against overweening displays of power or arrogance by any successful individual. Under the circumstances of hunting and gathering, great disparities of wealth or status were neither possible (since mobility precluded storage) nor desirable (since hunting required too much flexibility to be undertaken by the unwilling or the enslaved).

But strong community solidarity coexisted with violent inter-community rivalries. Although it used to be believed that hunter-gather communities were inherently peaceful, this is now known to be a myth (Ember 1978; LeBlanc 2003). Though trading links existed between different communities (including for the exchange of marriageable women), encounters between strangers or historic rivals were frequently violent, much as they are known to be often violent when groups of foraging chimpanzees of unequal strength encounter each other by chance in the wild (Ghiglieri 1999).

How did all this change? Beginning around ten thousand years ago, agriculture was independently invented in at least seven different places (Anatolia, Mexico, the Andes of South America, northern China, southern China, the eastern United States, and in sub-Saharan Africa at least once and possibly up to four times; see Richerson et al. 2001). The techniques of agriculture spread rapidly around the world (Bellwood 2005), not simply by emulation but by the migration of the farmers themselves (Cavalli-Sforza et al. 1994). It was all the more surprising that agriculture should catch on so fast because studies of the bones and teeth of some of the earliest agricultural communities of the Near East show that farmers had worse health, due to poorer nutrition, than the hunter-gatherers who preceded them (Cohen and Armelagos 1984). Increases in agricultural productivity in later millennia more than made up for this eventually, but even so the puzzle remains: what prompted agriculture to be adopted so quickly and often within a comparatively short space of time, if it did not achieve the one thing that a new agricultural technique surely ought to achieve – to leave people better fed than they were before?

Explanations for the paradox have included the depletion of game, which lowered the productivity of hunting (see hunting and gathering economies), and the fact that agriculture once adopted led to population growth and crowding, thereby reducing food availability and increasing disease (Bar-Yosef and Belfer-Cohen 1989; Robson 2005). Consistent with these views, and adding force to the view that agriculture might have been irresistible even if disadvantageous to the health and nutrition of its adopters, is the idea that sedentarism significantly increased the effort and the resources human societies had to devote to defence (Seabright 2006).

Those who are sedentary are also vulnerable. When enemies attack, farmers have much more to lose than hunter-gatherers, who can melt into the forest without losing houses, chattels and stores of food. So farmers need to spend time, energy and resources defending themselves – building walls, manning watchtowers, guarding herds, patrolling fields. This means less time and energy, fewer resources, devoted to making food. The greater productivity of the hours they spend growing and raising food could even be outweighed by the greater time they must spend defending themselves and the food they have grown – meaning that they produce less food in all.

But why should the first farmers have adopted agriculture at all? And why should this new technology have spread with such rapidity? Stunted farmers would hardly have been a good advertisement to their hunter-gatherer neighbours of the qualities of their new wonder diet. What is needed is an account that explains how agricultural adoption could have been individually rational even if perhaps collectively self-defeating, at least in the short run.

Agriculture dramatically raised the advantages to mankind of banding together for self-defence. Once constrained by a sedentary lifestyle and unable any longer to play hide-and-seek with its enemies, a large group is much more secure than its members could be in multiple smaller groups. But once the first farmers began to invest systematically in defence, they became a threat to their neighbours, including communities who were on the margins of adopting agriculture themselves. There is no such thing as a purely defensive technology. Even walls around a town can make it easier for attacking parties to travel out to raid nearby communities in the knowledge they have a secure retreat. The club that prehistoric man used to ward off attackers was the same club he used to attack others. Once a community has invested in even a modest army, whether of mercenaries or of its own citizens, the temptation to encourage that army to earn its keep by preying on weaker neighbours can become overwhelming. So, even if the first farming communities were not necessarily any better off than they would have been if no one had adopted agriculture, once the process had started many communities had an interest in joining in. These interactions could lead each to act ineluctably against the collective interests of all. It is a logic well known from the theory of contests (Becker 1983; Hirshleifer 1989).

However, the necessity of self-defence was also in time the mother of an astonishing array of technological and institutional mechanisms for keeping the peace and encouraging social cooperation, albeit much more effectively within communities than between them. Many of these mechanisms were subject to significant economies of scale, which encouraged the growth of cities even before their more subtle effects on economic development had been remarked (by Adam Smith among others). They led also to the accumulation of wealth, status and power by a minority of individuals within society who had access to land or capital, or to control of the means of inflicting physical violence. Added to the fact that agriculture could be carried out by slaves under constant surveillance, as hunting and gathering could not, this led to a dramatic increase in inequality. Almost no societies did not enslave others at some time in their history, with slavery becoming more likely the wealthier the society concerned, at least until it became wealthy enough to afford to take a stand against slavery on principle (Nieboer 1900; Fogel and Engerman 1974).

The institutions that now keep the peace in an urban environment are extraordinarily subtle, as the work of Jane Jacobs (1961) has notably emphasized. The police and courts are but the apex of an informal structure of eyes and ears that depends on the willing participation of citizens in a neighbourhood. Formal authority alone can never establish order, as Raymond Chandler recognized when in 1950 he wrote of a ‘world in which gangsters can rule nations and can almost rule cities’. The historian Peter Hall (1998) has also noted that the characteristics that turn some cities into crucibles of artistic creativity and economic innovation depend on subtle networks of interaction that are impossible to plan in detail. They are an organic outgrowth of human beings’ acquired capacity to build trust with strangers in a daily multitude of individually insignificant but collectively remarkable encounters.

In these and other ways the consequences of the developments in human psychology that permitted the adoption of agriculture were momentous for human life. A long-standing literature in political theory, going back to Ibn Khaldun (1377) and excellently discussed by Ernest Gellner (1994), considers the need to raise a surplus for defence as constituting the foundation of the division of labour and as giving rise to some of the most intractable problems of political organization. It can be said, therefore, to be at the root of both the most remarkable intellectual and economic achievements of human society and of its most deplorable cruelties and excesses.

See Also

▶ Hunting and Gathering Economies

Bibliography

Bar-Yosef, O., and A. Belfer-Cohen. 1989. The origins of sedentism and farming communities in the Levant. *Journal of World Prehistory* 3: 447–497.

Becker, G. 1983. A theory of competition among pressure groups for political influence. *Quarterly Journal of Economics* 98: 371–400.

Bellwood, P. 2005. *First farmers*. Oxford: Blackwell.

Binmore, K. 2005. *Natural justice*. Oxford: Oxford University Press.

Boehm, C. 1999. *Hierarchy in the forest: The evolution of egalitarian behavior*. Cambridge, MA: Harvard University Press.

Cavalli-Sforza, L., P. Menozzi, and A. Piazza. 1994. *The history and geography of human genes*. Princeton: Princeton University Press.

Cohen, M.N., and G.J. Armelagos. 1984. *Paleopathology at the origins of agriculture*. New York: Academic.

Cosmides, L., and J. Tooby. 1992. Cognitive adaptations for social exchange. In *The adapted mind*, ed. J. Barkow, L. Cosmides, and J. Tooby. New York: Oxford University Press.

Deacon, T. 1997. *The symbolic species: The co-evolution of language and the human brain*. London: Allen Lane.

De Waal, F. 1989. *Peacemaking among primates*. Cambridge, MA: Harvard University Press.

Ember, C. 1978. Myths about hunter-gatherers. *Ethnology* 17: 439–448.

Fogel, R., and S. Engerman. 1974. *Time on the cross: The economics of American negro slavery*. New York: Little Brown.

Frank, R. 1988. *Passions within Reason: The strategic role of the emotions*. New York: W.W. Norton.

Gellner, E. 1994. *Conditions of liberty: Civil society and its rivals*. London: Hamish Hamilton.

Ghiglieri, M. 1999. *The dark side of man: Tracing the origins of male violence*. Cambridge, MA: Perseus Publishing.

Gintis, H. 2000. Strong reciprocity and human sociality. *Journal of Theoretical Biology* 213: 103–119.

Gintis, H., et al. 2006. Symposium on Ken Binmore’s natural justice. *Politics, Philosophy and Economics* 5(1).

Hall, P. 1998. *Cities in civilisation*. London: Weidenfeld & Nicolson.

Hamilton, W. 1964. The genetical evolution of social behaviour. *Journal of Theoretical Biology* 7: 1–52.

Henrich, J., et al. 2004. *Foundations of human sociality*. Oxford: Oxford University Press.

Hirshleifer, J. 1989. Conflict and rent-seeking success functions: Ratio vs. difference models of relative success. *Public Choice* 63: 101–112.

Ibn Khaldun, A.Z. 1377. *The Muqadimmah,* trans. F. Rosenthal. Princeton, NJ: Princeton University Press. 1969

Jacobs, J. 1961. *The death and life of great American cities*. New York: Vintage Books. 1992.

Keeley, L. 1996. *War before civilization: The myth of the peaceful savage*. Oxford: Oxford University Press.

LeBlanc, S. 2003. *Constant battles*. New York: St Martin's Press.

Mithen, S. 1996. *The prehistory of the mind*. London: Thames & Hudson.

Nieboer, H.J. 1900. *Slavery as an industrial system*. New York: Burt Franklin. 1971.

Richerson, P., R. Boyd, and R. Bettinger. 2001. Was agriculture impossible during the pleistocene but mandatory during the holocene? A climate change hypothesis. *American Antiquity* 66: 387–411.

Robson, A. 2005. A bioeconomic view of the neolithic and recent demographic transitions. Mimeo, Simon Fraser University

Seabright, P. 2004. *The company of strangers: A natural history of economic life*. Princeton: Princeton University Press.

Seabright, P. 2006. Warfare and the multiple adoption of agriculture after the last ice age. Discussion paper, Centre for Economic Policy Research

Sterelny, K. 2003. *Thought in a Hostile world: The evolution of human cognition*. Oxford: Blackwell.

Tomasello, M. 1999. *The cultural origins of human cognition*. Cambridge, MA: Harvard University Press.

United Nations. 1999. *World urbanization prospects: The 1999 revision*. New York: Population Division, United Nations.

WHO (World Health Organization). 2002. *The World Health Report 2002*. Statistical Annex Table 2. Online. Available at http://www.who.int/whr/2002/en/annex_table2.xls. Accessed 28 Nov 2006.

Hunting and Gathering Economies

Vernon L. Smith

Keywords

Agricultural revolution; Commodity money; Human capital; Hunting and gathering economies; Leisure; Multilateral contracting

JEL Classifications

R0

Men and women (*Homo erectus*) who were culturally and biologically distinguishable from other hominoids have lived on the planet Earth for about 1.6 million years (Pilbeam 1984). It is likely that the biological changes since that time form a microevolutionary continuum: archaic *H. sapiens*, including the Neanderthal, appeared 125,000 years ago and anatomically modern *H. sapiens* appeared about 45,000 years ago. The record suggests that *H. erectus* fabricated and used tools, and his use of fire may have begun by 700,000 years ago. The changes identified in the prehistoric period appear only to distinguish less advanced from more advanced stone age technology. Consequently, the dominating message seems to be that over almost the whole of man's epoch on earth he lived successfully as an exceptionally well-adapted hunter. It is only recently, in the last 8000–10,000 years (less than 1 per cent of his time on Earth), that man abandoned the nomadic life of the hunter to begin growing crops, husbanding domesticated animals, and living in villages. It is difficult to exaggerate the importance of this agricultural or first economic revolution (North and Thomas 1977) in understanding who we are, and what we have become. Once man opted for the farmer–herder way of life it was but a short step to mankind's much more sophisticated development of specialization and exchange, greatly enlarged production surpluses, the emergence of the state, and finally the Industrial Revolution. Our direct knowledge of early man is confined to the record of the durables he left behind. Yet when combined with anthropological evidence from the study of recent hunter–gatherer economies the evidence can be interpreted as demonstrating that all the ingredients associated with the modern wealth of nations – investment in human capital, specialization and exchange, the development of property right or contracting institutions, even environmental 'damage' – had their development in the course of that vast prehistorical, pre-agricultural, period.

What accounts for this sudden abandonment of the nomadic hunting life? We do not know for we have no direct observations on the transformation from hunting to agriculture. This transformation is perhaps the pre-eminent scientific mystery, since all of that which we have called civilization, all the great achievements of industry, science, art and literature stem from that momentous event within the last few minutes of man's day on Earth. Yet there are common factors that dominated the evolution of man from his earliest form to modern *H. sapiens*, and his primary intellectual and social development, which suggest an underground continuity between the pre-agricultural, Paleolithic hunting period, and the agricultural and subsequent periods.

Man the Hunter–Gatherer

There are many widely held beliefs concerning the characteristic features of the hunter–gatherer way of life that stretch back several hundred years in academic writings, and persist as part of the folklore of contemporary man's misperception of his own prehistoric past; until recently these beliefs dominated even the anthropological view of hunter–gatherer 'subsistence'. These beliefs tend to obscure the striking continuity in man's ability to respond to changes in his environment by substituting new inputs (labour, capital and knowledge) for old, and develop new products to replace the old when effort prices were altered by the environment.

Ever since Hobbes there has prevailed the perception that life in the state of nature was 'solitary, poor, nasty, brutish and short'. A more accurate representation (if not strictly correct in all aboriginal societies) would argue that the hunter culture was the original affluent society (Lee and DeVore 1968). Extensive earlier data on extant hunter–gatherers show that with rare exceptions (such as the Netsilik Eskimos) their food base was at minimum reliable, at best very abundant. The African Kung Bushman inhabited the semi-arid north-west region of the Kalahari Desert, an inhospitable environment, characterized by drought every second or third year. These conditions had served more to isolate the Kung from their agricultural neighbours than to condemn them to a brutish existence. Adults typically worked 12–19 hours per week in getting food. As with all such societies, for the most part the women gathered, the men hunted. The caloric-protein returns exceeded several measures of nutritional adequacy. Gathering was the more reliable and productive activity with women producing over twice as many edible calories per hour as men. Both men and women bought leisure with this work schedule – resting, visiting, entertaining and (for the men) trance dancing. About 40 per cent of the population were children, unmarried young adults (15–25 years of age) or elderly (over 60 years of age), who did not contribute to the food supply and were not pressured to contribute.

A comparable macroeconomic picture applied to the Hazda in Tanzania. Large and small animals were numerous and all – with the exception of the elephant – were hunted and eaten by Hazda. Hunting was the speciality of men and boys, conducted as an individual pursuit that relied primarily on poisoned arrows. The Hazda spent on average no more than two hours a day hunting. The principal leisure activity of the men was gambling, which consumed more time than hunting.

Other hunting (or fishing) peoples of Africa, Australia, the Pacific Northwest, Alaska, Malaya and Canada have shown comparably effective adaptation to this form of livelihood. Malnutrition, starvation and chronic diseases were rare or infrequent, although accidental death was high in certain cases such as the Eskimo.

The argument that life in the Paleolithic must have been intolerably harsh is simply not borne out by the many ethnographic studies of extant hunting societies in the past century. With few exceptions such societies have fared well, and did not leap to embrace the agricultural or pastoral pursuits of their neighbours. Whether life in the Paleolithic mirrored this modern experience cannot be known with any assurance, but certainly there is no support for the proposition that hunting, per se, means an intolerably harsh existence. In fact, the Paleolithic hunting economy had demonstrably high survival value in a world far more plentifully endowed with game than has existed since the great megafaunal extinctions of the late Pleistocene, and therefore a world which might indeed have been marked by numerous original affluent economies.

