D

Dalton, Edward Hugh John Neale (1887–1962)

Alan Peacock

Keywords

Bank of England; Bernoulli's formula; Bonds; Cheap money; Dalton, E. H. J.; Public finance; Redistribution of income and wealth; Taxation

JEL Classifications

B31

British fiscal economist and prominent Labour politician, Hugh Dalton was a student of A.C. Pigou and J.M. Keynes. His main professional interest was in the use of taxation as an instrument for the redistribution of income and wealth, an interest inspired by Pigou's teaching and by his revulsion at the contrast between the sufferings inflicted on younger generations by the First World War and the material gains of those who financed or profited from the war itself. (Dalton spent four years on military service in France and Italy and lost several close friends, including the poet Rupert Brooke.) His main contribution was to investigate the properties of a modification of Bernoulli's formula $dw = dw/x$ where $w =$ economic welfare and $x =$ income but in which equal increases in welfare should correspond to more than proportionate increases in income, a condition satisfied by Dalton's formula $dw = dx/x^2$ so that $w = c - 1/x$ where c is a constant. Using this formula he concluded that economic welfare would be improved by transfers from rich to poor (Dalton 1935), a proposition that has excited the interest of 'modern' public finance theorists of the neo-utilitarian school (see Fishburn and Willig 1984). He elaborated his ideas in several works including his highly successful standard text *Principles of Public Finance* and in his lectures as Reader in Economics at the London School of Economics (1923–36). There he was responsible for teaching and for recommending Lionel Robbins to be Professor of Economics, a typical example of his desire not only to 'corrupt the young' (as he termed it) but also to promote the interests even of those with whom he disagreed.

Dalton combined teaching with a political career throughout the 1920s and 1930s, rising to political eminence as a member of Churchill's coalition government during the Second World War. As Minister of Economic Warfare he was responsible for setting up the famous sabotage team, the Special Operations Executive (SOE). Later as President of the Board of Trade he formulated plans for post-war distribution of industry designed to prevent mass unemployment. In the Attlee Labour government of 1945 he reached the pinnacle of his political career as Chancellor of the Exchequer, one of his first acts being to nationalize the Bank of England. His famous attempt to drive down interest rates

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

through a cheap money policy in order to float off an issue of Treasury stock at 2.5 per cent is a classic example of the failure of even an experienced and able economist to understand that, other than in the short run, governments can control either the price or the supply of bonds but not both.

Selected Works

1923. *Principles of public finance.* London: George Routledge & Sons.
1935. *The inequality of incomes.* 4th Impression, London: George Routledge & Sons, especially the Appendix.

Bibliography

Davenport, N. 1961–70. Hugh Dalton. In *Dictionary of national biography.* Oxford: Oxford University Press.
Fishburn, P.C., and R.D. Willig. 1984. Transfer principles in income redistribution. *Journal of Public Economics* 25: 323–328.
Pimlott, B. 1985. *Hugh Dalton*. London: Jonathan Cape.

Dantzig, George B. (1914–2005)

Richard W. Cottle, B. Curtis Eaves and Mukund N. Thapa

Abstract

George Dantzig is known as 'father of linear programming' and 'inventor of the simplex method'. This biographical sketch traces the high points of George Dantzig's professional life and scholarly achievements. The discussion covers his graduate student years, his wartime service at the US Air Force's Statistical Control Division, his post-war creativity while serving as a mathematical advisor at the US Air Force Comptroller's Office and as a research mathematician at the RAND Corporation, his distinguished career in academia – at UC Berkeley and later at Stanford University – and finally as an emeritus professor of operations research.

Keywords

Complementarity; Computational complexity; Convex programming; Convexity; Dantzig, G. B; Decomposition principle; Degeneracy; Distributed computation; Hurwicz, L; Integer programming; Interior point methods; Kantorovich, L.V; Koopmans, T.C; Lagrangian function; Leontief, W. W; Linear programming; Logarithmic barrier method; Mathematical programming; Neyman, J; Nonlinear programming; Operations research; Simplex method for solving linear programs; Stochastic programming with recourse; von Neumann, J; Wood, M.K

JEL Classification

B31

George Dantzig is known as the 'father of linear programming' and the 'inventor of the simplex method'. Employed at the Pentagon (the US government's defence establishment) in 1947 and motivated to 'mechanize' programming in large timestaged planning problems, George Dantzig gave a general statement of what is now known as a linear program, and invented an algorithm, the simplex method, for solving such optimization problems. By the force of Dantzig's theory, algorithms, practice, and professional interaction, linear programming flourished. Linear programming has had an impact on economics, engineering, statistics, finance, transportation, manufacturing, management, and mathematics and computer science, among other fields. The list of industrial activities whose practice is affected by linear programming is very long.

Over the subsequent half century, Dantzig remained a major contributor to the subject of linear programming as researcher, practitioner, teacher, mentor, and leader. The impact of linear programming and extensions on theory, business, medicine, government, the military, all in the broadest sense, is now hard to overstate. In the words of the editors of the Society for Industrial

and Applied Mathematics: 'In terms of widespread application, Dantzig's algorithm is one of the most successful of all time: linear programming dominates the world of industry, where economic survival depends on the ability to optimize within budgetary and other constraints...' (quoted in Dongarra and Sullivan 2000).

There were some significant contributions to what became linear programming prior to Dantzig's work. In their time, however, these results were not applied, linked together, or continued. In fact, they were nearly lost, perhaps because the prevailing historical setting was not favourable. As these contributions have been recognized, they have been drawn into the history of linear programming.

A Linear Program Defined

In mathematical terms, a linear program is most simply stated as the problem of minimizing a multivariate linear function constrained by linear inequalities. Dantzig's first formulation of a linear program was the equivalent problem of minimizing a linear function over non-negative variables constrained by linear (material balance) equations. In matrix notation such a linear program is:

$$LP(A,b,c): \quad \begin{array}{ll} \text{Minimize} & C^T x = z \\ x,z & \\ \text{subject to} & Ax = b, \quad x \geq 0. \end{array}$$

Here the given data are the $m \times n$ matrix A and vectors b and c; the unknowns to be determined are the objective scalar value z and the decision-variable vector x. The simplex method solves a linear program in a comprehensive sense; in particular, no conditions are imposed on the data (A, b, c). Dantzig assessed a linear program as the simplest optimization model with broad applicability.

The study, solution, and application of linear programs constitute the subject of linear programming. The use of the words 'programming' and 'program' has changed somewhat over time. The original idea was that 'programming' is the activity of deciding now upon a plan, called a program, for some system that would be executed later in time. The same meaning was subsequently adopted in computer programming where the system is a computer. (See linear programming.)

Early Life and Education

George Bernard Dantzig was born to Tobias Dantzig and his wife, Anja Ourisson, in Portland Oregon, on 8 November 1914. Tobias, a house-painter and pedlar in his early years in the United States, later held professional positions at John Hopkins University (1919–1920) and the University of Maryland (1927–1946) where he chaired the mathematics department from 1930 to 1941. He is best known for his book Number, *The Language of Science*, which is still in print (T. Dantzig 1930).

In 1936, George Dantzig both received an A.B.. in Mathematics and Physics at the University of Maryland and married Anne Shmuner (1917–2006), who at age 19 received an A.B.. in French at Maryland. In 1938 Dantzig received an MA in Mathematics at the University of Michigan; he was a Horace Rackham Scholar. In 1937–1939 Dantzig worked at the U.S. Bureau of Labor Statistics as a junior statistician. Inspired by a paper of J. Neyman on which he had been assigned to report, Dantzig wrote to Neyman, then at University College London, asking if he could study under his supervision. Neyman relocated to the University of California at Berkeley, and Dantzig became his student in 1939. As is now folklore, one day Dantzig arrived late for one of Neyman's theoretical statistics classes and proceeded to copy two problems from the blackboard. In a few weeks time, with some effort, Dantzig solved the problems and submitted his homework, whereupon it was tossed onto a large pile of papers on Neyman's desk. Early one Sunday morning, about two weeks later, George and Anne were awakened by a pounding on their apartment door. There was Neyman waving George's homework. As it turned out, the assumed homework problems were, in fact, important unsolved problems. Furthermore, Neyman continued, these solutions, suitably

presented, would suffice for George's Ph.D. dissertation. A. Wald independently obtained one of the same results, and the work was eventually published jointly in Dantzig and Wald (1951). Before Dantzig could complete his degree, Pearl Harbor was attacked, and he took leave of absence to work at the U.S. Air Force Comptroller's Office.

Dantzig in Washington, DC, 1941–1952

At the outbreak of the Second World War, Dantzig began working at the War Department, again as a junior statistician. By the war's end, he was in charge of the Combat Analysis Branch of the Statistical Control Division of the United States Air Force. His office collected and consolidated data with hand-operated mechanical desk calculators about sorties flown, tons of bombs dropped, planes lost, personnel attrition rates, and so on. By end of the war, Dantzig had a personnel force of 300 reporting to him.

In 1946 Dantzig returned to Berkeley for one semester to defend his thesis and complete his minor thesis in dimension theory. Throughout his life, Dantzig acknowledged a great debt to J. Neyman, his mentor. Dantzig nonetheless turned down a position in mathematics at UC Berkeley for the greater financial security of a position at the Pentagon. There he undertook the challenge to 'mechanize' the planning process. War planning required coordination of an entire nation and yet was executed with desk calculators; the need for mechanization was clear. To this end, a group in the Air Force was organized under the name Project SCOOP (Scientific Computation of Optimum Programs) and headed by M.K. Wood. Dantzig was a principal. Two movements suggested that progress was possible: Leontief's (1936) work and the emergence of the computer; indeed, Project SCOOP arranged for Pentagon support of computer development (see Dantzig 1947).

In early 1947, Dantzig formulated the general statement of a linear program. In June of that year he learned from T.C. Koopmans, who had been studying transportation problems (Koopmans 1947) that economists had no algorithm for solving a linear program. By July Dantzig had designed the simplex method, a name suggested by Leo Hurwicz (see simplex method for solving linear programs). Experiments with the simplex method in the following year at the Pentagon were encouraging. Linear programs were also solved with the simplex method at the National Bureau of Standards (NBS) in coordination with SCOOP. At NBS a 'large one', the diet problem, was undertaken by J. Laderman. It had been studied earlier by Stigler (1945). The question was: what selection of 77 foods produces a diet meeting nine nutritional criteria at the least cost? The problem was solved by the simplex method with five statistical clerks using desk calculators. According to (Dantzig 1963), 'approximately 120 man-days were required to obtain a solution'. The simplex method was gaining acceptance. Air Force applications of linear programming in years following included contract bidding, crew training, deployment scheduling, maintenance cycles, personnel assignments, and airlift logistics.

From special cases as a triangular model to the general algorithm, the simplex method was first implemented on a computer in 1949 by M. Mantalbano (NBS) on an IBM 602-A, in 1950 by C. Diehm on the SEAC, in 1951 by A. Orden (Air Force) and A. Hoffman (NBS) on the SEAC, and in 1952 by the Air Force for the Univac. The next generation of codes, circa 1952–1956, which achieved commercial quality, was developed by W. Orchard-Hays at the RAND Corporation on a sequence of IBM machines. For the matrix A of $LP(A, b, c)$ of size 200 by 1000, linear programs could be solved in five hours (Orchard-Hays 1954). In years following, there was a flood of computer implementations, both by commercial vendors and in research institutions. As of 2006, linear programs where both m and n exceed hundreds of thousands are routinely solved in hours by the simplex method on personal computers.

After describing and testing the simplex method, Dantzig had an audience with J. von Neumann at Princeton in 1947. Among world-class mathematicians, von Neumann had the broadest interests. Dantzig began his explanation

of linear programming with the 30-min version when von Neumann snapped 'Get to the point'. Dantzig began again, this time with his one-minute version. Von Neumann responded, 'Oh, that!' He envisioned an analogy with matrix games as developed in von Neumann and Morgenstern (1944). Extrapolating from what he knew about duality in matrix games, von Neumann expounded on what was to become known as duality in linear programming. As a by-product of the meeting, it was evident that any matrix game problem could be solved by a linear program. Volume VI of John von Neumann: Collected Works contains his previously uncirculated manuscript dated 15–16 November 1947 on duality in linear programming (von Neumann 1947). The following January, Dantzig (1948a) wrote 'A Theorem on Linear Inequalities'. This memorandum clarified his understanding of von Neumann's duality monologue. Von Neumann's (1947) paper is regarded as the earliest on this subject; Dantzig's memorandum is the second. A.W. Tucker, also at Princeton, took an interest in the relationship of linear programming and game theory and involved his students, D. Gale and H.W. Kuhn. These three subsequently wrote the definitive account of duality in linear programming (Gale, Kuhn and Tucker 1951).

First Linear Programming Conference, 1949

Koopmans organized a conference on 'linear programming' and economics in Chicago at the Cowles Commission for Research in Economics in 1949. Koopmans and others (including Dantzig) edited the conference proceedings volume *Activity Analysis of Production and Allocation* (1951). Dantzig's work was the focus of the proceedings; of the 25 papers, Dantzig co-authored a paper with M.K. Wood and authored four others, including the two leading papers which developed linear programming for time-staged planning. Earlier versions of these two papers appeared in *Econometrica* (1949). Four of the 20 contributors to these proceedings – K.J. Arrow, T.C. Koopmans, P.A. Samuelson, and H.A. Simon – were later to win Nobel Prizes. Hundreds of books on, or inspired by, linear programming followed over the years. Four of note are Dorfman, Samuelson and Solow (1958), Arrow, Hurwicz and Uzawa (1958), Dantzig (1963), and Schrijver (1986). The terminology 'linear programming' was not in regular use at the time of this conference; Koopmans had suggested it to Dantzig (1948b) in lieu of expressions like 'programming in a linear structure'. Even so, Koopmans (1951) observed, 'To many economists the term linearity is associated with narrowness, restrictiveness, and inflexibility of hypotheses'. R. Dorfman, at the Pentagon with Dantzig, had suggested the broader expression of 'mathematical programming'.

Nonlinear Programming, 1950

Following the early successes of linear programming, there was a natural inclination to generalize the model, the algorithm, and duality to results beyond linear functions to a next layer of difficulty such as differentiable, convex, quadratic, or polynomial functions. This body of research has become known as 'nonlinear programming'. As for optimality conditions and duality, the paper 'Nonlinear Programming' of Kuhn and Tucker (1951) was pivotal at the time: their investigation proceeded through the Lagrangian function and saddle points thereof with the duality in linear programming as a target. The Lagrangian had been used in equality-constrained optimization, and results obtained there were less general. Kuhn and Tucker cited the fundamental paper of John (1948), which includes inequality constraints. Some 25 years later, the master's thesis of Karush (1939) came to light in the mathematical programming community; Karush, as far as is known, was the first to lay down optimality conditions for a nonlinear (inequality constrained) program. Rockafellar (1970) carried the convex duality analysis to a new level. As for nonlinear programming algorithms, tens, and eventually hundreds, were forthcoming, many using ideas from the simplex method in one way or another.

Dantzig at RAND, 1952–1960

Reorganization of the Air Force preceded Dantzig's taking a position at the RAND

Corporation in Santa Monica, California, as a research mathematician. Awareness of the power of linear programming set the scene for a second growth. For the next few years most theoretical development of linear programming took place at RAND and Princeton. Dantzig's book *Linear Programming and Extensions* (1963) records his own (and collaborative) contributions during this period.

Transportation and Network Optimization Problems

The war years has seen interest in optimal transportation research. Historically significant papers from this period include Hitchcock (1941), Kantorovich (1942), Koopmans (1947, 1949), Kantorovich and Gavurin (1949), and Flood (1956). Flood, through M. Shiffman, had come upon the Kantorovich papers on translocation and transportation; however, linear programming launched the general analysis of optimal transportation. Dantzig made several contributions here, starting with Dantzig (1951). Dantzig, Fulkerson and Johnson (1954) is a seminal work on the travelling salesman problem. Others are Dantzig and Fulkerson (1954) on tanker routing, and Dantzig and Fulkerson (1955) on maximizing flow through a network. For networks with non-negative arc distances, Dantzig (1960a) stated an algorithm for shortest distances. Dijkstra (1959) produced similar results at about the same time. Flows in Networks by the RAND Corporation's Ford and Fulkerson (1962) was then the definitive work on the subject.

Large-Scale Methods and Decomposition

Dantzig and Orchard-Hays (1954) described the 'revised simplex method' as a more efficient version of the simplex method. As linear programming was applied to more applications and with a broader scope, including time and alternate scenarios, the size of linear programs that needed to be solved continued to grow. Dantzig was among the first to observe that large linear programs typically had two convenient features: sparsity and structure. Sparsity refers to the fact that a very small percentage, often less than one hundredth of one per cent, of the A data matrix is non-zero. Structure refers to the fact that the non-zeros typically occur an orderly pattern of submatrices of A. Dantzig (1955a) wrote the first paper on methods for large-scale linear programs addressing upper bounds, block triangular systems, and secondary constraints. Building on the Dantzig, Orden and Wolfe (1955) paper on generalized linear program, Dantzig and Wolfe (1960) devised a generalization of the simplex method, called the *decomposition principle*, for certain structured large-scale linear programs, wherein the problem is decomposed allowing for use for what is now called distributed computation.