Although it is natural to suppose that man's uniqueness derived from his intellectual superiority, what is more likely is that man's physical superiority was also important in giving him a superpredator's advantage over other species. His endowment of physical human capital would probably have been of significance even in the absence of his investment in tools and the human capital required to produce and use tools. As noted by J.B.S. Haldane, only man can swim a mile, walk 20, and then climb a tree. Add to this observation the four-minute mile, unsurpassed long-distance endurance running, the ability to carry loads in excess of body weight, high altitude performance, American Indian capacity literally to run down a horse or deer by pacing the animal, the incredible accomplishments of acrobats and gymnasts, and finally the finger agility and coordination required to milk a cow, and you are left with the physical portrait of an astonishingly superior species. It appears that man's basic foundation of physical superiority was laid by his upright stance, to which of course the addition of knowledge made him truly formidable, even in the presence of the various giant proboscidea (mastodon, mammoth, elephant) which early man did not hesitate to hunt and to kill on three continents.

The idea that primitive man was too puny and too few in number to have had a significant influence on his environment underestimates man's uniqueness as a tool using, fire using, highly mobile species who, with minor exceptions (Madagascar, New Zealand and Antarctica), had populated the world by 8000 BC. The archaeological record suggests that man was a big game hunter par excellence. He hunted mammoth, mastodon, horse, bison, camel, sloth, reindeer, shrub oxen, red deer, aurochs (wild cattle), and other large mammals, for perhaps a minimum of 30,000–40,000 years, ceasing only with the great megafaunal extinctions throughout much of the world some 8000–12,000 years ago. Paul Martin (1967) has argued the case for the overkill hypothesis that man was a significant causative factor in these extinctions. Essentially, the argument is that the alternatives to overkill, principally the climate hypothesis, fail to account for the worldwide pattern of these extinctions which appear to have begun in Africa and perhaps southeast Asia 40,000–50,000 years ago, spread north through Eurasia 11,000–13,000 years ago, jumped to Australia perhaps 13,000 years ago, and entered North America in the last 11,000 years, followed by South America 10,000 years before the present. The most recent extinctions are in New Zealand (numerous species of flightless moa birds) 900 years ago and in Madagascar 800 years ago, shortly after the remarkably late migration of man to those islands.

Man's use of fire as a tool in the management and control of natural resources must be counted as having a profound effect on his ecological environment. Numerous authors who have studied patterns of land burning by primitive peoples have concluded that most of the greatest grasslands of the world represent fire- vegetation that is man-made (see Heizer 1955, for a summary). Where tree growth is strongly favoured by climatic conditions, regular burning will select for certain species of tree such as the pine stands of southern New York and to the West, which have been attributed to Indian burning. Contemporary man's attempts to prevent fires, which today are almost entirely caused by lightning, has probably produced far more ecological damage than the controlled use of fire that has characterized aboriginal cultures. Recurrent fire prevents the accumulation of brush which then fuels the holocaust wildfire that destroys all forest vegetation.

A third source of ecological change produced by primitive peoples was their transportation of seed, in their migrations as hunter–gatherers, which introduced numerous botanical exotics into new regions. Archaeologists have frequently observed the association of various plants with ancient campsites and dwellings. For example, the wide distribution of wild squash, gathered for its seed, appears to be associated with man. The introduction of exotics can and has produced significant environmental changes in modern times, but the phenomenon has ancient origins and may have been considerably more disruptive as the first men moved from one 'pristine natural' region to another.

Success as a hunter–gatherer requires human capital usually associated only with agricultural and industrial man: learning, knowledge transfer, tool development and social organization. Comprehensive studies of the aboriginal use of fire for game and plant management show clearly that primitive men demonstrated extensive knowledge of the reproductive cycles of shrubs and herbaceous plants, and used fire to encourage the growth and flowering of the plants used in gathering, and to discourage the growth of undesirable plants (Lewis 1973). This required one to know when, where, how and with what frequency to apply the important tool of controlled burning for managing the resources that allow gathering to make an efficient, productive and sustainable contribution to living. Primitive men knew that the growing season can be advanced by spring burns designed to warm the earth, that in dry weather fires should be set at the top of hills to prevent wild fires, but in damp air they should be set in depressions to avoid being extinguished, that the burning of underbrush aided the growth of the oak whose acorns were eaten and attracted moose who avoid underbrush, and that deer and other animals congregate to feed on the proliferation of tender new plants that sprout following a fall burn.

To live by hunting is to be committed to an intellectually and physically demanding activity that requires technology, skill, social organization, some division of labour, knowledge of animal behaviour, the habit of close observation, inventiveness, problem solving, risk bearing, and high motivation, since the rewards are great and the penalties severe. Such exceptional demands could have been highly selective in man's long evolution, and disciplined the development of the intellectual and genetic equipment that facilitated his subsequent rapid creation of modern civilization. This natural selection could have been intensified by the widespread practice among aboriginals of rewarding superior hunters with many wives.

It was as a hunter that man learned to learn. In particular he understood that young boys must be imbued with the habit of goal-oriented observations, and with knowledge of animal behaviour and anatomy. To know that many ungulates travel in an arc meant that tracking success could be improved by transversing the chord. Knowledge of animal behaviour was a substitute for weapon development. Even the weapons of the later pre-agricultural period (spears, bow and arrow, harpoon) required the hunter to approach the prey within ten yards for a best shot. This might require hours crouched on the ground waiting for a shift in the wind, for just the right change in the animal's position, or for the mammoth to get deeper into the bog in a watering hole. The weapons changed with shifts to new prey. Thus the Clovis fluted point, widely distributed throughout North America, was used to kill mammoth and mastodon 11,000–12,000 years ago. The Folsom point was then developed and used to kill the large, now extinct *Bison antiquus*, which then gave way to the Scottsbluff point associated with the killing of the slightly smaller, now extinct *Bison occidentalis* (Haynes 1964; Wheat 1967). These observations suggest high specialization which required new forms of human and physical capital to meet the specialized demands of new prey.

The organizational requirements of the hunt are illustrated at the Olsen–Chubbuck site in Colorado, where the excavated remains of bones and projectile points of the Scottsbluff design show that about 8500 years ago some 200 *Bison occidentalis* were stampeded into an arroyo 5–7 feet deep. Armed hunters in the arroyo on each side of the stampede then slaughtered the injured or escaping animals with their weapons (Wheat 1967).

Primitive man has often been modelled as 'cultural' not 'economic' man, but the power and importance of the opportunity cost principle in conditioning the choice of all peoples was perceptively stated by the Kung Bushman, who, when asked why he had not turned to agriculture, replied, 'Why should we plant, when there are so many mongongo nuts in the world?' (Lee and DeVore 1968, p. 33). This Bushman, I would hypothesize, stated the answer to the scientific question: why did man the hunter tend to abandon that which appeared to serve him so well for 1.6 million years and to which he seems to have adapted ever more successfully, as indicated by the growing complexity of his tools and weapons as he evolved from *H. erectus* to anatomically modern *H. sapiens*? Man would not have given up the hunter–gatherer life had there not been a change in the terms of trade between man and nature that made the hunting way of life more costly relative to agriculture. This *hypothesis* does not leave 'culture' out of the equation. Thus to describe hunter–gatherers as directly seeking the cultural goal of prestige does not contradict the hypothesis that man, like nature, ever economizes. Attaching prestige to the hunt may simply be an astute means of advertising, teaching and propagating the discovery that hunting and its attendant technology is the best means of livelihood, with the result that each new generation does not have to rediscover this knowledge. Myths of the great hunter, of great rewards, of great penalties for lost technique, of killing the goose that lays golden eggs are part of the oral tradition by which the economy preserves this human capital.

The hypothesis that the agricultural revolution was due to a major decrease in the productivity of labour in hunting–gathering relative to agriculture (Smith 1975; North and Thomas 1977) is consistent with the observations that this cultural shift (*a*) occurred at different times in different parts of the world, with small aboriginal hunting enclaves still in existence, and (*b*) did not occur once and for all in every such tribe. With respect to (*a*), the great wave of terrestrial animal extinctions occurred over a period of several thousand years, and therefore the relative increase in the cost of hunting struck different regions at different times. Also different peoples in different environments with different opportunity costs would be expected to provide different mechanisms of adaptation, with some persisting as gatherers and small game hunters, and others turning to or perhaps persisting as fisherman (for example, the Aleutian Eskimos and the Pacific Northwest Indians) in regions unsuitable for agriculture. With respect to (*b*) the reintroduction of the horse in North America by the Spanish (in the hardy form of *Equus caballus* just 8000 years after other members of the genus became extinct in the Americas) had a major modifying impact on the economy of the plains Indians. In the northern plains the 'fighting' Cheyenne, as they were later to be termed by the Europeans, and the Arapahoe quickly abandoned their villages along with their pottery arts and horticulture to become nomadic Bison hunters (see the references in Smith 1975). Apparently, agricultural productivity was dominated by the enormous increase in the bison harvest made possible by a technological change that combined the horse with the bow and arrow. To the south, where the growing season was longer and the climate more favourable, the Pawnee preserved their maize agriculture when they turned to Bison hunting, creating a mixed agricultural–hunting economy. The south-western Apache, reported by Coronado in 1541 to be subsisting as bison hunters, simply adapted the horse to their pre-existing hunter culture. The vast bison-hide tepee encampments witnessed by the first Europeans to cross the plains were already the product of a technologically transformed native American, many of whom had only recently abandoned their agricultural economies.

Pleistocene Extinctions and the Rise of Agriculture

Here then is a model of the epoch of man: he arrives 1.6 million years ago as a hunter among hunters, but distinguishable in terms of his human capital endowment and his ability to invest in the development of human and physical capital. His tools become more complex and knowledge of the use of fire, perhaps his most significant tool, is added to his stock of human capital. There is a gradual improvement in weapons technology – clubs, stones, stone axes, spears, stone projectile points, the atlatl (which applies the leverage principle) and, in the late pre-agricultural period, the bow (which combines the leverage principle with temporary storage of energy for increased mechanical advantage). The combination of his physical superiority, tools and fire make him a superpredator without equal. At some unknown point this success brings relative affluence, and the important commodity 'leisure', which might have contributed to the development of language and other forms of investment in human and physical capital.

Although *H. erectus* and archaic *H. sapiens* were advanced hunters who apparently spread from Africa to Eurasia and Asia, it remained for modern *H. sapiens* to establish himself as a big game hunter par excellence, who populated most of the world by 8000 BC. Associated with this radiation is recorded a wave of extinction that was largely confined to the large terrestrial herbivores and their dependent carnivores and scavengers. (Other extinction episodes in the Earth's history had affected plants and marine life, as well as animals.) There appear to be no continents or islands where these accelerated late Pleistocene extinctions precede man's invasion (Martin 1967). Whether men caused these extinctions cannot be known with any certainty, but Martin's overkill hypothesis is clearly consistent with a common property resource model of the economics of megaherbivore hunting (Smith 1975). Thus the large gregarious animals that suffered extinction provided low search cost and high kill value. The lack of appropriation (branding or domestication) provided disincentives for conservation and sustained yield harvesting. There are numerous stampede kill sites (pitfalls and cliffs) in Russia, Europe and North America that indicate wastage killing in excess of immediate butchering requirements. Considering the complex of suitabilities necessary for the remains of such a site to have been preserved, it is likely that only the tip of such phenomena has been observed. Finally, the slow growth, long lives and long maturation of the megafauna made them more vulnerable than other animals to extinction by hunting pressure.

But our model of economizing man need not sustain such a controversial hypothesis as overkill. It is sufficient that the easy, valuable prey disappeared, precipitating a decline in the productivity of hunting. Substitution is to be expected, given a change in relative effort 'prices'. Hence, it is in this late pre-agricultural period that the archaeological record shows the appearance of bows and arrows, seed grinding stones, boiling vessels, boats, more advanced houses, even 'villages' (probably clan group abodes), animal-drawn sledges and the dog (almost certainly derived from domesticating the wolf). These developments strongly suggest the substitution of new tools and techniques for the old, which allowed new products to substitute for the loss of big game that could be harvested by stampeding and/or dispatch with thrusting or throwing weapons. Now the bow and arrow becomes adaptive, and gathering becomes more crucial to maintaining overall food productivity. Whereas formerly, gathering emphasized seeds and plants that could be eaten on the run, now some of the seeds gathered were inedible without grinding, soaking, boiling. All this paraphernalia implies more sedentary, less nomadic, hunting and gathering.

Hence the incentive to invest in facilities such as utensils, sledges and houses. The boat allows fishing, sealing and whaling. The wolf, also characterized by its capacity to apply organization to the hunt, is now enlisted with man in the hunting of the game still available. Perhaps more important, the wolf may have been the model for domesticating other animals since the dog was a companion and pet that enabled children to learn about domesticated animal behaviour. With a more sedentary life, and the accumulation of personal property and real estate, would come more complex property right and contracting arrangements. The study of pre-colonial aboriginal societies in Northwest America and Melanesia reveals the existence of elaborate *multilateral contracting* arrangements in the form of 'ceremonial exchanges' such as the potlatch, kula, moka and abutu (Dalton 1977). The use of valuables or commodity money (bracelets, pearl shells, cowries, young women) in these primitive societies was more complex than that of cash used in nation states with well-defined legal bases for exchange. These valuables not only bought other valuables in ordinary internal or external market exchange, they bought kinship ties with the exchange of women, military assistance when attacked, the right of refuge if invasion required the abandonment of homes, and emergency aid in times of poor harvest, hunting or fishing. In short they bought political stability, and a property right environment that made ordinary exchange and specialization possible. Property was owned by corporate descent lineages and included land, fishing sites, cemetery plots and livestock, but, interestingly, also public goods like crests, names, dances, rituals and trade routes, that could be assigned to many groups or individuals. These practices, which characterize stateless hunter–gatherer aboriginals, demonstrate that the phenomenon of multilateral contracting (Williamson 1983), so common to the market economy in nation states, has ancient origins which antedate the state and the agricultural revolution.

Man's long existence as a hunter had brought knowledge of animals; extinction brought a change in relative costs; gathering brought knowledge of seeds and eggs; life became more sedentary, with property, contracting and exchange becoming more important. Under these more stable conditions it was a short step for mankind to plant for harvest, and/or to husband some of the more docile game that had been hunted previously. With agriculture and herding came a more sophisticated development of the earlier hunter–gatherer institutions of contract, property, exchange and specialization; and ultimately the continuing industrial-communication revolution. But long before these sweeping changes can be seen the dim outline of continuity in the development of man's capacity to adapt by creating cheaper products and techniques to substitute for dearer ones.

See Also

- ▶ Economic Anthropology
- ▶ Hunters, Gatherers, Cities and Evolution
- ▶ Population and Agricultural Growth

Bibliography

Dalton, G. 1977. Aboriginal economies in stateless societies: Interaction spheres. In *Exchange systems in pre-history*, ed. J. Erickson and T. Earle. New York: Academic Press.

Haynes, C.V. 1964. Fluted projectile points: Their age and dispersion. *Science* 145: 1408–1413.

Heizer, R. 1955. Primitive man as an ecological factor. *Krober Anthropological Society Papers* 13: 1–31.

Lee, R.B., and I. DeVore. 1968. *Man the hunter*. Chicago: Aldine.

Lewis, H. 1973. *Patterns of Indian burning in California: Ecology and ethnohistory*. Anthropological Papers No. 1. Romana: Ballena Press.

Martin, P. 1967. Prehistoric overkill. In *Pleistocene extinctions*, ed. P.S. Martin and H.E. Wright Jr. New Haven: Yale University Press.

North, D.C., and R.P. Thomas. 1977. The first economic revolution. *Economic History Review* 30: 229–241.

Pilbeam, D. 1984. The descent of hominoids and hominids. *Scientific American* 250 (3): 60–69.

Smith, V.L. 1975. The primitive hunter culture, pleistocene extinction, and the rise of agriculture. *Journal of Political Economy* 83: 727–755.

Wheat, J.B. 1967. A Paleo-Indian bison kill. *Scientific American* 216 (1): 44–52.

Williamson, O. 1983. Credible commitments: Using hostages to support exchange. *American Economic Review* 73: 519–540.

Huskisson, William (1770–1830)

Roy Green

Huskisson is better remembered for the manner of his death than for his not inconsiderable achievements as a statesman and economist. While it is true that he enjoyed 'little success in public life compared with that which his rare abilities should have commanded' (*Dictionary of National Biography*), there were few major debates which were not enhanced by his contribution. Huskisson first entered Parliament in 1796 and remained a member, with only one short break, for over 30 years. He served in the cabinet from 1823, and held a number of key government posts, including Secretary of the Treasury, President of the Board of Trade and Secretary of State for War and the Colonies. He figured prominently in the Bullion controversy and the subsequent discussion on the resumption of cash payments; and he initiated the process of tariff reform which was to culminate in the repeal of the Corn Laws.

His abilities may be gauged by the tributes paid by his contemporaries. It was said that 'there is no man in Parliament, or perhaps out of it, so well versed in finance, commerce, trade or colonial matters' (Charles Greville, in Melville 1931, p. viii); and that 'the knowledge of theory and practice were never possessed by any one in so high a degree' (Kirkman Finlay, in Huskisson 1831, I, p. 161; also Alexander Baring and Henry Brougham, ibid., pp. 120–121). Indeed, according to some observers, Huskisson might easily have become Chancellor of the Exchequer, but for his almost disingenuous loyalty to George Canning and the offence which he regularly caused to traditional Tory interests. These 'failings' earned him a remarkably fulsome tribute from J.S. Mill: 'With the exception of Turgot, the history of the world does not perhaps afford another example of a minister steadfastly adhering to general principles in defiance of the clamours of the timid and interested of all parties ...' (*Westminster Review*, 1826, *cit*. Tucker introduction to Huskisson 1830, p. xv).

Even his closest supporters, however, could not pretend that Huskisson was an eloquent speaker; to his everlasting shame, he was born and brought up outside London and the Home Counties. As a consequence, no doubt, he was 'a wretched speaker with no command of words, with awkward motions, and a most vulgar, uneducated accent' (Sir Egerton Brydges, cit. Dictionary of National Biography).

Huskisson's interest in political economy began in Paris, where, as a young man, he moved in French liberal circles, and is said to have met Franklin and Jefferson. There, in 1790, he presented a paper on the currency to the monarchist 'Club of 1789'; once the French Government started issuing assignats, however, he resigned from the club and, shortly afterwards, returned to Britain. In 1810, Huskisson had an opportunity to make his mark on British financial policy; he did so in conjunction with Henry Thornton and Francis Horner in the Bullion Report, and then on his own in a pamphlet defending the report against its 'anti-bullionist' critics. This pamphlet, *The Question Concerning the Depreciation of our Currency* (1810), ran to several editions and drew praise not only from Ricardo, as might be expected, but also from the more critical Thomas Tooke (1838–1857, IV, p. 98); its main target was the 'real bills doctrine' pleaded by the Bank of England directors as an adequate principle of limitation even when the currency was inconvertible. In the Parliamentary debates on the Bullion Report, Huskisson likened the views of the Bank directors to those of John Law, and made a strong case for the resumption of cash payments (Fetter 1965, p. 43). After the passage of resumption legislation in 1819, however, Huskisson confessed to private doubts: 'The wheel of depreciation producing high prices, etc., was turning one way whereby many interests suffered and were ruined; to attempt to turn the wheel back, without some equitable adjustment ... has always appeared to me madness' (Letter to J.C. Herries, 20 December 1829, *cit*. Melville 1932, p. 312). The sharp decline in prices which followed resumption particularly affected agricultural products.

A Committee on Agriculture was formed in 1821 whose report – drafted mainly by Huskisson and Ricardo – accepted many of the arguments against the Act of 1819 but came down in favour of its retention. Thomas Attwood, after giving evidence to the Committee, wrote: 'The stupid landowners ... are all as dull as beetles, whilst Huskisson and Ricardo are as sharp as *needles* and as active as bees' (*cit*., Ricardo 1951–1973, VIII, p. 370). A year later, Huskisson headed off Western's motion to reopen the issue with an amendment in the same terms as Montague's resolution of 1696, 'That this House will not alter the Standard of Gold or Silver, in fineness, weight, or denomination'.

During the 1820s, Huskisson became an effective spokesman for the manufacturing interest, defending 'with singular success and ability, the general principles of commercial freedom' (Tooke 1838–1857, V, p. 414). He took part in debates on the silk trade, agricultural protection, tax reform, shipping and the repeal of the Combination Acts; and he was almost alone in foreseeing the crisis of 1825, expressing concern as early as March 1822, 'that this universal Jobbery in Foreign Stock will turn out the most tremendous Bubble ever known' (Hudson Gurney, *cit*. Fetter 1965, pp. 111–112). Having disregarded his warnings, the Bank of England directors sought to blame Huskisson for promoting the crisis:

> Such is the detestation in which he is held in the City that Ld L[iverpool] & Mr. Canning did not think it prudent to summon him to London till all the Cabinet were sent for &, in the discussions with the Bank, he is kept out of sight. He repays them with equal hatred . . .. (Mrs Arbuthnot, 17 December 1825, *cit*. Fetter 1965, p. 117)

In June 1827 Huskisson, responding to a memorandum circulated by James Pennington, wrote of the need to 'prevent . . . those alterations of excitement and depression which have been attended with such alarming consequences to this country'. He went on:

> This, for a long time, has appeared to me one of the most important matters which can engage the attention of the Legislature and the Councils of this country. The subject is certainly intricate and complicated; but the too great facility of expansion at one time, and the too rapid contraction of paper credit (I speak of it in the largest sense) at another, is unquestionably an evil of the greatest magnitude. (*cit*. Fetter 1965, p. 131; also Viner 1937, p. 224)

Huskisson asked Pennington for suggestions as to how these fluctuations could be minimized, and Pennington submitted a second memorandum which was to form the basis of the 'currency principle'.

Huskisson resigned from the government in 1828 over a seemingly trivial but symbolic issue – the allocation of a parliamentary seat to a sparsely populated rural hundred, instead of a manufacturing town. He died soon afterwards in unusual, not to say bizarre, circumstances. On 15 September 1830, he attended the opening ceremony of the Manchester and Liverpool Railway:

> At that moment several engines were seen approaching along the rails between which Huskisson was standing. Everybody made for the carriages on the other line. Huskisson, by nature uncouth and hesitating in his motions, had a peculiar aptitude for accident . . ..On this occasion he lost his balance in clambering into the carriage and fell back upon the rails in front of the Dart, the advancing engine. It ran over his leg . . . He lingered in great agony for nine hours, but gave his last directions calmly and with care, expiring at 9 P.M. (*Dictionary of National Biography*)

That would be the end of the story but for a fine piece of detective work by G.S.L. Tucker and his assistant, Helen Bridge, who in 1976 established beyond reasonable doubt that the author of an anonymously published 1830 tract, *Essays on Political Economy*, was none other than William Huskisson. In addition to the circumstantial evidence of style and argument, the publisher's Commission Ledger was signed by a certain 'George Robertson', a name unknown to political economy at that time. It was then demonstrated by Detective Sergeant D.G. Stuckey of the Document Examination Unit, New South Wales Police, that the signature belonged not to 'George Robertson' at all but to Huskisson's half-brother, Thomas, with whom he was on close terms (Fay 1951, pp. 300–301). Thomas Huskisson was a captain in the Royal Navy; and there is evidence that in return for career advancement (William Huskisson was treasurer of the Navy from 1823 to 1827), he would perform errands of this kind (ibid.).

Although the *Essays* had a poorer reception than if they had appeared under Huskisson's own name, he presumably felt that he could not take the risk of further embarrassing the government with his forthright views. The *Essays* are basically Smithian in approach, and, in most respects, were already superseded by Ricardo's *Principles*. They do, however, propose some important financial reforms (Huskisson 1830, pp. 149–151 and 152–153), repudiate the landowners' monopoly (ibid., p. 255) and, most notably, anticipate J.S. Mill's concept of a 'general glut' (ibid., pp. 448–452 and 454–455). Overall, they epitomize Huskisson's economic philosophy and were even cited approvingly by Marx (1867, p. 495n.); this philosophy was reflected clearly and consistently in a life of ceaseless activity: 'Whatever ridicule might be attempted to be thrown on the science of political economy', he said, 'that science could not be discredited. It was the result of general principles warranted by observation, and constituted the guide in the regulation of political measures' (Huskisson 1831, II, p. 128).

Selected Works

1830. *Essays on political economy*. Canberra: Australia National University, 1976.

1831. *The speeches of the right honourable William Huskisson*. London: John Murray.

References

Fay, C.R. 1951. *Huskisson and his age*. London: Longmans.

Fetter, F.W. 1965. *Development of British monetary orthodoxy, 1797–1875*. Reprinted, Fairfield: Kelley, 1978.

Marx, K. 1867. *Capital*, vol. 1. Moscow: Progress Publishers.

Melville, L. 1931. *The Huskisson papers*. London: Constable.

Ricardo, D. 1951–1973. In *The works and correspondence of David Ricardo*, ed. P. Sraffa. Cambridge: Cambridge University Press.

Tooke, T. 1838–1857. *A history of prices, and of the state of circulation, from 1792 to 1856*. London: P.S. King, 1928.

Viner, J. 1937. *Studies in the theory of international trade*. London: George Allen & Unwin.

Hutcheson, Francis (1694–1746)

Andrew Skinner

Abstract

Holder of the Chair of Moral Philosophy at Glasgow University, Hutcheson, counted Adam Smith among his pupils. His moral philosophy resembled Smith's in emphasizing the role of sentiment, though Smith rejected his notion of an internal moral sense. Hutcheson's economic analysis embraced the division of labour, property, and money. His theory of value, which stressed the role of subjective judgement as a determinant of value in exchange, was influenced by Pufendorf, but Hutcheson went beyond Pufendorf (and foreshadowed Smith) in arguing that goods exchange at a rate that is in part determined by the quantity of labour embodied in them.

Keywords

Barter; Benevolence; Carmichael, G.; Division of labour; Ferguson, A.; Hume, D.; Hutcheson, F.; Locke, J.; Money; Moral sense; Natural rights; Property; Pufendorf, S.; Self-interest; Smith, A.; Supply and demand; Value in exchange; Value theory; *Wealth of Nations*

JEL Classifications

B31

Biographical

Hutcheson was born on 8 August 1694. His father, John, was a Presbyterian minister in Armagh, Ireland, and Francis spent his early years at nearby Ballyrea. In 1702 Francis and his elder brother, Hans, went to live with their grandfather, Alexander Hutcheson, at Drumalig in order to further their schooling. At the age of 14 Francis moved to a small denominational academy at Killyleagh, County Down.

In 1711 Hutcheson matriculated at Glasgow University, where he was particularly influenced by Robert Simson (mathematics), Gerschom Carmichael (moral philosophy), Alexander Dunlop (Greek) and John Simpson (the 'heretical divine'). Hutcheson graduated in 1713 and embarked upon a course of study in theology under Simpson's guidance.

Hutcheson was back in Ireland in 1719 when he was licensed as a probationary minister but moved to Dublin where he established an academy of which he remained head until 1730. His reputation established, Hutcheson was elected to the Chair of Moral Philosophy in Glasgow, succeeding Carmichael. It was as a lecturer that he made his mark, brilliant and stylish, using English rather than Latin. Hutcheson's career as author and teacher amply confirms Adam Smith's famous reference to the 'abilities and virtues of the never-to-be-forgotten' master.

Hutcheson lectured five days a week on natural religion, morals, jurisprudence, and government – an order which was to be followed by Adam Smith on his appointment to the Chair of Moral Philosophy in 1752. On three days he lectured on classical theories of morality, thus contributing (with Dunlop) to a revival of classical learning in Glasgow, which formed an important channel for stoic philosophy; a philosophy which was to have an important influence on Adam Smith. Hutcheson died on 8 August 1746 (his birthday) and was buried in St Mary's churchyard in Dublin.

Social Order

Although this article is concerned primarily with Hutcheson's economic analysis it will be convenient to say a little regarding his ethical work.

Adam Smith identified two key questions which the moral philosopher must confront. First, wherein does virtue consist, and, secondly:

> how and by what means does it come to pass, that the mind prefers one tenor, of conduct to another, denominates the one right and the other wrong; considers the one as the object of approbation, honour and reward, and the other of blame, censure and punishment. (TMS, VII, i.2)

Hutcheson addressed both questions, identifying virtue with benevolence while explaining the processes of judgement in terms of a particular sense, the 'moral sense'. Smith was to reject Hutcheson's answer to the first question on the ground that while important, the emphasis on benevolence neglected the role of self- command and the 'inferior' virtue of prudence. In the same way, while welcoming his master's emphasis on sentiment rather than reason in explaining the means by which the mind forms judgements concerning what is fit and proper to be done or to be avoided, Smith rejected the notion of a special (internal) sense, the moral sense.

The common element evident in the work of Hutcheson, Hume and Smith is the emphasis on sentiment. But they also share another preoccupation, namely the attempt to explain the origins of social order; a crucially important element in the treatment, inter alia, of economic phenomena. The basic task was to explain how it was that a creature endowed with both self- and other-regarding propensities was fitted for the social state.

When we turn to Hutcheson it is to discover marked similarities with the work of his successor, especially in the context of his belief that 'We may see in our species, from the vary cradle, a constant propensity to action and motion' (*System*, I, p. 21). But in some respects the position is subtler than that stated by Smith. To begin with, Hutcheson argued that man has powers of perception which 'introduce into the mind all the materials of knowledge' and which are associated with 'acts of the understanding' (*System*, I, p. 7). Acts of the understanding assist in the isolation of objects to be attained (for example, sources of pleasure) or to be avoided, and culminate in acts of will.

Acts of will, which may be calm or turbulent, were divided in turn into the selfish or the benevolent. Benevolent acts of will which may be described as calm, tend towards the 'universal happiness of others' while the turbulent include 'pity, condolence, congratulation, gratitude'.

Acts of will which are selfish but calm include 'an invariable constant impulse towards one's own perfection and happiness of the highest kind' (*System*, I, p. 9) and do not rule out 'deliberate purposes of injury' (*System*, I, p. 73). The turbulent and selfish embrace 'hunger, thirst, lust, passions for sensual pleasure, wealth, power or fame' (*System*, I, pp. 11–12).