Quadratic Programming

A most natural first extension of a linear program is a quadratic program, that is, a linear program except that the objective is a quadratic function such as $x^T Qx + q^T x$. A convex quadratic program is one with a convex objective function to be minimized. Following the success of linear programming, there was a proliferation of studies on convex quadratic programming and associated algorithms.

Convex Programming

Convex programming is also a natural extension of linear programming. Here a convex function is minimized over a convex region; the latter is specified by convex inequality constraints. If the feasible region is bounded, the convex program can be approximated as close as desired by a linear program, and one can improve the approximation as the simplex method runs. A special case of a convex program is one having linear inequality constraints and a *separable* objective function, that is, a function that is the sum of univariate convex functions. Charnes and Lemke (1954) and Dantzig (1956) solved such problems with linear programming approximations.

Stochastic Programming with Recourse

Linear programming offered a breakthrough for mathematical approximation and solution of planning problems. Dantzig knew that to move to the next level of approximation of planning, an accommodation of uncertainty and of discrete

variables was needed; he made inroads on each. Linear programming has been extended in a number of directions to incorporate uncertainty. An elementary example is a linear program where the costs $c = (c_1, c_2, \ldots, c_n)$ are random variables and the desire is to minimize the expected value. In this case the problem is solved as the linear program where the costs are simply taken as their expected value. More interesting is the Markowitz (1956) portfolio selection where quadratic programming is used to obtain at least a desired level of expected return while minimizing risk.

Dantzig's early work on stochastic programming was stimulated by his work with A.R. Ferguson on the assignment of aircraft to routes, where a deterministic formulation proved insufficient, and so uncertain demand needed to be considered (Ferguson and Dantzig 1955). Subsequently, Dantzig (1955b) applied linear programming to solve multistage decision problems sequenced amidst uncertainty; this topic is often referred to as stochastic programming with recourse. Such a multistage problem concerns the optimization of a sequence of decisions in time where each decision depends on random events which in turn are dependent on previous decisions. The vision in this paper was truly extraordinary, and has been reprinted as one of the ten most influential papers in management science since the mid-1950s in Hopp (2004).

Integer Programming and Cuts

An integer program is a linear program except that some, or all, of the variables $x_1, x_2, \ldots, x_n$ are required to take on integer values, as in $x_i = 0, 1, 2, \ldots$. Dantzig, Fulkerson and Johnson (1954) took the first steps towards obtaining integer solutions for a large problem with the simplex method. They addressed an instance of the travelling salesman problem: find the shortest route, by car, through major cities of the 48 states and Washington, DC. Let a directed network represent the available roads and let costs represent distances. The variables are flows on each link of the network. Constrain for one unit of flow into each capital, constrain for one unit of flow out of each capital, constrain for conservation of flow at other nodes, and find the minimum cost flow. The linear programming solutions here, which yield flows of 0 or 1, are deficient as a solution for the travelling salesman problem in that isolated loops of flow may occur. To combat such loops, Dantzig et al. sequentially and dynamically (as the simplex method was stopped and continued) introduced additional constraints, called cuts, which would prohibit those loops which had occurred in a solution of the expanding linear program, without constraining out desired solutions. The concept of a cut or cutting planes was so conceived. In addition, this study revealed the inherent difficulty of the travelling salesman problem. Over the following decades, aspects of this matter would grow to become a major issue in applied mathematics. There is a vast difference between linear constraints and linear inequality constraints (both with unconstrained variables); there is an even larger difference between real variables and integer variables. Subsequently, Gomory (1958), at Princeton, began the design of several general purpose cutting plane algorithms for solving integer programs, and gave proofs for finite convergence. These algorithms did not work well for a reason not understood at the time namely, that general integer programs are inherently hard to solve.

Other Edge Path Descent Algorithms

By 1955 the simplex method was regarded as the algorithm for solving linear programs. Indeed, the simplex method inspired dozens of related fundamental ideas for algorithms, and hundreds of variations. In particular, there was steady research on variations of edge path descent algorithms, that is, those which accept the simplex method strategy but strive to improve upon it. One target was to reduce computation time by reducing the number of pivots and the work per pivot. Example contributions include: the dual simplex method of Lemke (1954), the parametric method of Orchard-Hays (1954), the primal-dual method of Dantzig, Ford and Fulkerson (1956) and the parametric objective method of Gass and Saaty (1955). In a slightly different direction were the column generation and the decomposition method of Dantzig and Wolfe (1960, 1961). Essentially all of these variants of the simplex method have proved valuable for various specialized tasks

related to linear programs, and sometimes nonlinear programs. For nonlinear programs the main ideas of the simplex method have been adopted; here one can think of solving linear or quadratic programs that are approaching the nonlinear program. It is interesting to note that as late as the early 1970s an eminent speaker of a plenary session of a national mathematical programming conference said that the simplex method was the best algorithm for linear programming and that it always would be; the statement was accepted, without objection.

Problem Reduction

The mathematical subject of computational complexity aims to categorize problems by their solution difficulty. Several of Dantzig's papers (1957, 1960b, 1968) contributed to the foundation of this subject. A basic technique of computational complexity is the reduction of one class of problems to another. For the reduction of discrete problems, Dantzig focused on problems in mixed binary form, MBP, and the related relaxed form *RMBP* obtained by replacing binary constraints with corresponding interval constraints. *MBP*(*A, b, c*) is a linear program *LP*(*A, b, c*) plus the discrete constraints $x_i = 0, 1 \quad \text{for} \quad i = 1, \ldots, k \quad \text{for some} \quad k \leq n$. *RMBP*(*A, b, c*) is the linear program *LP*(*A, b, c*) plus the linear inequalities $0 \leq x_i \leq 1$ for $i = 1, \ldots k$. For emphasis, *MBP*(*A, b, c*) is not a linear program whereas *RMBP*(*A, b, c*) is.

A few problem classes of form *MBP* can be solved as the corresponding linear program *RMBP*; that is if (*x, z*) is an extreme point solution, as the simplex method would generate, of *RMBP*, then (*x, z*) is a solution to *MBP*. Problem classes *MBP* which can be so solved by *RMBP* include the assignment problem, shortest route problems with non-negative distances, and the tanker scheduling problem. Other problems, such as the empty container problem, most scheduling problems, fixed charge problems, and travelling salesman problems, do not permit such solution; nevertheless, the corresponding *RMBP* can be most helpful in solving or approximately solving *MBP*. As time and theory have revealed, general problems of type *MBP* are difficult to solve.

Let *C** be the convex hull of all feasible solutions of *MBP* and let *C* be the set of all solutions of RMBP. Then *C** is a subset of *C*, and all extreme points of *C** are extreme points of *C*; the issue is, however, that there are extreme points of *C* that are not in *C**. Note that, if there is but one binary variable, then *MBP* can be solved as two linear programs, one with $x_1 = 0$ and one with $x_1 = 1$; but for general *k*, this scheme requires the solution of an exponential number 2^k of linear programs. For reducing problems to the *MBP* form, Dantzig (1960b) illustrated a number of examples such as: (*a*) dichotomies, (*b*) discrete variables, (*c*) piecewise linear objective functions, (*d*) conditional constraints, and (*e*) the fixed charge problem.

Recognition of Earlier Work, 1958–1960

Towards the end of the 1950s, the mathematical programming community became aware of three relevant works from the past. The first two are pertinent to the simplex method and the third relevant to the formulation of real problems as linear programs. Fourier (1826) had also written on the idea of descending from vertex to adjacent vertex in the polyhedron defined by linear inequalities for minimizing a linear error over linear inequalities. De la Vallée Poussin (1911), independently of Fourier's work, made a similar suggestion and gave two examples. There appears to have been no follow-up on their suggestions. Also, neither Fourier nor de la Vallée Poussin described his ideas fully enough to reveal any awareness of degeneracy considerations and corresponding non-convergence possibilities, much less any procedures for coping with the matter. Made aware of Kantorovich's transportation papers by Flood (1956), Koopmans (1960) corresponded with Kantorovich. In due course, an English translation of Kantorovich's remarkable 1939 paper was made available to the West as 'Mathematical Methods of Organizing and Planning Production' (Kantorovich 1960). Therein Kantorovich had formulated a collection of problems as what we now call linear programs. These problems were: machine utilization, production planning, scrap management, refinery scheduling, fuel utilization, construction planning, and arable land distribution. Using the Minkowski separation theorem, Kantorovich

proved in this work that optimal multipliers exist. He suggested some ideas based on 'resolving multipliers' (essentially dual variables, or marginal costs) towards an algorithm, but none has emerged following this line of thought. According to Dantzig (1963), 'Kantorovich should be credited with being the first to recognize that certain important broad classes of production problems had well-defined mathematical structures which, he believed, were amenable to practical numerical evaluation and could be numerically solved'. But although Koopmans (1960) argued that, with a suitable transformation, one of Kantorovich's problems had the generality of Dantzig's linear program, Koopmans's conclusion was not justified as the argument did not and could not cover the possibilities of infeasibility and an unbounded objective, a point made by Charnes and Cooper (1962). Koopmans's argument notwithstanding, the statement of a general linear program belongs to Dantzig.