In Hutcheson's case, the problem is that of attaining degree of balance between the turbulent and the calm, the selfish and the benevolent:

> the general tenor of human life is an incoherent mixture of many social, kind, innocent actions, and of many selfish, angry, sensual ones; as one or other of our natural dispositions happens to be raised, and to be prevalent over others. (*System*, I, p. 37)

While Smith was correct in identifying Hutcheson with that school of thought which found virtue to consist in benevolence, there is equally no doubt that he (Hutcheson) gave a prominent place to self-love:

> Our reason can indeed discover certain bounds, within which we may not only act from self-love consistently with the good of the whole; but every mortal's acting thus within these bounds for his own good, is absolutely necessary for the good of the whole; and the want of self-love would be universally pernicious . . . But when self-love breaks over the bounds above mentioned, and leads us into actions detrimental to others, and to the whole; or makes us insensible of the generous kind affections; then it appears vicious, and is disapproved. (1725, III.v)

As in the case of Smith, what is critically important is man's desire to be approved of:

> an high pleasure is felt upon our gaining the approbation and esteem of others for our good actions, and upon their expressing their sentiments of gratitude; and on the other hand, we are cut to the very heart by censure, condemnation, and reproach. (*System*, I, p. 25)

On Hutcheson's argument an important source of control is represented by a capacity for judgement, including moral judgement, which is linked to man's deployment of internal senses such as the 'sympathetic' which differ from external senses such as sight, sound, or taste, and 'by which, when we apprehend the state of others, our hearts naturally have a fellow-feeling with them' (*System*, I, p. 19).

It was Hutcheson's contention that men were inclined to, and fitted for, society: 'their curiosity, communicativeness, desire of action, their sense of honour, their compassion, benevolence, gaiety and the moral faculty, could have little or no exercise in solitude' (*System*, I, p. 34).

This discussion was to lead to Hutcheson's treatment of natural rights and of the state of nature in a manner which is reminiscent of Locke. He also advances the Lockian claim that the state of nature is a state not of war but of inconvenience which can only be resolved by the establishment of government in terms of a complex double contract.

This has been described as the 'Real Whig position' (Winch 1978, p. 46; Robbins 1968) and may explain the considerable influence of Hutcheson's political ideas in the American colonies (Norton 1976). Hutcheson's 'warm love of liberty' was attested by Principal Leechman in his introduction to the *System* (I, pp. xxxv–xxxvi); a sentiment which was echoed by Hugh Blair (Winch 1978, pp. 47–8) in a contemporary review of the book.

While agreeing that an essential precondition of social stability is some system of 'magistracy' (TMS, VII.iv.36), Adam Smith (like Hume) was to emerge as a critic of the contract theory. In addition, he criticized Hutcheson for seeming to imply that self-love was 'a principle which could never be virtuous in any degree or in any direction' (TMS, VII.ii.3.12). But for the economist it is important to note that Hutcheson distinguished often more clearly than did Smith between approval and moral approbation. As Hutcheson put it:

> A penetrating genius, capacity for business, patience of application and labour . . . are naturally admirable and relished by all observers, but with quite a different feeling from moral approbation. (*System*, I, p. 28)

Whatever the differences of emphasis and of analysis which are disclosed in the writings of Hutcheson and Smith, the arguments reviewed in this section are or should be important to the economist for three reasons. First, it appears that social order as a basic precondition for economic activity depends in part upon a capacity for moral judgement. Secondly, it is alleged that the psychological drives which explain economic activity must be seen in a context wider than the economic.

Finally, the argument suggests that all forms of activity are subject to the scrutiny of our fellows.

Economic Analysis

There are five major topics covered in Hutcheson's *System*, which is generally assumed to follow closely the content of his lecture course as a whole. The economic analysis is not given in the form of a single coherent discourse, but rather woven in the broader treatment of jurisprudence. Perhaps for this reason Hutcheson's work did not attract a great deal of attention from early historians of economic thought. But the situation was transformed as a result of Edwin Cannan's discovery of Smith's *Lectures on Jurisprudence*. Cannan recalled that:

> On April 21, 1895, Mr Charles C Maconochie, Advocate, whom I then met for the first time, happened to be present when, in course of conversation with the literary editor of the Oxford Magazine, I had occasion to make some comment about Adam Smith. Mr Maconochie immediately said that he possessed a manuscript report of Adam Smith's lectures on jurisprudence, which he regarded as of considerable interest. (1896, p. xv)

While Cannan's reaction may be imagined, the lectures had the effect of confirming Hutcheson's influence upon his pupil on a broad front, but especially in the area or economic analysis (as distinct from policy). For what Cannan discovered was that the *order* of a large part of Smith's course and its content corresponded closely with what Hutcheson was believed to have taught. It is this correspondence which served to renew interest in Hutcheson's economics with remarkable speed. Quite apart from

Cannan's introduction to the *Lectures*, the same theme is elaborated in his introduction to the *Wealth of Nations* (1904). The link had also been noted, following the publication of the *Lectures*, in the Palgrave *Dictionary of Political Economy* (1896) and received its most elaborate statement in W.R. Scott's *Francis Hutcheson* (1900). The most modern treatment of this kind is to be found in W.L. Taylor's influential work *Francis Hutcheson and David Hume as Predecessors of Adam Smith* (1965).

But Cannan noted something else, namely that it may be that the 'germ of the *Wealth of Nations*' is to be found in Hutcheson's treatment of value (1896, p. xxvi). It is this topic which forms the central feature of the remainder of the present argument although it will be convenient to begin with Hutcheson's views on the division of labour where his influence on Smith may be particularly obvious.

But before we pass on to these subjects, it should be noted that Hutcheson's work on economic topics has its own history. It is evident that he admired the work of his immediate predecessor in the Chair of Moral Philosophy – Gershom Carmichael (1672–1729), and especially his translation of, and commentary on, Samuel Pufendorf. In Hutcheson's address to the 'students in Universities' (Taylor 1965, p. 25) the *Introduction to Moral Philosophy* (1742) is described thus:

> The learned will at once discern how much of this compound is taken from the writing of others, from Cicero and Aristotle, and to name no other moderns, from Pufendorf's smaller work, *De Officio Hominis et Civis Juxta Legem Naturalem* which that worthy and ingenious man the late Professor Gerschom Carmichael of Glasgow, by far the best commentator on that book has so supplied and corrected that the notes are of much more value than the text.

Carmichael's influence as a student of ethics and of jurisprudence has been frequently celebrated, notably by Sir William Hamilton who stated that he may be regarded 'on good grounds, as the true founder of the Scottish school of philosophy' (Taylor 1965, p. 253). But it is to W.L. Taylor that we are indebted for the reminder that Carmichael (and Pufendorf) may have shaped Hutcheson's economic ideas. Taylor concluded that:

> The interesting point for the development of economic thought in all this is the very close parallelism between Pufendorf's *De Officio* and Hutcheson's *Introduction to Moral Philosophy*. Each man covered almost exactly the same field . . . The inescapable conclusion is that Francis Hutcheson took over almost in whole, from Carmichael, the economic ideas of Pufendorf. (1965, pp. 28–2)

The Division of Labour

A key issue for both Hutcheson and Pufendorf arose from the comparison of the social as distinct from the solitary state; or, as Pufendorf put it,

> it would seem to have been more wretched than that of any wild beast, if we take into account with what weakness man goes forth into this world, to perish at once, but for the help of others; and how rude a life each would lead, if he had nothing more than what he owed to his own strength and ingenuity. On the contrary, it is altogether due to the aid of other men, that out of such feebleness, we have been able to grow up, that we now enjoy untold comforts, and that we improve mind and body for our own advantage and that of others. And in this sense of natural state is opposed to a life improved by the industry of men. (*De Officio* 1682, II, pp. 8–9)

This broad line of argument was developed in the *System* (II, p. 4) where Hutcheson offered two specific economic applications. First, he noted that the '*joint* labours of twenty men will cultivate forests, or drain marshes, for farms to each one, and provide houses for habitation, and enclosures for their stocks, much sooner than the *separate* labours of the same number' (*System*, II, p. 289).

Secondly, Hutcheson drew attention to the importance of the *division* of labour:

> Nay 'tis well known that the produce of the labours of any given number, twenty, for instance, in providing the necessaries or conveniences of life, shall be much greater by assigning to one, a certain sort of work of one kind, in which we will soon acquire skill and dexterity, and to another assigning work of a different kind, than if each one of the twenty were obliged to employ himself, by turns in all the different sorts of labour requisite for his subsistence, without sufficient dexterity in any. In the former method each procures a great quantity of goods of one kind, and can exchange a part of it for such goods obtained by the labours of others as he shall stand in need of. One grows expert in tillage, another

> in pasture and breeding cattle, a third in masonry, a fourth in the chace, a fifth in iron-works, a sixth in the arts of the loom, and so on throughout the rest. Thus all are supplied by means of barter with the works of complete artists. In the other method scarce any one could be dextrous and skilful in any one sort of labour. (*System*, II, pp. 288–9)

Property

The discussion of the division of labour implied that members of society are interdependent in respect of the satisfaction of their wants. It also led to two further analytical developments: security of property and the problem of value in exchange (see especially Brown 1987).

Much of the discussion in Book 2, Chapter 6 of the *System* is concerned with 'the right of property'. But Hutcheson also noted that:

> If we extend our views further and consider what the common interest of society may require, we shall find the right of property further confirmed. Universal industry is plainly necessary for the support of mankind. Tho' men are naturally active, yet their activity would rather turn toward the lighter and pleasanter exercises, than the slow, constant, and intense labours requisite to procure the necessaries and conveniences of life, unless strong motives are presented to engage them to these severer labours. Whatever institution therefore shall be found necessary to promote universal diligence and patience, and make labour agreeable or eligible to mankind, must also tend to the public good; and institutions or practices which discourage industry must be pernicious to mankind. Now nothing can so effectually excite men to constant patience and diligence in all sorts of useful industry, as the hopes of future wealth, ease, and pleasure to themselves, their offspring, and all who are dear to them, and of some honour too to themselves on account of their ingenuity, and activity, and liberality. All these hopes are presented to men by securing to every one the fruits of his own labours, that he may enjoy them, and dispose of them as he pleases.
>
> Nay the most extensive affections could scarce engage a wise man to industry, if no property ensued upon it. (*System*, II, pp. 320–1)

Hutcheson attached a great deal of importance to freedom of choice and in fact concluded this phase of the argument by rejecting any suggestion that 'magistrates' may be involved, passages that may well have attracted the attention of the youthful Smith (*System*, II, pp. 322–3).

The Theory of Value

It is Hutcheson's treatment of value that shows most clearly the influence of Pufendorf and of Carmichael where the latter observed that:

> In general we may say that the value of goods depends upon these two elements, their *scarcity*, and the difficulty of acquiring them. Furthermore, *scarcity* is to be regarded as combining two elements, the number of those demanding, and the *usefulness* thought to adhere in the good or service, and which can add to the utility of human life. (Quoted in Taylor 1965, p. 65)

Pufendorf's analysis received its most elaborate statement in the *De Jure*, in the long chapter 'On Price' (Book 5, Chapter 1). The most succinct statement, on which Carmichael commented, is to be found in Book 1, Chapter 14, of *De Officio*.

Hutcheson opened his analysis of the problem by pointing out that the 'natural ground of all value or price is some sort of use which goods afford in life', adding that 'by the use causing a demand we mean not only a natural subserviency to our support, or to some natural pleasure, but any tendency to give any satisfaction by prevailing custom or fancy, as a matter of ornament or distinction' (*System*, II, pp. 53–4). He continued:

> But when some aptitude to human use is presupposed, we shall find that the prices of goods depend on these two jointly, the *demand* on account of some use or other which many desire, and the *difficulty* of acquiring, or cultivating for human use. When goods are equal in these respects men are willing to interchange them with each other; nor can any artifice or policy make the values of goods depend on any thing else. When there is no *demand*, there is no price, where the *difficulty* of acquiring never so great: and where there is no *difficulty* or labour requisite to acquire, the most universal *demand* will not cause a price; as we see in fresh water in these climates. Where the demand for two sorts of goods is equal, the prices are as the difficulty. Where the difficulty is equal, the prices are as the demand. (*System*, II, p. 54)

Hutcheson then added two points which are reminiscent of Pufendorf in commenting on issues that affect supply price and the rate of exchange. First, he argued:

> In like manner by difficulty of acquiring, we do not only mean great labour or toil, but all other circumstances which prevent a great plenty of the goods or performances demanded. Thus the price is increased by the rarity or scarcity of the materials in nature, or such accidents as prevent plentiful crops or certain fruits of the earth; and the great ingenuity and nice taste requisite in the artists to finish well some works of art, as men of such genius are rare. The value is also raised, by the dignity of station in which, according to the custom of the country, the men must live or provide us with certain goods, or works of art. Fewer can be supported in such stations than in the meaner; and the dignity and expense of their stations must be supported by the higher prices of their goods or services. Some other singular considerations may exceedingly heighten the values of goods to some men, which will not affect their estimation with others. These above mentioned are the chief which obtain in commerce. (*System*, II, pp. 54–5)

As regards the *rate of exchange*, Hutcheson commented:

> In commerce it must often happen that one may need such goods of mine as yield a great and lasting use in life, and have cost a long course of labour to acquire an cultivate, while yet he has none of those goods I want in exchange, or not sufficient quantities; or what goods of his I want, may be such as yield but a small use, and are procurable by little labour. In such cases it cannot be expected that I should exchange with him. I must search for others who have the goods I want, and such quantities of them as are equivalent in use to my goods, *and require as much labour to produce them*; and the goods on both sides must be brought to some estimation or value. (*System*, II, p. 53)

But although these positions do not differ significantly from those of Pufendorf, Hutcheson does seem to have taken notice of two additional points. First, he seems to suggest, as the above quotation indicates, that goods will exchange at a rate that will be in part determined by the quantity of labour embodied in them (a point later taken up by Smith). Secondly, he noted in a passage that may have been ‘foreshadowed’ by Pufendorf, that some commodities: ‘of great use have no price, either because they are naturally destined for community, or cannot come into commerce but as appendages of something else, the price of which may be increased by them, though they cannot be separately estimated’ (Hutcheson 1742b; quoted in Taylor 1965, p. 66).

Money

The discussion of value in exchange led Hutcheson on quite logically to consider the medium of exchange, namely money, and here too he followed an old tradition which had already been commented upon by Pufendorf. In Book I, Chapter 14 of *De Officio* he noted the inconvenience of exchange by barter:

> But after men departed from their primitive simplicity and various kinds of gain were introduced, it was readily understood that common value alone was not sufficient for the transactions of men’s affairs and their increased dealings.

Once more, Hutcheson followed suit in explaining the problems of barter and the need to establish a standard or ‘common measure’ when settling the ‘values or goods for commerce’.

> The qualities requisite to the most perfect standard are these: it must be something generally desired so that men are generally willing to take it in exchange. The very making of any goods the standard will of itself give them this quality. It must be *portable*; which will often be the case if it is rare, so that small quantities are of great value. It must be *divisible* without loss into small parts, so as to be suited to the values of all sorts of goods: and it must be *durable*, not easily wearing by use, or perishing in its nature. One or other of these prerequisites in the standard, shews the inconvenience of many of our commonest goods for that purpose. The man who wants a small quantity of my corn will not give me a work-beast for it, and his beast does not admit division. I want perhaps a pair of shoes, but my ox is of far greater value, and the other may not need him. I must travel to distant lands, my grain cannot be carried along for my support, without insufferable expense, and my wine would perish in the carriage. ‘Tis plain therefore that when men found any use for the rarer metals, silver and gold, in ornaments and utensils, and thus a demand was raised for them, they would soon also see that they were the fittest standards of commerce, on all the accounts above-mentioned. (*System*, II, pp. 55–6)

The familiar arguments concerning the need for coinage and the dangers of debasement follow

(*System*, II, ch. 12), while there is also a hint of the need to find an invariable measure of value at least over long periods of time.

> We say indeed commonly, that the rates of labour and goods have risen since these metals grew plenty; and that the rates of labour and goods were low when the metals were scarce; conceiving the value of the metals as invariable, because the legal names of the pieces, the pounds, shillings, or pence, continue to them always the same till a law alters them. But a days digging or ploughing was as uneasy to a man a thousand years ago as it is now, tho' he could not then get so much silver for it: and a barrel of wheat, or beef, was then of the same use to support the human body, as it is now when it is exchanged for four times as much silver. Properly, the value of labour, grain, and cattle, are always pretty much the same, as they afford the same uses of life, where no new inventions of tillage, or pasturage, cause a greater quantity in proportion to the demand. 'Tis the metal chiefly that has undergone the great change of value, since these metals have been in greater plenty, the value of the coin is altered tho' it keeps the old names. (*System*, II, p. 58)

The analytical section of the work is concluded in the following chapter where Hutcheson demonstrated the need for *interest*, since if it were prohibited 'none would lend' (*System*, II, p. 72). He argued that the rate would be determined 'by the state of trade and the quantity of coin, recognizing that 'as men can be supported by smaller gains upon proportion upon their large stocks, the profit made upon any given sum employed is smaller, and the interest the trader can afford must be less' (*System*, II, p. 72). Hutcheson was well aware of the relationship between interest and other forms of return, such as rent, and also introduced an allowance for risk. In sum, an interesting and often sophisticated analysis, taken as whole, which is likely to have made an impression of the youthful Smith.

Conclusion

This article has pursued a number of themes. First, it endeavours to establish a link between Hutcheson and Pufendorf. Secondly, the argument has elaborated on the parallel between Hutcheson's order of argument and that developed by Adam Smith as suggested by W.R. Scott (1900, 1932), Cannan (1896, 1904) and W.L. Taylor (1965). While these parallels are important, it is noteworthy that Smith's treatment of economic topics is worked out as a single discourse, while Hutcheson's treatment is woven into the broader fabric of his analysis of jurisprudence. Finally, the argument has sought to give prominence to the role of subjective judgement as regards the determinants of value in exchange.

Edwin Cannan, as we have seen, considered that Hutcheson's emphasis on the utility of goods to be acquired and on the effort (disutility) involved in creating the goods to be exchanged, with the attendant emphasis on demand and supply considerations, provided the 'kernel' of the *Wealth of Nations*. Taylor, on the other hand, suggested that Smith's concern with material welfare served to obscure the line of argument set out by Hutcheson. Robertson and Taylor concluded that:

> It is evident that the *magnum opus* was cast in a mould of a powerful unifying conception. Now within this framework it is evident that the measurement, in real terms, of the wealth of nations, and in particular of its progress would seem to call for some unvarying standard of value which would enable valid comparisons to be made through time ... for this reason, if for no other, it does not appear inexplicable that Adam Smith no longer paid so much attention to the lines of argument taken over from Hutcheson, which had served well enough in the *Lectures*. (1957, pp. 194–5)

What Robertson and Taylor did not note was that Smith's preoccupation with a real measure of value may also have owed much to Hutcheson (Skinner 1996, 148–50).

Selected Works

All citations of *System* in the text refer to *A System of Moral Philosophy* (1755). Hutcheson's *Collected Works* are published by Georg Olms, Hildesheim, 1969.

1725. *Inquiry into the original of our ideas of beauty and virtue*. London, 2nd edn, 1726.

1725–26. Reflections upon laughter and remarks on the fable of the bees. *Dublin Journal* No. 11 (5 June 1725); No. 12 (12 June); No.13 (19 June). Also in Hibernicus, *Letters* (1729).

1728a. *Essay on the nature and causes of the passions, with illustrations upon the moral sense*. London and Dublin.

1728b. Letters between the late Mr. G. Burnet and Mr. Hutcheson. *London Journal*.

1742a. *The meditations of Marcus Aurelius. Newly translated from the Greek, with notes and an account of his life*. Glasgow.

1742b. *A short introduction to moral philosophy in three books, containing the elements of ethics and the law of nature*. Glasgow.

1755. *A system of moral philosophy in three books. Published from the original Ms. by his son. Francis Hutcheson MD to which is prefaced Some account of the life, writings and character of the author. By the Reverend William Leechman, D.D., Professor of Divinity in the same University*. Glasgow.

Bibliography

Brown, V. 1987. Value and property in the history of economic thought: An analysis of the emergence of scarcity. *Oeconomia* 7: 85–112.

Campbell, T.D. 1982. Francis Hutcheson. In *The origins and nature of the Scottish enlightenment*, ed. R.H. Campbell and A.S. Skinner. Edinburgh: John Donald.

Cannan, E. 1896. *Adam Smith's lectures on justice, police, revenue and arms*. Oxford: Clarendon Press.

Cannan, E., ed. 1904. *The wealth of nations*. London: Methuen.

Hutcheson, T. 1988. *Before Adam Smith: The emergence of political economy 1622–1776*. Oxford: Blackwell.

Kaye, F.B., ed. 1924. *The fable of the bees*. Oxford: Clarendon Press.

McCosh, J. 1875. *The Scottish philosophy from Hutcheson to Hamilton*. Princeton: Princeton University Press.

Meek, R.L. 1976. New light on Adam Smith's Glasgow lectures on Jurisprudence. *History of Political Economy* 8: 439–477.

Naldi, N. 1993. Gershom Carmichael on demand and difficulty of acquiring. *Scottish Journal of Political Economy* 40: 456–470.

Norton, D.F. 1976. Francis Hutcheson in America. *Studies on Voltaire and the Eighteenth Century* 154: 1547–1568.

Pesciarelli, E. 1986. On Adam Smith's Glasgow lectures on Jurisprudence. *Scottish Journal of Political Economy* 33: 74–85.

Robbins, C. 1954. When is it that colonies may turn independent: An analysis of the environment and politics of Francis Hutcheson. *William and Mary Quarterly* 11: 214–251.

Robbins, C. 1968. *The eighteenth century Commonwealth man*. New York: Atheneum.

Robertson, H.M., and W.L. Taylor. 1957. Adam Smith's approach to the theory of value. *Economic Journal* 67: 181–198.

Scott, W.R. 1900. *Francis Hutcheson: His life, teaching and position in the history of philosophy*. Cambridge: Cambridge University Press.

Scott, W.R. 1932. Francis Hutcheson. In *Encyclopaedia of the social sciences*, ed. E.R.A. Seligman, vol. 7. New York: Macmillan.

Skinner, A.S. 1996. *A system of social sciences: Papers relating to Adam Smith*. Oxford: Clarendon Press.

Skinner, A.S. 2006. Francis Hutcheson. In *A history of Scottish economic thought*, ed. Alexander Dow and Sheila Dow. London: Routledge.

Smith, A. 1759. In *The theory of moral sentiments (TMS)*, ed. D.D. Raphael and A.L. Macfie. Oxford: Clarendon Press. 1976.

Smith, A. 1776. In *The wealth of nations*, ed. R.-H. Campbell, A.S. Skinner, and W.B. Todd. Oxford: Clarendon Press. 1976.

Taylor, W.L. 1965. *Francis Hutcheson and David Hume as predecessors of Adam Smith*. Durham: University of North Carolina Press.

von Pufendorf, S. 1682. *De Officio et Civis Juxta Legem Naturalem Libri Duo*. Trans. F.G. Moore, 2 vols. Oxford: Oxford University Press, 1927.

von Pufendorf, S. 1688. *De Jure Naturae et Gentium Libri Octo*. Trans. C.H. W.A. Oldfather. Oxford: Clarendon Press, 1934.

Winch, D. 1978. *Adam Smith's politics: An essay in historiographic revision*. Cambridge: Cambridge University Press.

Hutchison, Terence Wilmot (1912–2007)

Roger E. Backhouse

Keywords

Economic policy; Economic theory; History of economics; Hutchison, Terence Wilmot; Machlup, Fritz; Ricardo, David; Robbins, Lionel; Robinson, Joan; Statistics; Truth in economics

JEL Classifications

B31

Terence Hutchison, a specialist in economic methodology and the history of economic thought, defended the idea that, if economics was to make progress, economic propositions needed to be testable and confronted with evidence. This, together with scepticism about theory based on the assumption of perfect knowledge, informed not only his methodological writing but also his work on the history of economics.

Career

Hutchison was born in Bournemouth on 13 August 1912, and attended Tonbridge School. He went to Cambridge in 1931, to read classics, but switched to economics in which he had Joan Robinson as his tutor, obtaining a first in 1934. Though much of his subsequent work can be seen as a rebellion against his tutor's economics and her politics, he acknowledged her role in training him to think. In his final year he picked up some of Wittgenstein's ideas from two of his friends, to whom Wittgenstein was dictating the lectures that comprised his *Blue Book*. Hutchison attended the now-famous lectures in which John Maynard Keynes worked his way towards the *General Theory*, and later rued the loss of his lecture notes in his wartime travels.

After a year spent going to lectures at the London School of Economics and reading widely, in 1935 he obtained a job as *Lektor* in Bonn, where his main duty was to give lectures which could be on any subject, so long as they were in good English. He remained there for around three years, learning German and developing the interest in German economic and methodological writing, the latter having been stimulated by his undergraduate exposure to Wittgenstein, that ran through all his work. While there he married. As his wife was German, they decided not to move to England, but to Baghdad, where he taught at a teacher training college. With the coming of a pro-Nazi government which wanted to reduce British influence there, he managed to get his family out, via Basra, to Bombay. A while later, he was allowed out to join them, and he joined up. He served on the Northwest Frontier and later in Egypt where he worked as an intelligence officer. He spent the last years of the war in Delhi, at one point working with All India Radio.

Hutchison's British university career began, in 1946, with a year at Hull, after which he moved to the London School of Economics. There, working alongside Lionel Robbins, who shared and stimulated his interest in continental European writing, he taught courses on the history of economic thought since 1870 and on the history of economic controversies. In 1956 he was appointed Mitsui Professor of Economics at the University of Birmingham, the position he held until his retirement in 1978. He taught the history of economic thought until 1980, when university regulations forced him to stop. In retirement, his research continued unabated till only a few years before his death.

Away from his academic pursuits, he had a passion for cricket. He played the game in Egypt during the war, and in the 1950s became a good club cricketer. He first visited Lords (Middlesex versus the Australians) with his mother in 1921, and during the final match between England and Australia in 2005, he appeared on television to give an account of the corresponding game in 1926 (perhaps he was by then the only person alive who had seen all four days of that match).

Economic Methodology and the History of Economic Thought

Hutchison's reputation was established with his first book, *The Significance and Basic Postulates of Economic Theory* (1938). This was a response to the recently published *Essay on the Nature and Significance of Economic Science* (1932/1935) in which Lionel Robbins had defended economic theory as a body of propositions deduced from the assumption of scarcity. Hutchison argued that most economic theory comprised tautologies that said nothing about the real world. Economists should instead seek to develop testable propositions and confront them with evidence. The book's significance lay partly in its being the first attempt systematically to apply to economics philosophical ideas being developed in the 1930s, the most prominent of which went under the label of logical positivism.

Hutchison was particularly critical of any theorizing based on the assumption of perfect knowledge. The book received unexpected attention when it was the subject of a 32-page review article, '"What is truth" in economics?' in the *Journal of Political Economy* for 1940 by the eminent Chicago economist Frank Knight, to which Hutchison replied from wartime Baghdad (Knight 1940; Hutchison 1941).

Though Hutchison continued to emphasize testability and the limitations of theorizing based on perfect knowledge, one strand of his methodological work involved engaging with ideas coming from the philosophy of science. In the 1950s he became involved in an exchange with Fritz Machlup, after being described as an 'ultra-empiricist' (Machlup 1955, 1956; Hutchison 1956). The framework within which this debate, over the extent to which propositions needed to be testable, took place reflected the concerns of the so-called 'received view, then dominant in the philosophy of science. In the 1970s, Hutchison brought detailed knowledge of the history of economics to bear on the question of whether economics had exhibited revolutionary changes corresponding to those that Thomas Kuhn and Imre Lakatos claimed to have identified in the history of science (Hutchison 1976, 1978, chapter 3).

This knowledge of the history of economics was first demonstrated in *A Review of Economic Doctrines, 1870–1929* (1953b), a book that arose out of the course Hutchison taught at LSE, which provided a systematic coverage of the subject from the date of the so-called marginal revolution to the onset of the Great Depression. It was unjustly overshadowed by the appearance of Joseph Schumpeter's posthumous magnum opus a year later. Methodological themes were never far from the surface. Interestingly, the book concluded with a discussion of the growth of economic statistics, on which what he thought 'the most spectacular progress in economic knowledge was necessarily being founded' (p. 427). This view that the development of economic statistics was the main example of progress in economics was one that he maintained throughout his career (see, for example, Hutchison 1977, chapter 2; 1992, 1994, chapter 8). He became increasingly critical of theoretical work that was not grounded in empirical work, criticizing the 'crisis of abstraction' of the 1970s (Hutchison 1977) and later the 'formalist revolution' (Hutchison 1992, 2000) and the literature that developed from around the 1980s, dismissing a focus on prediction as outdated positivism.

The other strand in Hutchison's methodological work was analysis of policy. "*Positive*" *Economics and Policy Objectives* (1964), though a methodological book that sought to bring clarity to policy discussions through applying the positive-normative discussion, had a strong historical dimension, analysing economists' statements over several centuries. Most prominent, however, was *Economics and Economic Policy in Britain, 1946–1966* (1968). This examined what economists had said on economic policy, in some instances contrasting this with what they later claimed to have said. He followed this up with an essay, 'Economic knowledge and ignorance in action', which showed that, despite claims to the contrary, economists simply did not agree on the questions of whether sterling should have been devalued in the 1960s, or whether Britain should have entered the European Community (Hutchison 1977, chapter 5). He clearly delighted in pointing out how reviewers considered it an outrage to hold economists to account for claims they had made in newspaper articles or correspondence columns and the suggestion that this was, somehow, merely journalism. His own view was that to understand the policy process it was necessary to take account of economists' views, wherever they were published.

Though concerned throughout with methodological questions and with what had shaped modern economics, his interests extended much further back. *Before Adam Smith* (1985) was the first English-language work to analyse systematically the entire century of economic writing before Adam Smith's *Wealth of Nations*. As his use of the phrase 'contentious essays' in one of his book titles suggests, he never shirked controversy, often challenging widely accepted beliefs about major figures in economics. As with his work on economists' policy advice, he repeatedly pointed out inconsistencies in the statements of economists who upheld dogmatic views. A particular target was the Marxian ideology of his former teacher, Joan Robinson, and Maurice Dobb, and the way it coloured their interpretation of the past. He believed that readers of their historical interpretations should be informed about their views on Stalin's Soviet Union and Mao's cultural revolution (Hutchison 1981, chapter 3). He argued that early 'marginalists' were not unqualified supporters of laissez-faire, concerned to defend capitalism against Marxist critics, but supporters of extensive pragmatic government intervention in economic activity. Similarly, he pointed out that in the early 1930s the differences between A. C. Pigou and Keynes were slight: Pigou advocated fiscal cures for unemployment and Keynes attributed part of the problem to the rigidity of money incomes (Hutchison 1978: 179).