Dantzig Returns to UC Berkeley, 1960–1966

Dantzig left RAND to become a professor in the industrial engineering department at the University of California at Berkeley. There, that year, he established the Operations Research Center. Operations research (OR) was a term that emerged in the Second World War to describe the activity of studying an operation (process, system, and so on) with mathematical methods with the intent of improving performance. In 1963 Dantzig completed his classic *Linear Programming and Extensions*. The book was based on his research which began at the Pentagon and continued through RAND and UC Berkeley. By the time Dantzig left UC Berkeley in 1966, he had produced 11 Ph.D. students, and written about 25 research papers on the theory and practice of linear programming and extensions (integer, nonlinear, stochastic, and so on). As a mentor of Ph.D. students, Dantzig was among the very best. Within a course or two he could bring students to the frontier on some aspect of linear programming. His new book offered a full perspective of linear programming right up to 1963. Dantzig supplied the time, inspiration, guidance, knowledge, and example that students needed. He lived and breathed research.

Interest in the study of linear and nonlinear complementarity problems, as such, began in the early 1960s. Dantzig's second student, Cottle (1964), wrote on this topic, and his work was extended in Cottle and Dantzig (1968). Problems in this category can be viewed as abstractions of optimality conditions or of (economic or physical) equilibrium conditions. In a complementarity problem, one has a mapping W of R^N into itself and seeks a solution z of the conditions $W(z) \geq 0$, $z \geq 0$, $z^T W(z) = 0$. In the linear complementarity problem, the mapping would be of the form $W(z) = Mz + q$. The linear complementarity problem is related to the minimization of $z^T(Mz + q)$ subject to the constraints $Mz + q \geq 0$ and $z \geq 0$. This would be easy enough to solve as a quadratic program, if the objective function were convex. However, the excitement arose from the fact that the problem could be solved, effectively, in the absence of convexity. From the classic paper of Lemke (1965) followed the computation of points in the core of a balanced game and the computation of economic equilibria (Scarf 1967, 1973), the computation of fixed points with piecewise linear homotopies (Eaves 1972), and the computation with differentiable functions (Smale 1976).

Dantzig at Stanford University, 1966–1996

Dantzig joined the Stanford faculty in 1966, half-time in the inter-departmental Operations Research Program and half-time in Computer Science. In 1967 the OR Program became the Department of Operations Research in the School of Engineering; this is where Dantzig conducted his work. He was away for two years: in 1973–1974 at the International Institute for Applied Systems Analysis in Austria, and in 1978–1979 at the Center for Advanced Study in the Behavioral Sciences on the Stanford campus. In 1973 he was appointed to the C.A. Criley Professorship in Transportation Science. While at Stanford, Dantzig produced 41 Ph.D. students

and published about 115 research papers on the theory and applications of mathematical programming. Dantzig's Ph.D. progeny, if Berkeley and Stanford graduates and subsequent generations are counted, as of 2006 exceeded 200. Dantzig had long felt that the development of good software was key to widespread usage of linear programming in industry. This vision led him to create the Systems Optimization Laboratory (SOL) at Stanford for research and development of numerical algorithms for mathematical programming. Under the SOL banner were the PILOT and planning under uncertainty programs (see Dantzig et al. 1973; Gill et al. 2007).

Stochastic Programming with Recourse, Continued, 1989–2005

Cognizant of the potentially enormous size of multi-stage stochastic linear programs, Dantzig and Madansky (1961) suggested the incorporation of statistical sampling of uncertainties together with approximating time-staged models to solve the full problem. Following this avenue some 30 years later, Dantzig and Glynn (1990) brought together decomposition, Monte Carlo sampling, and multiprocessing to solve time-staged linear programs (see also Infanger 1991; Dantzig and Infanger 1992). In a series of papers, importance sampling was used to estimate second-stage costs and Benders cuts. Portfolio optimization and electric power planning were among the applications envisioned; the latter problems, with 39 uncertain parameters leading to 15 million scenarios, were solved to high accuracy with a confidence level of 95%; in equivalent deterministic form, such problems would have more than four billion constraints. However, Dantzig, to the end, regarded stochastic linear programming as a major unresolved problem.

Computational Complexity, 1972–2006

Since its inception, the question of the number of steps required by the simplex method for a given linear program has been of interest. In the 1970s the field of 'computational complexity' emerged; a theory of problem difficulty which draws a sharp distinction between categories of problems that could be solved in polynomial time (number of steps) in the size of their data, and those which could not. How did linear programming fit into this scheme? Klee and Minty (1972) produced a worst case example of a simple linear program on which the simplex method takes an exponential number of iterations. But the expected number of pivots of the simplex method over a random selection of problems was shown to be polynomial in (m, n) (Smale 1983). This raised the question: could a linear program be solved in polynomial time? Khachiyan (1979) defined a polynomial time algorithm for linear programs based on a sequence of convergent ellipsoids; however, unexpectedly according to computational complexity, the algorithm was very slow, and certainly no competitor of the simplex method. Later, Karmarkar (1984) gave a polynomial time interior point algorithm for linear programs which was claimed to be superior to the simplex method in the sense of solving linear programs much faster on a computer; the method required the linear program to be expressed in a special form with an optimal objective value of zero and viewed each iterate as being at the centre of a polyhedron in a different coordinate system. The method typically required considerably fewer iterations than the simplex method, but each iteration required significantly more computations. The method was patented by AT&T and published as a theoretical result. There was considerable secrecy associated with the particulars of its implementation; and, thus, no independent verification was possible regarding its claimed superiority in computational speed over the simplex method. It was later shown to be equivalent under the same special form to the logarithmic barrier method, a method traceable back to Frisch (1955) and Fiacco and McCormick (1968). The logarithmic barrier method, however, could be applied to a linear program in standard form. The logarithmic barrier method was in the public domain and so allowed researchers to focus on computational improvements. Today, it is known that there are problems for which the logarithmic barrier method is superior to the simplex method; notable are those very large problems for which AA^T is sparse. For a survey of interior point methods, see Todd (1996). It is also interesting to note that most practical interior-point algorithms

include an option to move the ε-optimal interior point solution to the nearest extreme point, a procedure requiring a significant number of simplex-type pivots. A technique to do this was proposed by Dantzig (1963, ch. 6, exercise 11). As of 2006, the simplex method (and various realizations thereof) remains the algorithm of choice for the majority of linear programs.

Dantzig in Retirement, 1996–2005

Dantzig was retired from Stanford in stages, each firmly resisted. He was formally retired from the regular faculty at age 65 in 1980, but was recalled until age 82 in 1996. Until that year he remained as active as formal members of the faculty. After that he met at home with all who wished to consult him: students, colleagues, and strangers. Whenever presented with an idea, Dantzig would respond, as always, with something of value. Until around 2001 he continued to travel and present papers. At his 90th birthday celebration, he attended a full day of presentations followed by a banquet and additional talks. He was full of energy, enthusiasm, keen observations, and wit. Dantzig's mind was razor-sharp up to the end.

In retirement, Dantzig's principal project was the writing of a multi-volume book on linear programming and extensions. Dantzig had always felt that software was a key element that would contribute to the success of linear programming usage. He wanted to write another book on linear programming that incorporated software to aid students in learning both the theory and the practice of linear programming, and in particular in learning how to implement the simplex method and other algorithms for commercial use. In 1985 he invited M.N. Thapa to coauthor such a book. As work on the book progressed, it became apparent to the authors that the amount of material required a really huge book. One volume became two, and two became four. In the end only two volumes were completed (Dantzig and Thapa 1997, 2003). Dantzig continued to be fascinated by interior point methods; von Neumann's and Karmarkar's algorithms were reanalysed and included in the second volume. According to M. Thapa, Dantzig never tired of editing and re-editing to improve proofs and readability. He would say: 'it is like polishing a stone; the more you polish it, the more it will shine.' Dantzig also continued his work with G. Infanger on planning under uncertainty. In addition to their research together, Dantzig and Infanger consulted on financial portfolio design. They intended to edit a collection of papers (including work of their own) on planning under uncertainty. Dantzig was convinced that the way to get further exposure for, and research into, planning under uncertainty was to set up an institute; to no avail, he tried at Stanford, tried at EPRI, and finally tried to create a stand-alone non-profit organization. In addition to these projects, Dantzig reworked the text of a science fiction novel he had begun in 1980.