Hutchison's most controversial target was David Ricardo, who he saw as the source of the excessively abstract theorizing that plagued modern economics (1952, 1953a, 1978, 1994). When reviewing Piero Sraffa's edition of David Ricardo's collected works, he feigned surprise that its sponsor had been the Royal Economic Society, not the Moscow State Publishing house (Hutchison 1952: 421). He questioned not only the Marxist interpretation of Ricardo but, even more controversially, made the heretical suggestion that Ricardo was less original and less central to the history of economics than was commonly assumed. Decades later (1994, chapter 5), he ridiculed the idea that this believer in the sanctity of private property was, despite his influence on Marx, a man of the left. Ricardo was, he claimed, 'something of an innocent abroad, whose inconsistent ideas ... fell into the hands of people too keen on exploiting them for their own ideological purposes, and who had to pretend that these inconsistencies were not there' (Hutchison 1994: 99).

However, his criticisms were not just directed against those on the left. He also raised questions about Friedrich Hayek and the Austrians (Hutchison 1981, 1994). The common theme running through his writing was the need for clear thinking informed by knowledge of what economists had actually said.

See Also

- ▶ Falsificationism
- ▶ History of Economic Thought
- ▶ Methodology of Economics
- ▶ Philosophy and Economics

Selected Works

1938. *The significance and basic postulates of economic theory*. London: Macmillan.

1941. The significance and basic postulates of economic theory: A reply to professor knight. *Journal of Political Economy* 49(5): 732–750.

1952. Some questions about Ricardo. *Economica* 19: 415–432.

1953a. Ricardo's correspondence. *Economica* 20: 263–273.

1953b. *A review of economic doctrines, 1870–1929*. Oxford: Clarendon Press.

1956. Professor Machlup on verification in economics. *Southern Economic Journal* 22(4): 476–483.

1964. "*Positive*" *economics and policy objectives*. London: George Allen and Unwin.

1968. *Economics and economic policy in Britain, 1946–1966*. London: George Allen and Unwin.

1976. On the history and philosophy of science and economics. In *Method and appraisal in economics*, ed. S.J. Latsis, 181–206. Cambridge: Cambridge University Press.

1977. *Knowledge and ignorance in economics*. Oxford: Basil Blackwell.

1978. *On revolutions and progress in economic knowledge*. Cambridge: Cambridge University Press.

1981. *The politics and philosophy of economics: Marxians, Keynesians and Austrians*. Oxford: Basil Blackwell.

1985. *Before Adam Smith: The emergence of political economy, 1662–1776*. Oxford: Basil Blackwell.

1992. *Changing aims in economics*. Oxford: Basil Blackwell.

1994. *The uses and abuses of economics: Contentious essays on history and method*. London: Routledge.

2000. *On the methodology of economics and the formalist revolution*. Cheltenham: Edward Elgar.

Acknowledgments This article draws on an obituary written by the author for The Times, published on December 5, 2007. See Coats (1983a, b) for appraisals of Hutchison's work, and the interviews recorded in Tribe (1997) and Hart (2002).

Bibliography

Coats, A.W. 1983a. Half a century of methodological controversy in economics: As reflected in the writings of T.W. Hutchison. In *Methodological controversy in economics: Historical essays in honor of T.W. Hutchison*, ed. A.W. Coats, 1–42. Greenwich: JAI Press.

Coats, A.W. 1983b. T.W. Hutchison as a historian of economics. In *The craft of the historian of economic thought. Research in the history of economic thought and methodology: A research annual*, ed. W.J. Samuels, vol. 1, 187–208. Greenwich: JAI Press.

Hart, J. 2002. A conversation with Terence Hutchison. *Journal of Economic Methodology* 9 (3): 359–377.

Knight, F.H. 1940. 'What is truth?' in economics. *Journal of Political Economy* 48 (1): 1–32.

Machlup, F. 1955. The problem of verification in economics. *Southern Economic Journal* 22 (1): 1–21.

Machlup, F. 1956. Rejoinder to a reluctant ultra-empiricist. *Southern Economic Journal* 22 (4): 483–493.

Robbins, L.C. 1932. *Essay on the nature and significance of economic science*. London: Macmillan. 2nd ed., 1935.

Tribe, K. 1997. Terence Hutchison. In *Economic careers: Economics and economists in Britain, 1930–1970*. London: Routledge, chapter 8.

Hymer, Steven Herbert (1934–1974)

David M. Gordon

Hymer was born on 15 November 1934 in Montreal, Canada, and died tragically at the age of 39 in a car accident, returning from a winter holiday, on a New York State thruway in February 1974.

Hymer began his study of economics as an undergraduate at McGill University and then received his PhD in economics from MIT in 1960. He worked in Ghana for several years in the early 1960s and then returned to the United States to teach at Yale from 1964 to 1970. He moved increasingly in radical and then Marxian directions in the late 1960s. Having been denied tenure by Yale – a common fate at elite US graduate schools for leftists of his generation – he moved to the Graduate Faculty of the New School for Social Research, where he helped found and then foster a political economy programme until his sudden death in 1974.

Hymer's main analytic contributions flowed from his analyses of foreign direct investment by multinational corporations. As early as his seminal dissertation (1960), Hymer broke away from international trade theory, viewing foreign direct investment as a consequence of the particular internal contradictions of multinational enterprises and their drive to extend territorial control. Despite his short productive working life, Hymer's work in this area had wide-ranging influence in both the advanced and developing worlds in shaping both analysis and policy discussions.

Though less widely known for this work, Hymer was also making important contributions in his last several years to the articulation of a modern, complex, analytically rigorous Marxian political economy. Some of his most original and provocative papers in this effort, along with his best essays on multinationals and the global economy, were posthumously collected and published in *The Multinational Corporation* (1979).

Selected Works

1960. *The international operations of national firms: A study of direct foreign investment*. PhD dissertation. Cambridge, MA: MIT Press, 1976.

1979. In *The multinational corporation: A radical approach*, ed. R.B. Cohen et al. New York: Cambridge University Press.

Hyndman, Henry Mayers (1842–1921)

Anthony Wright

A British Marxist theorist and politician, Hyndman was born in London to a prosperous merchant family of staunchly Conservative politics; he was educated at Trinity College, Cambridge. He became in turn a journalist, imperial traveller and financial adventurer. An enthusiast for Empire, he stood for Parliament in 1880 as an independent on a Tory radical programme but withdrew from the contest. Increasingly acquainted with continental socialism, he read Marx's *Capital* in French in 1880 and became personally acquainted with Marx in London. This began the process whereby Hyndman, during the 1880s, emerged as the pioneer of British Marxism, the founder and leader of a Marxist party (the Social Democratic Federation) and the leading theorist and propagandist of Marxism in Britain. In essentials, this was the role he continued to play for the rest of his life.

Hyndman has had a bad press. In part this may be attributed to the easy caricature of him as the Marxist cricketer, stockbroker and national chauvinist, armed with top hat and frock coat. In part, too, it derives from his overbearing personality and sectarian political leadership. However, it is also directly related to the nature of his presentation of Marx's economic theory. He set himself the task of explaining this theory to the British public, relating it to British conditions, and drawing the appropriate political lessons from it. His *England for All* (1881), with its indirect tribute to Marx's work but omission of his name (thereby beginning the personal breach with Marx and Engels), began this task, which was then taken further in his best book, *The Historical Basis of Socialism in England* (1883), with its application of Marxist economic theory to the economic history of England since the 15th century. Its preface recorded his 'indebtedness to the famous German historical school of political economy headed by Karl Marx, with Friedrich Engels and Rodbertus immediately following'.

Hyndman's presentation of Marxist economics was narrowly literal and inflexible, which meant that he could neither develop it creatively nor defend it against its critics with sufficient rigour. When he departed from Marx's own position this was not because of any intention to do so but because he had either failed to understand Marx on the point, or had access only to a limited range of Marx's work, or because when he cited other economic authorities (such as Rodbertus and Lassalle) he was unaware of Marx's disagreements with them. Hence his exposition of the Lassallean 'iron law of wages' as Marxist orthodoxy. In the 1880s, on the basis of Marx's work then available, it was certainly possible to present this as Marx's own position, but Hyndman's later and most developed discussion of economics in his *Economics of Socialism* (1896) showed him still substantially attached to a theory by then repudiated in Marx's mature work. It was on the basis of this doctrinal position that Hyndman poured scorn on the trade unions for the futility of their economic activities.

Similar limitations prevented Hyndman from defending a tenable version of Marx's theory of value when this came under criticism and discussion in the 1880s, especially in Wicksteed's critique of it in terms of Jevonian marginal utility theory. If Fabian intellectuals like Shaw and Webb could respond to this critique by restating the economic case against capitalism in terms of a theory of economic rent rather than of Marxist surplus value, Hyndman lacked the equipment to mount an effective counter-offensive of his own. He continued to be a vigorous propagandist for what he understood as Marxist economic orthodoxy, but the intellectual battle was lost and it was left to a later generation of British Marxist economists to take the argument further.

Selected Works

1883. The historical basis of socialism in England. London: Kegan Paul.

1896. The economics of socialism. London/Boston: Small, Maynard and Co.

Bibliography

Collins, H. 1971. The Marxism of the social democratic federation. In *Essays in labour history 1886–1823*, ed. A. Briggs and J. Saville. London: Macmillan.

Hobsbawm, E. 1964. Hyndman and the SDF. In *Labouring men*, ed. E. Hobsbawm. London: Weidenfeld & Nicolson.

Tsuzuki, C. 1961. *H.M. Hyndman and British socialism*. Oxford: Oxford University Press.

Hyperinflation

Juan Pablo Nicolini

Abstract

A hyperinflation occurs when price indexes of broadly defined baskets of goods increase at extremely high rates. As such, hyperinflations are rare. However, the few known cases share many things in common. First, they can occur only in paper currency systems that are not pegged by the central bank to any good. Second, they occur when the quantity of paper currency also grows at extremely high rates. Finally, the force behind the process is always a fiscal imbalance that is financed by issuing currency.

Keywords

Bretton Woods system; Budget deficits; Commodity money; Convertibility; Fiat money; Fisher, I.; German hyperinflation; Gold standard; Hyperinflation; Inflation; Laffer curve; Paper money; Price control; Price stability; Quantity theory of money; Seigniorage; Stabilization policy; Wage control

JEL Classifications

D4; D10

Price stability shares with a healthy knee a particular feature: both are precious, but you do not realize how much until you miss them. When you run, your knees perform amazing functions, without you even being aware of them. That is price stability. Under some circumstances, one of your knees may be under some stress and you may be forced to use medication to be able to run well. While you run, you are aware of your knee. That is inflation. Eventually, your knee hurts so much you can only walk. That is high inflation. Finally, in the worst case, your knee is broken and you must lie in bed. That is hyperinflation.

Money – that is, a commodity that is widely used as a medium of exchange – has been in use in the world since commerce became a social activity. However, to the extent that money was a particular commodity or was paper money but pegged to a commodity like silver or gold, there was no risk of long-run inflation.

From the point of view of the theory, this premise comes from the quantity equation that was first formalized by Irving Fisher (1934). He argued that the general level of prices was a constant proportion to the ratio of the supply of currency and some index of the total quantity of goods that are traded in a year. Thus, there cannot be long-run inflation without long-run growth of the net supply of the commodity that serves as money or that backs the paper money in circulation, where by ‘net supply’ I mean the rate of growth of money in excess of the growth rate of the index of total goods.

The first known example of inflation occurred during the 16th century in Europe, precisely because of the increase in the supply of gold and silver that came from South America after the Spanish conquest. It is interesting to recall, however, that this first inflation was roughly 100 per cent during the whole century or, equivalently, 0.7 per cent a year. According to the theory, this means that the net supply of gold and silver doubled in 100 years. (The ability of Fisher’s quantity framework to explain low inflation events during relatively short periods of time like a few years has been rightly called into question. However, for the kind of episodes that I discuss here, which involve very high inflation rates, this conceptual framework is perfectly suitable. See Marcet and Nicolini 2005, and all references therein.)

By 1900 paper money was the norm, but all economies were functioning under some form of commodity standard in the sense that money was backed by some commodity, typically gold. Governments would suspend convertibility in some circumstances, like wars, but would eventually restore it. Thus, the ability to increase the net supply of paper money depended on the ability of the issuer to accumulate the commodity that backed it. As a consequence, the economic history of the world does not have records of persistent increases in the general level of prices up to the 20th century, except for the cases of the exceptional gold and silver inflows after the Spanish conquest of America mentioned above.

In a seminal paper, Cagan (1956) defined monthly inflation rates that exceed 50 per cent a month as hyperinflations. To generate a hyperinflation according to this definition, Columbus would have had to double Europe's net supply of gold and silver in a little less than two months!

The 20th century witnessed, among other things, a key change in the functioning of our monetary systems. Today, almost without exception, all modern economies function under fiat money arrangements in the sense that paper money circulates, is widely accepted and used in transactions, and is not backed to any particular commodity. Thus, the size of its net supply depends only on the will of the issuer.

All episodes of hyperinflation we observed during the 20th century, no matter how we define them, and with absolutely no exception, occurred during periods of unbacked paper money. All of them, no matter how we define them and with absolutely no exception, occurred during periods in which the net supply of paper money increased at enormous rates. And all of them occurred in times of substantial fiscal imbalances, represented by excessive government expenditures, inadequate government revenues or a huge government debt burden – or a combination of these.

The first burst of hyperinflations occurred in the 1920s in countries that lost the First World War, most notably Germany and Hungary. Sargent (1992) provides a very neat description of the causes and remedies for each of the cases. It is remarkable that the only cases registered in the first half of the century were highly concentrated in time and space: all occurred between 1922 and 1923 and in central Europe. A common story can be told about those episodes: political instability, large fiscal imbalances due, in part, to war and huge increases in the money supply.

It is also interesting to note that the first half of the century was still characterized mainly by convertible monetary systems. The four hyperinflationary experiences described by Sargent occurred during temporary suspensions of the gold standard. By the mid-1970s, however, after the fall of the Bretton Woods arrangement, the world moved to a fiat money system, in which no commodity serves as backing.

The second half of the century also witnessed hyperinflationary episodes. Somewhat surprisingly, the second wave of hyperinflationary episodes was concentrated in the period 1985–94. And they were concentrated in two regions; it would appear, though, that the temporal coincidence was just random. The countries involved were Argentina, Brazil, Bolivia and Peru in Latin America, and Yugoslavia and Poland in central Europe. Again, a common story could be told: in the first years of the 1980s, the four Latin American countries experienced major financial crises, including default in international debt markets. As a consequence, the ability of the governments to smooth temporary fiscal shocks via credit markets was severely restricted. The four countries had experienced in the previous decade substantial political instability, including military dictatorships and weak democratic governments. On the other hand, both Poland and Yugoslavia were undergoing substantial political and economic transformation after the fall of the USSR. In all cases, there were major fiscal imbalances: government deficits were chronic and volatile. As consequence, money printing became the only source of revenues and major bursts in inflation rates occurred.

It is interesting to note that other Latin American countries (Colombia, Uruguay, Mexico) also suffered financial and debt crises, but did not experience inflation rates of this magnitude, and other central European countries underwent major political and economic transformation and did not have hyperinflations.