Dantzig's Honours

In 1975 L.V. Kantorovich and T.C. Koopmans received the Nobel Prize in Economics for 'their contributions to the theory of optimum allocation of resources'. Both mentioned Dantzig in their Nobel Lectures. That Dantzig did not participate in this prize came as a great shock and disappointment to those familiar with his contributions. Himself aside, Dantzig regarded Leontief, Kantorovich, von Neumann, and Koopmans as the principal early contributors to linear programming.

Dantzig, the man, and his contributions have nevertheless been honoured extensively. His honours include distinguished memberships, prizes, honorary doctorates, and dedications. He was elected to membership in the National Academy of Sciences, the National Academy of Arts and Sciences, and the National Academy of Engineering. He was a fellow of the Econometric Society, the Institute of Mathematical Statistics, the Association for the Advancement of Science, the Operations Research Society, IEEE, and the Omega Rho Society. He was awarded the War Department Exceptional Civilian Service Medal, the National Medal of Science, the John von Neumann Theory Prize, the NAS Award in Applied Mathematics and Numerical Analysis, the Harvey Prize (Technion), the Silver Medal of Operational

Research Society (England), the Adolph Coors American Ingenuity Award, the Special Recognition Award of Mathematical Programming Society (MPS), the Harold Pender Award, and the Harold Lardner Memorial Prize (Canada). He received honorary doctorates from the Israel Institute of Technology (Technion), University of Linkøping (Sweden), University of Maryland, Yale University, Université Catholique de Louvain (Belgium), Columbia University, the University of Zurich, and Carnegie-Mellon University. Dantzig was also honoured as the dedicatee of a symposium of MPS, in two volumes of *Mathematical Programming*, in the first issue of the *Journal of Optimization of the Society for Industrial and Applied Mathematics* (SIAM), with the joint MPS-SIAM Dantzig Prize, and with the INFORMS Dantzig Prize for students. In 2006, a fellowship in his name was established in the Department of Management Science and Engineering at Stanford University.

See Also

- ▶ Kantorovich, Leonid Vitalievich (1912–1986)
- ▶ Koopmans, Tjalling Charles (1910–1985)
- ▶ Leontief, Wassily (1906–1999)
- ▶ Linear Programming
- ▶ Simplex Method for Solving Linear Programs
- ▶ von Neumann, John (1903–1957)

Selected Works

1947. *Prospectus for the AAF electronic computer*. Unpublished manuscript.

1948a. *A theorem on linear inequalities*. Unpublished manuscript, 5 January.

1948b. Programming in a linear structure. Washington, DC: Comptroller, USAF. February. *Abstract in Econometrica* 17(1949), 73–74.

1951. Application of the simplex method to a transportation problem. In *Activity analysis of production and allocation*, ed. T.C. Koopmans. New York: Wiley.

1951. (With A. Wald.) On the fundamental lemma of Neyman and Pearson. *Annals of Mathematical Statistics* 22, 87–93.

1954. (With D.R. Fulkerson.) Minimizing the number of tankers to meet a fixed schedule. *Naval Research Logistics Quarterly* 1, 217–222.

1954. (With D.R. Fulkerson and S.M. Johnson.) Solution of a large scale traveling-salesman problem. *Journal of Operations Research Society of America* 2, 393–410.

1954. (With W. Orchard-Hays.) The product form for the inverse in the simplex method. *Mathematical Tables and Other Aids to Computation* 8, 64–67.

1955a. Upper bounds, secondary constraints, and block triangularity in linear programming. *Econometrica* 23, 174–183.

1955b. Linear programming under uncertainty. *Management Science* 1, 197–206.

1955. (With A.R. Fergusson.) The problem of routing aircraft. *Aeronautical Engineering Review* 14(4), 51–55. RAND Research Memorandum RM1369, 1954.

1955. (With D.R. Fulkerson.) Computation of maximal flows in networks. *Naval Research Logistics Quarterly* 2, 277–283.

1955. (With A. Orden, A. and P. Wolfe.) The generalized simplex method for minimizing a linear form under linear inequality restraints. *Pacific Journal of Mathematics* 5(2), 183–195. RAND Research Memorandum RM-1264, 1954.

1956. Recent advances in linear programming. *Management Science* 2, 131–144.

1956. (With L.R. Ford, Jr. and D.R. Fulkerson.) A primal-dual algorithm for linear programs. In *Linear inequalities and related systems*, Annals of Mathematics Study No. 38, ed. H.W. Kuhn and A.W. Tucker. Princeton, NJ: Princeton University Press.

1957. Discrete variable extremum problems. *Operations Research* 5, 226–277.

1960a. On the shortest route through a network. *Management Science* 6, 187–190. RAND memorandum P-1345, 1959.

1960b. On the significance of solving linear programming problems with some integer variables. *Econometrica* 28, 30–44.

1960. (With P. Wolfe.) Decomposition principle for linear programs. *Operations Research* 8, 101–111.

1961. (With A. Madansky.) On the solution of two-stage linear programs under uncertainty. In

Proceedings, Fourth Berkeley Symposium on Mathematical Statistics and Probability, I, ed. J. Neyman. Berkeley University of California Press. RAND memorandum P-2039, 1960.
1961. (With P. Wolfe.) The decomposition algorithm for linear programming. *Econometrica* 29, 767–778.
1963. *Linear Programming and Extensions*. Princeton, NJ: Princeton University Press.
1968. Large-scale linear planning. In *Mathematics of the Decision Sciences*, vol. 1, ed. G.B. Dantzig and A.F. Veinott, Jr. Providence, RI: American Mathematical Society.
1968. (With R.W. Cottle.) Complementary pivot theory of mathematical programming. *Linear Algebra and its Applications* 1, 103–125.
1973. (With others.) On the need for a System Optimization Laboratory. In *Mathematical Programming*, ed. T.C. Hu and S.M. Robinson. New York: Academic Press.
1982. Reminiscences about the origins of linear programming. *Operations Research Letters* 1(2), 43–48.
1990. (With P.W. Glynn.) Parallel processors for planning under uncertainty. *Annals of Operations Research* 22, 1–21.
1991. Linear programming. In *History of Mathematical Programming: A Collection of Personal Reminiscences*, ed. J.K. Lenstra, A.H.G. Rinnooy Kan and A. Schrijver. Amsterdam: North-Holland.
1992. (With G. Infanger.) Large-scale stochastic linear programs: importance sampling and Benders decomposition. In *Computational and Applied Mathematics I – Algorithms and Theory, Proceedings of the 13th IMACS World Congress*, ed. C. Brezinski and U. Kulisch. Amsterdam: North-Holland.
1997. (With M.N. Thapa.) *Linear Programming, 1. Introduction*. New York: Springer.
2003. (With M.N. Thapa.) *Linear Programming 2: Theory and Extensions*. New York: Springer.

Acknowledgments The authors are grateful to David Dantzig, Jessica Dantzig Klass, and many of Dantzig's friends and colleagues who have contributed to this biographical article. These include A.J. Hoffman, G. Infanger, E. Klotz, J.C. Stone, M.J. Todd, J.A. Tomlin and M.H. Wright. This article has also benefited from other writings on G.B. Dantzig's life, namely: Albers and Reid (1986), Albers, Alexanderson and Reid (1990), Cottle (2003, 2005, 2006), Cottle and Wright (2006), Dantzig (1982, 1991), Dorfman (1984), Gill et al. (2007), Kersey (1989), Lustig (2001), Gass (1989, 2002, 2005).