Indeed, what we have learned (see Bruno et al. 1988, 1991) is that major political and economic crises are a necessary condition for hyperinflations to occur. But crisis will lead to hyperinflation if, and only if, the crisis manifests itself in serious fiscal imbalances that are financed by the central bank issuing unbacked paper money. There is a wide consensus in the literature about this.

Although we know very precisely the conditions under which hyperinflations are almost unavoidable, it is difficult to tell exactly when the burst will start and how large it will be.

The subtlety of hyperinflationary dynamics has been explored in a sequence of papers (Eckstein and Leiderman 1992; Zarazaga 1993; Marcet and Nicolini 2003) that can be seen as complementary. All these models share the property, supported by evidence, that hyperinflations can occur only in economies with large and persistent fiscal deficits that are purely financed by printing money, or seigniorage. In all the models, the problem arises because the required seigniorage is close to the maximum revenue that can be raised, given the demand for real money, that is, the maximum of the Laffer curve. Eckstein and Leiderman (1992) argue that if the elasticity of money demand with respect to the inflation rate approaches one form above, when average seigniorage is very high, very small shocks to it can generate drastic changes in the required inflation rate.

Zarazaga (1993) introduces a decentralized government with a common pool of resources and private information on the shock to the spending opportunities of each member of the government. Hyperinflations occur when there are too many positive expenditure shocks, there is too much demand for resources, and the required seigniorage is too high. When this happens, a price war-type strategy follows in which all agencies become excessively demanding and the central bank ends up issuing enormous amounts of currency. Finally, Marcet and Nicolini (2003) introduce very small departures from rationality and show that the dynamics of the most simple seigniorage model change in a way that fits the evidence surprisingly well.

From the point of view of inflation stabilization policies, the debate has taken three routes. The first claims that the key for a successful stabilization policy is to correct the fundamentals, this is, to make a drastic and permanent change in fiscal policy so as to eliminate the need to print money. This kind of policy is called ‘orthodox’. The second puts the emphasis on ‘heterodox’ policies, that is, a combination of nominal anchors like fixing the nominal exchange rate – eventually moving towards a gold or strong currency standard – and price and wage controls. Finally, a third approach points to the need to combine the other two policies. From the point of view of experience and the theory, it is clear that no attempt to stabilize the economy without orthodox policies has any chance of success in the medium term. And it appears from experience that in most successful cases (although there has been some debate on whether this was true in all of them), some type of nominal anchor, typically the exchange rate, was also important. While not all theoretical models put much weight on the nominal anchor (Marcet and Nicolini 2003, is the most notable exception), in all of the models these policies are either harmless or good for the success of the stabilization effort.

A final word regarding Cagan’s (1956) definition: as with any definition, it is arbitrary. Had we taken a lower inflation rate per month, like 25 per cent, the number of experiences would have been greater, and many more countries would have been involved in our discussion. However, the general lessons one learns are essentially the same. Quantity theory predictions work extremely well, and the most appropriate policies to deal with these experiences are the same.

See Also

▶ German Hyperinflation
▶ Inflation

Bibliography

Bruno, M., G. Di Tella, R. Dornbusch, and S. Fisher. 1988. *Inflation stabilization: The experience of Argentina, Brazil, Israel and Mexico*. Cambridge, MA: MIT Press.

Bruno, M., S. Fisher, E. Helpman, and N. Liviatan. 1991. *Lessons of economic stabililzation and its aftermath*. Cambridge, MA: MIT Press.

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

Eckstein, Z., and L. Leiderman. 1992. Seigniorage and the welfare cost of inflation. *Journal of Monetary Economics* 29: 389–410.

Fisher, I. 1934. *Stable money: A history of the movement*. New York: Adelphi.

Marcet, A., and J. Nicolini. 2003. Recurrent hyperinflations and learning. *American Economic Review* 93: 1476–1498.

Marcet, A., and J. Nicolini. 2005. Money and prices in models of bounded rationality in high-inflation economies. *Review of Economic Dynamics* 8: 452–479.

Sargent, T. 1992. The ends of four big inflations. In *Rational expectations and inflation*, 2nd ed. New York: Harper and Row.

Zarazaga, C. 1993. *Hyperinflation and moral hazard in the appropriation of seigniorage*. Working paper, Federal Reserve Bank of Philadelphia.

Hypothesis Testing

Gregory C. Chow

Testing Restrictions on Parameters

For those who believe that economic hypotheses have to be confirmed by empirical observations, hypothesis testing is an important subject in economics. As a classical example, when an economic relation is represented by a linear regression model:

$$Y = X\beta + \varepsilon \tag{1}$$

where Y is a column vector of n observations on the dependent variable y, X is an $n \times k$ matrix with each column giving the corresponding n observations on each of k explanatory variables (which typically include a column of ones), β is a column of k regression coefficients and ε is a vector of n independent and identically distributed residuals with mean zero and variance σ^2, it is of interest to test a hypothesis consisting of m linear restrictions on β:

$$R\beta = r \tag{2}$$

where R is $m \times k$ and r is $m \times 1$. A most common case occurs when there is only one restriction ($m = 1$) and (2) is reduced to $\beta_i = 0$, the hypothesis being that the ith explanatory variable has no effect on y.

Among the statistical tests often employed in economic research are the likelihood ratio (LR) test, the Wald test and the Lagrangian multiplier (LM) test. The LR test, due to Neyman and Pearson (1928), uses as the test statistic the likelihood ratio:

$$\mu = \frac{L\left(Y, \widehat{\theta}^*\right)}{L\left(Y, \widehat{\theta}\right)} \tag{3}$$

where L is the likelihood function, $\widehat{\theta}^*$ is the maximum-likelihood estimator of a parameter vector θ under the null hypothesis to be tested, or subject to a vector $h(\theta) = 0$ of m restrictions such as (2), and $\widehat{\theta}$ is the ML estimator of θ without imposing the restrictions. A high value of the likelihood ratio μ favours the null hypothesis. The Wald test, proposed by Wald (1943), uses the test statistic:

$$W = h\left(\widehat{\theta}\right)'\left[\mathrm{Cov}\ h\left(\widehat{\theta}\right)\right]^{-1} h\left(\widehat{\theta}\right) \tag{4}$$

where Cov denotes covariance matrix. The null hypothesis $h(\theta) = 0$ will be accepted if the vector $h\left(\widehat{\theta}\right)$ is sufficiently close to zero, or if the statistic W is sufficiently small. Wald (1943) has shown that under general conditions, the statistics W and -2 ln μ have the same asymptotic distribution.

The LM test, suggested by Silvey (1959), uses the Lagrangian multiplier $\widehat{\lambda}$: obtained by maximizing the Lagrangian expression:

$$n^{-1}\ln L(Y, \theta) + \lambda' h(\theta) \tag{5}$$

or by solving the associated first-order conditions for $\widehat{\theta}^*$ and $\widehat{\lambda}$:

$$\begin{aligned} n^{-1}\frac{\partial \ln L\left(Y,\widehat{\theta}^*\right)}{\partial\theta} + H_\theta\widehat{\lambda} &= 0 \\ h\left(\widehat{\theta}^*\right) &= 0 \end{aligned} \qquad (6)$$

where H_θ denotes the $k \times m$ matrix $\partial h'(\theta) / \partial\theta$. The solution of (5) gives the maximum-likelihood estimator $\widehat{\theta}$ subject to the restriction $h(\theta) = 0$ and the associated Lagrangian multiplier $\widehat{\lambda}$. Under the null hypothesis $h(\theta) = 0$, $\sqrt{n}\widehat{\lambda}$ has a normal limiting distribution with mean zero and a certain covariance matrix $- R$. Hence the statistic $-\widehat{\lambda}' R^{-1}\widehat{\lambda}[$ is distributed asymptotically as $\chi^2(m)$. This statistic can be rewritten as a score statistic (see Chow 1983, pp. 286–9):

$$\begin{aligned} -n\widehat{\lambda}'\widehat{R}^{-1}\widehat{\lambda} &= \left[\partial \ln L\left(Y,\widehat{\theta}^*\right)/\partial\theta'\right] \\ &\times \left[-\partial^2 \ln L\left(Y,\widehat{\theta}^*\right)/\partial\theta\partial\theta'\right]^{-1}\left[\partial \ln L\left(Y,\widehat{\theta}^*\right)/\partial\theta\right] \end{aligned} \qquad (7)$$

As is well known, under the null hypothesis ∂ ln $L(Y, \theta) / \partial\theta$ has mean zero and covariance matrix $- E\partial^2$ ln $L / \partial\theta\partial\theta'$. If the vector ∂ ln $L\left(Y,\widehat{\theta}^*\right)/\partial\theta'$ is very different from zero, as measured by the statistic (6), one would be inclined to reject the null hypothesis. Silvey (1959) has shown that under fairly general assumptions:

$$\begin{aligned} -p\lim\ (2\log\mu) &= p\lim W \\ &= -p\lim\ n\widehat{\lambda}' R^{-1}\widehat{\lambda} \end{aligned} \qquad (8)$$

and that the LR test, the Wald test and the LM test are asymptotically equivalent in the sense that their test statistics have the same asymptotic distribution. The equivalence for testing the hypothesis (2) in the linear regression case with normal residuals is shown in Chow (1983, pp. 290–291).

An example of (2) often encountered in practice is the hypothesis that certain subsets of coefficients in two linear regressions are equal. The test serves to detect whether certain economic parameters have changed from one sample period to another or whether they are different in two different situations (see Chow 1960). Let the two samples of n_1 and n_2 observations be represented by:

$$\begin{aligned} Y_i &= X_i\beta_i + \varepsilon_i \\ &= Z_i\gamma_i + W_i\delta_i + \varepsilon_i, \quad (i = 1, 2) \end{aligned} \qquad (9)$$

We wish to test H_0: $\gamma_1 = \gamma_2$, each with k_1 elements. A linear regression model for both samples can be written as:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} Z_1 & 0 & W_1 & 0 \\ 0 & Z_2 & 0 & W_2 \end{bmatrix}\begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \delta_1 \\ \delta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix} \qquad (10)$$

The null hypothesis $\gamma_1 = \gamma_2$ can be written as a set of k_1 linear restrictions:

$$R\beta = \begin{bmatrix} I & -I & 0 & 0 \end{bmatrix}\begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \delta_1 \\ \delta_2 \end{bmatrix} = 0 \qquad (11)$$

When the elements of ε_1 and ε_2 are normal, the test statistic is:

$$\frac{(A - B/k_1}{B/(n_1 + n_2 - 2k)} \qquad (12)$$

where A is the sum of squared residuals of (9) estimated by imposing the k_1 restrictions (10) and B is the sum of squared residuals estimated without imposing the restrictions. Under H_0, the statistic (11) has an $F(k_1,\ n_1 + n_2 - 2\ k)$ distribution.

Much useful information concerning economic relations can be ascertained by testing hypotheses about the parameters of economic models. For example, one question in applying the regression model (1) to time-series data is whether the elements ε_t are serially correlated. One may postulate a first-order autoregressive model $\varepsilon_t = \rho\ \varepsilon_{t-1} + \eta_t$ for the residuals, where ε_t is assumed to be independent and identically distributed. The hypothesis of interest is ρ = 0. As another example, one may ask whether the relation between y and a certain explanatory variable x_j is linear. A partial answer is given by introducing powers of x_j in the regression and testing whether their coefficients are significantly different from zero.

Testing Non-nested Hypotheses

In the last section, the hypothesis to be tested consists of a set of restrictions $g(\theta) = 0$ on the parameter vector θ. Since the null hypothesis states that the parameter θ lies in a subspace of a parameter space, it is nested within a more general hypothesis. Comparing a more general alternative hypothesis with a more restrictive null hypothesis nested within the former is to test a nested hypothesis. When the two hypotheses to be compared are not nested, we are testing *non-nested* hypotheses. One important example of non-nested hypothesis consists of two regression models, (1) and:

$$Y = Z_{\gamma} + u \tag{13}$$

where Z is an $n \times p$ matrix including a different set of explanatory variables from those included in X of model (1). X and Z may have some variables in common, but neither hypothesis can be derived from restricting the values of the parameter vector permitted by the other hypothesis. In general, one may wish to choose between two non-nested hypotheses represented by two density functions $f_1(y, \theta_1)$ and $f_2(y, \theta_2)$ for generating y.

For the purpose of choosing between two competing density functions, Cox (1961, 1962) suggests combining them in the model:

$$h(y;\theta_1,\theta_2,\lambda) = kf_1(y,\theta_1)^{\lambda} f_2(y,\theta_2)^{1-\lambda} \tag{14}$$

If the maximum-likelihood estimate of λ is close to 1, choose f_1; if it is close to zero, choose f_2; if neither, the result is inconclusive. Quandt (1974) proposes an alternative way of combining the two density functions, namely:

$$\begin{aligned} h(y;\theta_1,\theta_2,\lambda) = {} & \lambda f_1(y_1,\theta_1) \\ & + (1-\lambda) f_2(y,\theta_2) \end{aligned} \tag{15}$$

For choosing between two normal linear regression models (1) and (12), all parameters in (14) are identifiable, whereas for (13) one cannot separately identify $\lambda, \beta, \gamma, \sigma_1^2 = E\varepsilon_i^2$ and $\sigma_2^2 = Eu_i^2$.

A common approach to choosing between non-nested models is to formulate a more general model nesting them and reduce the problem to one of testing a nested hypothesis, as exemplified by the methods just described. As another example, to choose between linear regression models (1) and (12), one may formulate a more general linear regression model including both sets of explanatory variables X and Z. If this general model is assumed to be the true model, then both (1) and (12) may be false. Nevertheless, one may still ask which has a smaller error in predicting y by testing the null hypothesis that the residual variances of these models are equal. The residual variance of the regression of Y on X is:

$$n^{-1}\left[E(Y'Y) - (EY)'X(X'X)^{-1}X'(EY)\right]$$

and similarly for the regression of Y on Z. In the general model, let $EY = [X\ Z]\alpha$. The equality of these two residual variances means:

$$\begin{aligned} & \alpha'[X \quad Z]'\left[X(X'X)^{-1}X' - Z(Z'Z)^{-1}Z'\right] \\ & \qquad [X \quad Z]\alpha \equiv \alpha' H \alpha = 0 \end{aligned} \tag{16}$$

This is a quadratic restriction on the coefficient vector α in a linear regression model. It can be tested by the methods of (3), (3) and (4). See Chow (1980, 1983, pp. 278–284). Some other works on testing non-nested hypotheses are cited in Chow (1983, pp. 284–286).

Testing Model Specifications

When an economist wishes to find out whether a certain model is correctly specified, tests of model specification can be used. The situation here differs from that of section “Testing Non-nested

Hypotheses" in having no specific model to compete with the model in question. It differs from that of section "Testing Restrictions on Parameters" in not singling out, at least in the first instance, certain parameters as the likely sources of model misspecifications. If one believes that an omitted variable in a regression model may be the culprit, one would test whether its coefficient is significantly different from zero. If one believes that the residuals may be serially correlated, one might add an autoregressive structure to the residual and test the significance of its coefficients. Likewise, one may drop certain explanatory variables by testing the significance of their coefficients. In tests of model specifications, the alternatives are less specific. The tests aim at detecting misspecifications of a model against a variety of alternatives.