Bibliography

Albers, D.J., and C. Reid. 1986. An interview with George B. Dantzig: The father of linear programming. *College Mathematics Journal* 17: 292–314.
Albers, D.J., G.L. Alexanderson, and C. Reid. 1990. *More mathematical people*. New York: Harcourt Brace Jovanovich.
Arrow, K.J., L. Hurwicz, and H. Uzawa. 1958. *Studies in linear and non-linear programming*. Stanford: Stanford University Press.
Benders, J.K. 1962. Partitioning procedures for solving mixed-variables programming problems. *Numerische Mathematik* 4: 238–252.
Charnes, A., and W.W. Cooper. 1962. On some works of Kantorovich, Koopmans and others. *Management Science* 8: 246–263.
Charnes, A., and C.E. Lemke. 1954. Minimization of non-linear separable convex functionals. *Naval Research Logistics Quarterly* 1: 301–312.
Cottle, R.W. 1964. *Nonlinear programs with positively bounded Jacobians*. Ph.D. thesis. Department of Mathematics. University of California at Berkeley.
Cottle, R.W. 2003. *The basic George B. Dantzig*. Stanford: Stanford University Press.
Cottle, R.W. 2005. George B. Dantzig: Operations research icon. *Operational Research* 53: 892–898.
Cottle, R.W. 2006. George B. Dantzig: A life in mathematical programming. *Mathematical Programming* 105: 1–8.
Cottle, R.W., and M.H. Wright. 2006. Remembering George Dantzig. *SIAM News* 39(3): 2–3.
Dantzig, T. 1930. *Number, The language of science*. New York: Macmillan. 4th edition, ed. J. Mazur, republished New York: Pi Press, 2005.
de la Vallée Poussin, M.C.J. 1911. Sur la méthode de l'approximation minimum. *Annales de la Société Scientifique de Bruxelles* 35: 1–16.
Dijkstra, E. 1959. A note on two problems in connection with graphs. *Numerische Mathematik* 1: 269–271.
Dongarra, J., and F. Sullivan. 2000. The top 10 algorithms. *Computing in Science and Engineering* 2(1): 22–23.
Dorfman, R. 1984. The discovery of linear programming. *Annals of the History of Computing* 6: 283–295.
Dorfman, R., P.A. Samuelson, and R.M. Solow. 1958. *Linear programming and economic analysis*. New York: McGraw-Hill.
Eaves, B.C. 1972. Homotopies for the computation of fixed points. *Mathematical Programming* 3: 1–22.
Fiacco, A.V., and G.P. McCormick. 1968. *Nonlinear programming: Sequential unconstrained minimization techniques*. New York: Wiley.

Flood, M.M. 1956. The traveling-salesman problem. *Operational Research* 4: 61–75.
Ford Jr., L.R., and D.R. Fulkerson. 1962. *Flows in networks*. Princeton: Princeton University Press.
Fourier, J.B.J. 1826. Solution d'une question particulière du calcul des inégalités. *Nouveau Bulletin des Sciences par la Société Philomatique de Paris*, 99–100. Reprinted in *Oeuvres de Fourier, Tome II*, ed. G. Darboux. Paris: Gauthier, 1890.
Frisch, R.A.K. 1955. *The logarithmic potential method of convex programs*. Oslo: University Institute of Economics.
Gale, D., H.W. Kuhn, and A.W. Tucker. 1951. Linear programming and the theory of games. In *Activity analysis of production and allocation*, ed. T.C. Koopmans. New York: Wiley.
Gass, S.I. 1989. Comments on the history of linear programming. *IEEE Annals of the History of Computing* 11(2): 147–151.
Gass, S.I. 2002. The first linear-programming shoppe. *Operational Research* 50: 61–68.
Gass, S.I. 2005. In Memoriam, George B. Dantzig. 2005. Online. Available at http://www.lionhrtpub.com/orms/orms-8-05/dantzig.html. Accessed 12 Jan 2007.
Gass, S.I., and T.L. Saaty. 1955. The computational algorithm for the parametric objective function. *Naval Research Logistics Quarterly* 2: 39–45.
Gill, P.E., Murray, W., Saunders, M.A., Tomlin, J.A. and Wright, M.H. 2007. *George B. Dantzig and systems optimization*. Online. Available at http://www.stanford.edu/group/SOL/GBDandSOL.pdf. Accessed 2 Feb 2007.
Gomory, R.E. 1958. Essentials of an algorithm for integer solutions to linear programs. *Bulletin of the American Mathematical Society* 64: 256.
Hitchcock, F.L. 1941. The distribution of a product from several sources to numerous localities. *Journal of Mathematics and Physics* 20: 224–230.
Hoffman, A. 1953. *Cycling in the Simplex Algorithm*. Report No. 2974. Washington, DC: National Bureau of Standards.
Hopp, W.J., ed. 2004. Ten most influential papers of *Management Science's* first fifty years. *Management Science* 50(12 Supplement), 1764–1769.
Infanger, G. 1991. Monte Carlo (importance) sampling within a Benders decomposition algorithm for stochastic linear programs. *Ann. Oper. Res.* 39: 41–67.
John, F. 1948. Extremum problems with inequalities as side conditions. In *Studies and essays, Courant anniversary volume*, ed. K.O. Friedrichs, O.E. Neugebauer, and J.J. Stoker. New York: Wiley-Interscience.
Kantorovich, L.V. 1942. Translocation of masses. *Doklady Akademii Nauk SSSR* 37: 199–201. Reprinted in Management Science 5 (1958–59), 1–4.
Kantorovich, L.V. 1960. Mathematical methods of organizing and planning production. *Management Science* 6: 363–422. English translation of original monograph published in 1939.
Kantorovich, L.V. and Gavurin, M.K. 1949. The application of mathematical methods to freight flow analysis. *Problems of Raising the Efficiency of Transport Performance*. Akademiia Nauk, USSR. (Kantorovich confirmed the paper was completed and submitted in 1940, but publication delayed by the Second World War.)
Karmarkar, N. 1984. A new polynomial-time algorithm for linear programming. *Combinatorica* 4: 373–395.
Karush, W. 1939. *Minima of functions of several variables with inequalities as side constraints*. MSc dissertation, Department of Mathematics, University of Chicago.
Kersey, C. 1989. *Unstoppable*. Naperville: Sourcebooks, Inc.
Khachiyan, L.G. 1979. A polynomial algorithm in linear programming in Russian. *Doklady Akademii Nauk SSSR* 244, 1093–6. English translation: *Soviet Mathematics Doklady* 20 (1979), 191–4.
Klee, V., and G.J. Minty. 1972. How good is the simplex algorithm? In *Inequalities III*, ed. O. Shisha. New York: Academic.
Koopmans, T.C. 1947. Optimum utilization of the transportation system. *Proceedings of the International Statistical Conferences, 1947*, vol. 5. Washington D.C. Also in Econometrica 16 (1948), 66–8.
Koopmans, T.C. 1949. Optimum utilization of the transportation system. *Econometrica* 17(Supplement): 136–146.
Koopmans, T.C. (ed.). 1951. *Activity analysis of production and allocation*. New York: Wiley.
Koopmans, T.C. 1960. A note about Kantorovich's paper, 'Mathematical Methods of Organizing and Planning Production'. *Management Science* 6: 363–365.
Kuhn, H.W., and A.W. Tucker. 1951. Nonlinear Programming. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, ed. J. Neyman. Berkeley: University of California Press.
Lemke, C.E. 1954. The dual method of solving the linear programming problem. *Naval Research Logistics Quarterly* 1: 36–47.
Lemke, C.E. 1965. Bimatrix equilibrium points and mathematical programming. *Management Sciences* 11: 681–689.
Leontief, W.W. 1936. Quantitative input and output relations in the economic system of the United States. *Review of Economic Statistics* 18: 105–125.
Lustig, I. 2001. e-optimization.com. *Interview with G.B. Dantzig*. Online. Available at http://e-optimization.com/directory/trailblazers/dantzig. Accessed 29 Dec 2006.
Markowitz, H.M. 1956. The optimization of a quadratic function subject to linear constraints. *Naval Research Logistics Quarterly* 3: 111–133. RAND Research Memorandum RM-1438, 1955.
Orchard-Hays, W. 1954. *A composite simplex algorithm-II*, RAND Research Memorandum RM-1275. Santa Monica: RAND Corporation.
Rockafellar, R.T. 1970. *Convex analysis*. Princeton: Princeton Press.
Scarf, H. 1967. The core of an N-person game. *Econometrica* 35: 50–69.
Scarf, H. 1973. *The computation of economic equilibria*. New Haven: Yale University Press.

Schrijver, A. 1986. *Theory of linear and integer programming*. Chichester: Wiley.
Smale, S. 1976. A convergent process of price adjustment and global Newton methods. *Journal of Mathematical Economics* 3: 1–14.
Smale, S. 1983. The problem of the average speed of the simplex method. In *Mathematical programming: The state of the art*, ed. A. Bachem, M. Grötschel, and B. Korte. Berlin: Springer.
Stigler, G.J. 1945. The cost of subsistence. *Journal of Farm Economics* 27: 303–314.
Todd, M.J. 1996. Potential-reduction methods in mathematical programming. *Mathematical Programming* 76: 3–45.
von Neumann, J. 1947. Discussion of a maximum problem. Unpublished working paper dated 15–16 November. In *John von Neumann: Collected works*, vol. 6, ed. A.H. Taub. Oxford: Pergamon Press, 1963.
von Neumann, J., and O. Morgenstern. 1944. *Theory of games and economic behavior*. Princeton: Princeton University Press.