One approach to specification testing, initiated by Wu (1973) and studied by Hausman (1978), is based on comparing two estimators of a parameter vector which are both consistent and asymptotically normal if the model is correctly specified. One estimator $\widehat{\gamma}^0$ is asymptotically efficient if the model is correctly specified but is inconsistent if the model is incorrectly specified. The second estimator $\widehat{\gamma}$ is consistent even if the model is incorrectly specified. If the difference $\widehat{q} = \widehat{\gamma} - \widehat{\gamma}^{0}$ is large, one tends to reject the null hypothesis that the model is correctly specified. Let $V(\widehat{q})$ be the covariance matrix of the asympototic distribution of $\sqrt{n(\widehat{q})}$ and $\widehat{V}(\widehat{q})$ be a consistent estimate of $V(\widehat{q})$. Then under the null hypothesis, which implies $p \lim q = 0$:

$$n\widehat{q}'\widehat{V}(\widehat{q})^{-1}\widehat{q} \tag{17}$$

will have $\chi^2(k)$ as its asymptotic distribution, k being the number of elements of $\widehat{q}$. As an example, consider testing whether X is correlated with ε in model (1). Under the null hypothesis $p \lim n^{-1} X' \varepsilon = 0$, an asymptotically efficient estimator is the least-squares estimator $\widehat{\beta}^0$. Even if the null hypothesis does not hold a consistent estimator is the instrumental variable estimator $\widehat{\beta} = (W'X)^{-1} Y$ where we assume $p \lim n^{-1} W'X$ to be a nonsingular matrix and $p \lim n^{-1/2} W' \varepsilon$ to converge in distribution to k-variate normal with zero mean. A $\chi^2(k)$ statistic can be constructed to test the null hypothesis, using the difference $\widehat{q} = \widehat{\beta} - \widehat{\beta}^0$ and its covariance matrix. Another example is to test the correct specification of simultaneous equations by comparing a three-stage least-squares estimator $\widehat{\gamma}^0$ and a two-stage least-squares estimator $\widehat{\gamma}$.

A convenient framework of Newey (1985) views specification testing as choosing some function $m(y, \theta)$ which satisfies the moment condition:

$$E[m(y, \theta_0)] = 0 \tag{18}$$

if the model $f(y, \theta)$ is correctly specified, and testing this condition by using the sample moment $\sum_{t=1}^{n} m\left(y_1, \widehat{\theta}\right)/n$. For example, the information matrix text of White (1982) compares two estimates of the information matrix and uses as elements of the vector function $m(y, \theta)$:

$$m_h(y,\theta) = \frac{\partial \ln f(y,\theta)}{\partial \theta_i} \cdot \frac{\partial \ln f(y,\theta)}{\partial \theta_j} + \frac{\partial^2 \ln f(y,\theta)}{\partial \theta_i \partial \theta_j}$$
$$(h = i + j - 1;\ \ i = 1, \ldots, j;\ \ j = 1, \ldots, k) \tag{19}$$

where k is the number of parameters. The Hausman test using (16) is shown by Newey (1985) to be asymptotically equivalent to a particular moment-condition test.

Economists using various specification tests should be reminded that these tests serve the same purpose as the many diagnostic checks for statistical models used in the literature. Examples are the diagnostic checks of Box and Jenkins (1970) for time-series models and those of Belsley, Kuh and Welsch (1980) for regression models.

Model Selection Criteria

The statistical tests presented so far are based on the notion that if a model is true (an assumption to be tested), it will be chosen. This nation might be questioned because the true model can be very complicated and in practice one may prefer to use a simpler model for estimation or prediction purposes. Consider the choice between model (1), with $X\beta = X_1\beta_1 + X_2\beta_2$ and normal, and the smaller linear model using X_1 alone as explanatory variables, where X_1 is $n \times k_1$ and X_2 is $n \times k_2$. The standard treatment using the methods of section "Testing Restrictions on Parameters" is to test the null hypothesis $\beta_2 = 0$, but a question remains as to what level of significance to use. An alternative viewpoint is to choose the model which is estimated to have smaller prediction errors. Specifically, let n future, out-of-sample, observations be:

$$\tilde{Y} = \tilde{X}\beta + \tilde{\varepsilon} \tag{20}$$

under the assumption that the larger model (1) is the true model. Let the model be selected which has a smaller expected sum of squared prediction errors.

Using the small model with X_1 alone and denoting the corresponding maximum-likelihood estimate of β by $\widehat{\beta}_1$ [consisting of $(X_1{}'X_1)^{-1}\ X_1{}'Y$ and 0], one easily evaluates $E\left(\widehat{\beta}_1 - \beta\right)\left(\widehat{\beta}_1 - \beta\right)'$. Then using the estimated small model and the predictor $E\left(\widehat{\beta}_1 - \beta\right)\left(\widehat{\beta}_1 - \beta\right)'$ for $\tilde{y}$, one finds the expected sum of squared prediction errors to be:

$$\begin{aligned} &E\left(\tilde{X}\tilde{\beta}_1 - \tilde{Y}\right)'\left(\tilde{X}\tilde{\beta}_1 - \tilde{Y}\right) \\ &\quad = E\left(\tilde{\beta}_1 - \beta\right)'\tilde{X}'\tilde{X}\left(\tilde{\beta}_1 - \beta\right) + E\tilde{\varepsilon}'\tilde{\varepsilon} \\ &\quad = k_1\sigma^2 + \beta_2'X_2'\left[I - X_1\left(X_1'X_1\right)^{-1}X_1'\right]X_2\beta_2 + n\sigma^2 \end{aligned} \tag{21}$$

Using the large model (1) and letting $\widehat{\beta} = (X'X)^{-1}$ X', we have:

$$\begin{aligned} &E\left(X\widehat{\beta} - \tilde{Y}\right)'\left(\tilde{X}\beta - \tilde{Y}\right) \\ &\quad = (k_1 + k_2)\sigma^2 + n\sigma^2 \end{aligned} \tag{22}$$

Comparing (20) and (21), we find that the small model, though not being the true model, should be used if and only if:

$$\beta_2'X_2'\left[I - X_1\left(X_1'X_1\right)^{-1}X_1'\right]X_2\beta_2 \equiv \beta_2'X_{2.1}'X_{2\cdot 1}\beta_2 < k_2\sigma^2 \tag{23}$$

where $X_{2.1}$ is the matrix of residuals of the regression of X_2 on X_1. To apply the criterion (22), one may replace $\beta_2{}'X_{2.1}{}'X_{2.1}\beta_2$ by its unbiased estimate β_2' $X_{2.1.}'X_{2.1}\widehat{\beta}_2 - k_2\sigma^2$, and replace σ^2 in the resulting inequality by the unbiased estimate s^2 to yield:

$$\begin{aligned} &\widehat{\beta}_2'X_{2.1}'X_{2\cdot 1}\widehat{\beta}_2 < 2k_2s^2 \\ &\qquad \equiv 2k_2\left(Y - X\widehat{\beta}\right)'\left(Y - X\widehat{\beta}\right)/(n - k_1 - k_2) \end{aligned} \tag{24}$$

as the condition for selecting the small model. This criterion amounts to setting the critical value of the F ratio $\widehat{\beta}_2'X_{2.1}'X_{2\cdot 1}\widehat{\beta}_2/k_2s^2$ for testing the null hypothesis $\beta_2 = 0$ equal to 2. It is the C_p criterion of Mallows (1973) and is motivated by the desire for more accurate prediction. Comparing (20) and (21) we observe that omitting the variables X_2 might yield a better model for prediction even when (1) is the true model and $\beta_2 \neq 0$.

The information criterion of Akaike (1973, 1974) is also motivated by the desire for more accurate prediction. However, instead of using the expected squared prediction errors, one uses the following expected information:

$$E\left[\ln g\left(\tilde{Y},\ \theta_0\right) - \ln f(\tilde{y},\ \theta)\right] \tag{25}$$

to measure how good the density function $f(\cdot)$ of the model used for predicting a future observation y is, as compared with the true model $g(\cdot)$. Akaike has implemented this criterion by estimating (24), suggesting the criterion for selecting a model if its maximum log likelihood minus the number of estimated parameters is the highest among the competing models. A model having more parameters will tend to have a higher value for its maximum log likelihood, but this value has to be reduced by the number of parameters estimated. Sawa (1978) has provided a better estimate of (24) for linear regression models while Chow (1981a, b) has provided better estimates of (24) for general statistical models and simultaneous-equation models.

The Posterior-Probability Criterion

Another criterion for selecting models is the Jeffrey–Bayes posterior-probability criterion. Let $p(M_j)$ be the prior probability for model M_j to be correct and $p(\theta|M_j)$ be the prior density for the k_j-dimensional parameter vector $\widehat{\theta}_j$ conditioned on M_j being correct. Assume that a random sample of n observations $(y_1, y_2, \ldots, y_n) = Y$ is available. By Bayes's theorem the posterior probability of the jth model being correct is:

$$\begin{aligned} p(M_j|Y) &= \frac{p(M_j)p(Y|M_j)}{p(Y)} \\ &= \frac{p(M_j)p(Y|M_j)}{\sum\limits_j p(M_j)p(Y|M_j)} \end{aligned} \tag{26}$$

where

$$p(Y|M_j) = \int L_j(Y,\theta)p(\theta|M_j)\mathrm{d}\theta \tag{27}$$

with $L_j(Y, \theta_j)$ denoting the likelihood function for the jth model. Since $p(Y)$ is a common factor for all models, the model with the highest posterior probability of being correct is the one with the maximum value for:

$$p(M_j)p(Y|M_j) = p(M_j)\int L_j(Y,\theta)p(\theta|M_j)\mathrm{d}\theta$$

If the prior probabilities $p(M_j)$ are equal for the models, the one with the highest $p(Y|M_j)$ will be selected.

To evaluate $p(Y|M_j)$ for large samples we apply a theorem of Jeffreys (1961, pp. 193ff.) on the posterior density $p(\theta|Y, M_j)$ of θ_j given model M_j:

$$p(\theta|Y,M_j) = \frac{L_j(Y,\theta)p(\theta|M_j)}{p(Y|M_j)} = (2\pi)^{-k_j/2}|S|^{1/2}\exp\left[-\tfrac{1}{2}\left(\theta-\hat{\theta}_j\right)'S\left(\theta-\hat{\theta}_j\right)\right]\times\left[1+0\left(n^{-1/2}\right)\right] \tag{28}$$

where is the maximum-likelihood estimate of θ_j and the inverse covariance matrix is $S = -\left[\partial^2 \mathrm{ln} L_j)/(\partial\theta\partial\theta')\right]_{\widehat{\theta}} \equiv \ 3\mathrm{pt} nR_j \cdot 0\left(n^{-1/2}\right)$ is a function of order $n^{-1/2}$. Thus, for large samples, the posterior density of a parameter vector θ in model j is asymptotically normal with mean equal to the maximum-likelihood estimate $\widehat{\theta}_j$ and covariance matrix which can be approximated by the inverse of S. Evaluating both sides of (27) at and taking natural logarithms, we obtain, nothing $|S| = |nR_j| = n^{kj}\,|R_j|$,

$$\begin{aligned} \ln p(Y|M_j) &= \ln L_j\left(Y,\widehat{\theta}_j\right) - \frac{k_j}{2}\ln\ n - \tfrac{1}{2}\log|R_j| \\ &\quad + \frac{k_j}{2}\ln\ 2\pi + \ln p\left(\widehat{\theta}_j|M_j\right) + 0\left(n^{-1/2}\right) \end{aligned} \tag{29}$$

If we retain only the first two terms $p(Y|M_j)$ and $-k_j(\tfrac{1}{2}\ln n)$ in (28), we obtain the formula of Schwarz (1978) for approximating $\log p\ (Y|M_j)$.

In practice $\ln p(Y|M_j)$ may not be well approximated by using only the first two terms of (28), as it will depend on the prior density $p(\theta|M_j)$ of the parameter vector chosen for each model M_j. Bayesian statisticians, including Jeffreys (1961), Pratt (1975), and Leamer (1978), among others, have recognized the difficult problem of choosing a prior distribution $p(\theta|M_j)$ for the parameters of each model to be used to computer $p(Y|M_j)$. Unlike the estimation of parameters by Bayesian methods, even for large samples the choice of models by the posterior-probability criterion is very sensitive to the prior distribution $p(\theta|M_j)$ assumed for each model.

In this essay I have summarized some of the important ideas and methods employed in hypothesis testing and model selection in econometrics. The choice of an econometric model is a complicated subject. Many approaches have to be explored in practice for choosing and evaluating econometric models. Some of these approaches are discussed in Chow and Corsi (1982) and in Belsley and Kuh (1986).

See Also

- ▶ Econometrics
- ▶ Information theory
- ▶ Likelihood
- ▶ Non-nested hypotheses
- ▶ Regression and correlation analysis
- ▶ Statistical inference

Bibliography

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. In *Proceedings of the 2nd international symposium for information theory*, ed. B. Petrov and F. Cśaki. Budapest: Akademiai Kiadó.

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* AC-19: 716–723.

Belsley, D., and E. Kuh. 1986. *Model reliability*. Cambridge, MA: MIT Press.

Belsley, D., E. Kuh, and R. Welsch. 1980. *Regression diagnostics*. New York: Wiley.

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

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

Chow, G. 1980. The selection of variates for use in prediction: a generalization of Hotelling's solution. In *Quantitative econometrics and development*, ed. L. Klein, M. Nerlove, and S.C. Tsiang. New York: Academic Press.

Chow, G. 1981a. A comparison of the information and posterior probability criteria for model selection. *Journal of Econometrics* 16: 21–33.

Chow, G. 1981b. Evaluation of econometric models by decomposition and aggregation. In *Methodology of macro-economic models*, ed. J. Kmenta and J. Ramsey. Amsterdam: North-Holland.

Chow, G. 1983. *Econometrics*. New York: McGraw-Hill.

Chow, G., and P. Corsi (eds.). 1982. *Evaluating the reliability of macro-economic models*. London: Wiley.

Cox, D. 1961. Tests of separate families of hypotheses. In *Proceedings of the 4th Berkeley symposium on mathematical statistics and probability*. Berkeley: University of California Press.

Cox, D. 1962. Further results on tests of separate families of hypotheses. *Journal of the Royal Statistical Society Series* B24: 406–424.

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

Jeffreys, H. 1961. *Theory of probability*, 3rd ed. Oxford: Clarendon Press.

Leamer, E. 1978. *Specification searches*. New York: Wiley.

Mallows, C. 1973. Some comments on C_p. *Technometrics* 15: 661–675.

Newey, W. 1985. Maximum likelihood specification testing and conditional moment tests. *Econometrica* 53: 1047–1070.

Neyman, J., and E. Pearson. 1928. On the use of interpretation of certain test criteria for the purpose of statistical inference. *Biometrika* 20A, Part I: 175–240; Part II: 263–294.

Pratt, J. 1975. Comments. In *Studies in Bayesian econometrics and statistics*, ed. S. Fienberg and A. Zellner. Amsterdam: North-Holland.

Quandt, R. 1974. A comparison of methods for testing nonnested hypotheses. *Review of Economics and Statistics* 56: 92–99.

Sawa, T. 1978. Information critiera for discriminating among alternative regression models. *Econometrica* 46: 1273–1292.

Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics* 6: 461–464.

Silvey, S. 1959. The Lagrangian multiplier test. *Annals of Mathematical Statistics* 30: 389–407.

Wald, A. 1943. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society* 54: 426–482.

White, H. 1982. Maximum likelihood estimation of misspecified models. *Econometrica* 50: 1–25.

Wu, D. 1973. Alternative tests of independence between stochastic regressors and disturbances. *Econometrica* 41: 733–750.