Data Filters

Timothy Cogley

Abstract

Empirical economists often filter data prior to analysis to remove features that are a nuisance from the point of view of their theoretical models. Examples include trends and seasonals. This article describes how data filters work and the rationale that lies behind them. It focuses on the Baxter–King and Hodrick–Prescott filters, which are popular for measuring business cycles.

Keywords

ARIMA models; Band-pass filters; Baxter–King filter; Business cycle measurement; Cramer's representation theorem; Data filters; Deterministic linear trends; Gaussian log likelihood; Generalized method of moments; Granger causation; High-pass filters; Hodrick–Prescott filter; Impulse response function; Rational-expectations business-cycle models; Seasonal adjustment; Seasonal fluctuations; Shocks; Spurious cycle problem; Stochastic general equilibrium models; Stochastic trends; Trend reversion; Vector autoregressions

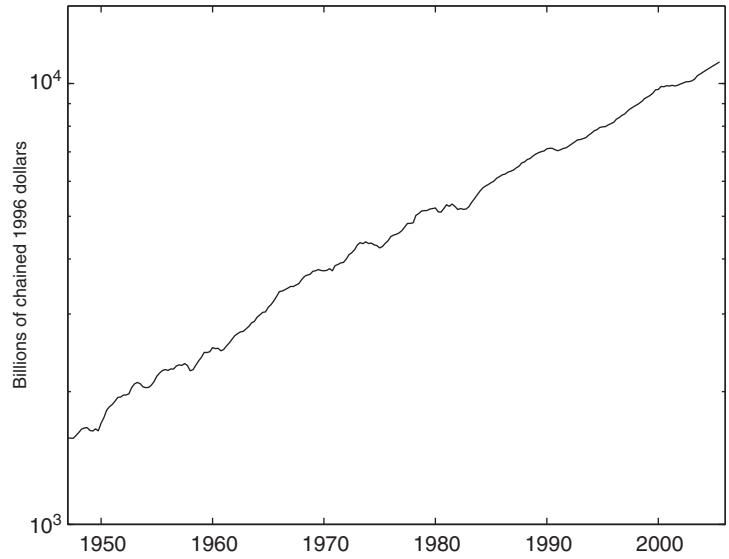
JEL Classifications

C2; C4

Economic models are by definition incomplete representations of reality. Modellers typically abstract from many features of the data in order to focus on one or more components of interest. Similarly, when confronting data, empirical economists must somehow isolate features of interest and eliminate elements that are a nuisance from the point of view of the theoretical models they are studying. Data filters are sometimes used to do that.

For example, Fig. 1 portrays the natural logarithm of US GDP. Its dominant feature is sustained growth, but business cycle modellers often abstract from this feature in order to concentrate on the transient ups and downs. To relate business cycle models to data, empirical macroeconomists frequently filter the data prior to analysis to remove the growth component. Until the 1980s, the most common way to do that was to estimate and subtract a deterministic linear trend. Linear de-trending is conceptually unattractive, however, because it presupposes that all shocks are neutral in the long run. While some disturbances – such as those to monetary policy – probably are neutral in the long run, others probably are not. For instance, a technical innovation is likely to remain relevant for production until it is superseded by another, later technical innovation.

The desire to model permanent shocks in macroeconomic time series led to the development of a variety of stochastic de-trending methods. For example, Beveridge and Nelson (1981) define a stochastic trend in terms of the level to which a time series is expected to converge in the long run. Blanchard and Quah (1989) adopt a more structural approach, enforcing identifying restrictions in a vector autoregression that separate permanent shocks that drive long-run movements from the transitory disturbances which account for cyclical fluctuations.



Data Filters, Fig. 1 Real US GDP, 1947–2006 (*Source*: Federal reserve economic database)

Another popular way to measure business cycles involves application of band-pass and high-pass filters. Engle (1974) was one of the first to introduce band-pass filters to economics. In the business cycle literature, the work of Hodrick and Prescott (1997) and Baxter and King (1999) has been especially influential. Figure 2 illustrates measures of the business cycle that emerge from the Baxter–King and Hodrick–Prescott filters.

In this article, I describe how data filters work and explain the theoretical rationale that lies behind them. I focus on the problem of measuring business cycles because that is one of the principal areas of application. Many of the issues that arise in this context are also relevant for discussions of seasonal adjustment. For a review of that literature, see Fok et al. (2006).

How Data Filters Work

The starting point is the Cramer representation theorem. Cramer's theorem states that a covariance stationary random variable x_t can be expressed as

$$x_t - \mu_x = \int_{-\pi}^{\pi} \exp(i\omega t) dZ_x(\omega), \qquad (1)$$

where μ_x is the mean, t indexes time, $i = \sqrt{-1}$, ω represents frequency, and $dZ_x(\omega)$ is a mean zero, complex-valued random variable that is continuous in ω. The complex variate $dZ_x(\omega)$ is uncorrelated across frequencies, and at a given frequency its variance is proportional to the spectral density $f_{xx}(\omega)$. If we integrate the spectrum across frequencies, we get the variance of x_t,

$$\sigma_x^2 = \int_{-\pi}^{\pi} f_{xx}(\omega) d\omega. \qquad (2)$$

This theorem provides a basis for decomposing x_t and its variance by frequency. It is perfectly sensible to speak of long- and short-run variation by identifying the long run with low-frequency components and the short run with high-frequency oscillations. High frequency means that many complete cycles occur within a given time span, while low frequency means the opposite.

Baxter and King (1999) define a business cycle in terms of the periodic components $dZ_x(\omega)$. They partition x_t into three pieces: a trend, a cycle, and irregular fluctuations. Inspired by the NBER business cycle chronology, they say the business cycle consists of periodic components whose frequencies lie between 1.5 and 8 years per cycle. Those whose cycle length is longer than 8 years are

identified with the trend, and the remainder are consigned to the irregular component.

The units for ω are radians per unit time. A more intuitive measure of frequency is units of time per cycle, which is given by the transformation $\lambda = 2\pi/\omega$. Often we work with quarterly data. To find the ω corresponding to a cycle length of 1.5 years, just set $\lambda_h = 6$ quarters per cycle and solve for $\omega_h = 2\pi/6 = \pi/3$. Similarly, the frequency corresponding to a cycle length of 8 years is $\omega_l = 2\pi/32 = \pi/16$. Baxter and King define the interval $[\pi/16, \pi/3]$ as 'business cycle frequencies'. The interval $[0,\pi/16)$ corresponds to the trend, and $(\pi/3, \pi]$ defines irregular fluctuations. One nice feature of the Baxter–King filter is that it can be easily adjusted to accommodate data sampled monthly or annually, just be resetting ω_l and ω_h.

To extract the business cycle component, we need to weigh the components $dZx(\omega)$ in accordance with Baxter and King's definition and integrate across frequencies,

$$x_t^B = \int_{-\pi}^{\pi} B(\omega)\exp(i\omega t)\, dZ_x(\omega), \tag{3}$$

where

$$B(\omega) = 1 \text{ for } \omega \in [\pi/16, \pi/3] \text{ or } [-\pi/3, -\pi/16], \\ = 0 \text{ otherwise.} \tag{4}$$

In technical jargon, $B(\omega)$ is an example of a 'band-pass' filter: the filter passes periodic components that lie within a pre-specified frequency band and eliminates everything else. The Baxter–King filter suppresses all fluctuations that are too long or short to be classified as part of the business cycle and allows the remaining elements to pass through without alteration.

Many economists are more comfortable working in time domain, and for that purpose it is helpful to express the cyclical component as a two-sided moving average,

$$x_t^B = \sum_{j=-\infty}^{\infty} \beta_j \left(x_{t+j} - \mu_x\right). \tag{5}$$

The lag coefficients can be found by solving

$$\beta_j = \frac{1}{2\pi}\int_{-\pi}^{\pi} B(\omega)\exp(i\omega j)\, d\omega. \tag{6}$$

The solution is

$$\beta_0 = \frac{\omega_h - \omega_l}{\pi}, \\ \beta_j = \frac{\sin(\omega_h j) - \sin(\omega_l j)}{\pi j} \text{ for } j \neq 0. \tag{7}$$

Notice that an ideal band-pass filter cannot be implemented in actual data samples because it involves infinitely many leads and lags. In practice, economists approximate x_t^B with finite-order moving averages,

$$x_t^B \doteq \sum_{j=-n}^{n} \tilde{\beta}_j \left(x_{t+j} - \mu_x\right). \tag{8}$$

Baxter and King (1999) and Christiano and Fitzgerald (2003) analyse how to choose the lag weights $\tilde{\beta}_j$ in order to best approximate the ideal measure for a given n.

For real-time applications, the two-sided nature of the filter is a drawback because the current output of the filter depends on future values of x_{t+j}, which are not yet available. Kaiser and Maravall (2001) address this problem by supplementing the filter with an auxiliary forecasting model such as a vector autoregression or univariate ARIMA model, replacing future x_{t+j} with forecasted values. This substantially reduces the approximation error near the end of samples.

That the filter is two-sided is also relevant for models that require careful attention to the timing of information. Economic hypotheses can often be formulated as a statement that some variable z_t should be uncorrelated with any variable known in period $t-1$ or earlier. These hypotheses can be examined by testing for absence of Granger causation from a collection of potential predictors to z_t. The output of a two-sided filter should never be included among those predictors, however, for that would put information about present and

D

future conditions on the right-hand side of the regression and bias the test towards a false finding of Granger causation. Similar comments apply to the choice of instruments in generalized-method-of-moments problems. For applications like these, one-sided filters are needed in order to respect the integrity of the information flow.

While Baxter and King favour a three-part decomposition, other economists prefer a two-part classification in which the highest frequencies also count as part of the business cycle. The trend component is still defined in terms of fluctuations lasting more than eight years, but the cyclical component now consists of all oscillations lasting eight years or less. To construct this measure, we define a new filter $H(\omega)$ such that

$$H(\omega) = 1 \text{ for } \omega \in [\pi/16, \pi] \text{ or } [-\pi, -\pi/16], \\ = 0 \text{ otherwise.} \tag{9}$$

This is known as a 'high-pass' filter because it passes all components at frequencies higher than some pre-specified value and eliminates everything else. If we use this filter in the Cramer representation, we can extract a new measure of the business cycle by computing

$$x_t^H = \int_{-\pi}^{\pi} H(\omega) \exp(i\omega t) dZ_x(\omega). \tag{10}$$

Once again, this corresponds to a two-sided, infinite-order moving average of the original series x_t,

$$x_t^H = \int_{j=-\infty}^{\infty} \gamma_j (x_{t+j} - \mu_x), \tag{11}$$

with lag coefficients $\gamma_0 = 1 - \omega_l/\pi$ and $\gamma_j = -\sin(\omega_l j)/\pi j$. As before, this involves infinitely many leads and lags, so an approximation is needed to make it work. The approximation results of Baxter and King (1999) and Christiano and Fitzgerald (2003) apply here as well.

Hodrick and Prescott (1997) also seek a two-part decomposition of x_t. They proceed heuristically, identifying the trend τ_t and the cycle c_t by minimizing the variance of the cycle subject to a penalty for variation in the second difference of the trend,

$$\min_{\{\tau_t\}} \left\{ \sum_{t=-\infty}^{\infty} \left[(x_t - \tau_t)^2 + \phi(\tau_{t+1} - 2\tau_t + \tau_{t-1})^2 \right] \right\}. \tag{12}$$

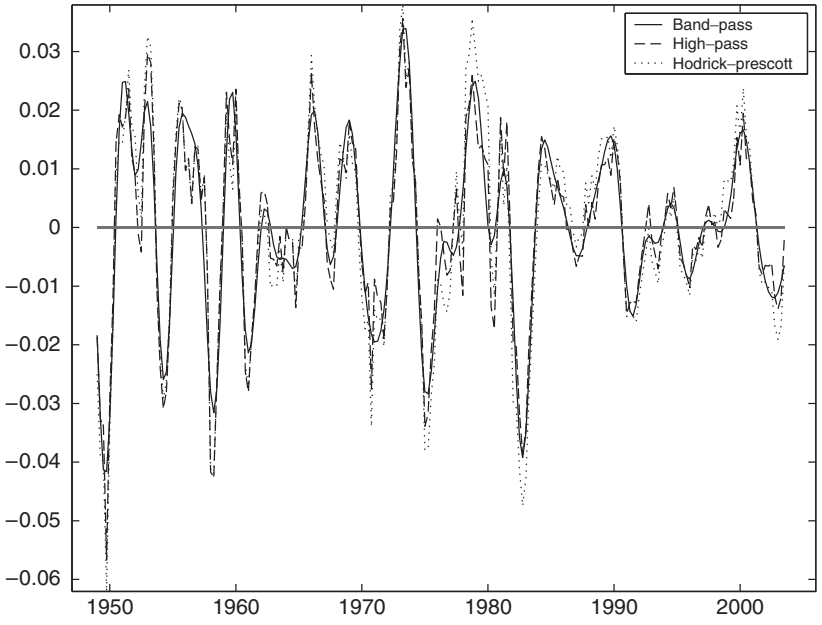
The Lagrange multiplier ϕ controls the smoothness of the trend component. After experimenting with US data, Hodrick and Prescott set $\phi = 1600$, a choice still used in most macroeconomic applications involving quarterly data. After differentiating (12) with respect to τ_t and rearranging the first-order condition, one finds that c_t can be expressed as an infinite-order, two-sided moving average of x_t,

$$c_t = HP(L)x_t \\ = \frac{\phi(1-L)^2(1-L^{-1})^2}{1+\phi(1-L)^2(1-L^{-1})^2} x_t, \tag{13}$$

where L is the lag operator. Although Hodrick and Prescott's derivation is heuristic, King and Rebelo (1993) demonstrate that $HP(L)$ can be interpreted rigorously as an approximation to a high-pass filter with a cut-off frequency of eight years per cycle. The close connection between the two filters is also apparent in Fig. 2, which shows that high-pass and Hodrick–Prescott filtered GDP are highly correlated.

Data Filters for Measuring of Business Cycles?

While data filters are very popular, there is some controversy about whether they represent appealing definitions of the business cycle. For one, there is a disconnect between the theory and macroeconomic applications, for the theory applies to stationary random processes and applications involve non-stationary variables. This is not critical, however, because the time-domain filters $\beta(L)$, $\gamma(L)$, and $HP(L)$ all embed difference operators, so business cycle components are stationary even if xt has a unit root.



Data Filters, Fig. 2 Filtered GDP, 1949–2003 (*Sources*: Federal reserve economic database and author's calculations)

A more fundamental criticism concerns the fact that the Baxter–King definition represents a deterministic vision of the business cycle. According to a theorem of Szego, Kolmogorov, and Krein, the prediction error variance can be expressed as

$$\sigma_{\varepsilon}^2 = 2\pi \exp\left[\frac{1}{2\pi}\int_{-\pi}^{\pi} \log f_{BC}(\omega)d\omega\right], \qquad (14)$$

where $f_{BC}(\omega)$ is the spectrum for the business-cycle component (see Granger and Newbold 1986, pp. 135–6). For an ideal band-pass filter, the spectrum of x_t^B is

$$f_{BC}(\omega) = |B(\omega)|^2 f_{xx}(\omega). \qquad (15)$$

Since $B(\omega) = 0$ outside of business cycle frequencies, it follows that $f_{BC}(\omega) = 0$ on a measurable interval of frequencies. But then Eq. (14) implies $\sigma_{\varepsilon}^2 = 0$, which means that x_t^B is perfectly predictable from its own past. The same is true of measures based on ideal high-pass filters. A variable that is perfectly predictable based on its own history is said to be 'linearly deterministic'. Thus, according to the Baxter–King definition, the business cycle is linearly deterministic.

In practice, of course, measured cycles are not perfectly predictable because actual filters only approximate the ideal. But this means that innovations in measured cycles are due solely to approximation errors in the filter, not to something intrinsic in the concept. The better the approximation, the closer the measures are to determinism.

How to square this deterministic vision with stochastic general equilibrium models is not obvious. Engle (1974); Sims (1993) and Hansen and Sargent (1993) suggest one rationale. They were interested in estimating models that are well specified at some frequencies but mis-specified at others. Engle studied linear regressions and showed how to estimate parameters by band-spectrum regression. This essentially amounts to running regressions involving band-pass filtered data, but band-pass filtering induces serial correlation in the residuals, and Engle showed how to adjust for this when calculating standard errors and other test statistics. He also developed

methods for diagnosing mis-specification on particular frequency bands.

Sims (1993) and Hansen and Sargent (1993) are interested in fitting a rational-expectations model of the business cycle to data that contain seasonal fluctuations. They imagine that the model abstracted from seasonal features, as is common in practice, and they wonder whether estimates could be improved by filtering the data with a narrow band-pass filter centred on seasonal frequencies. They find that seasonal filtering does help, because otherwise parameters governing business cycle features would be distorted to fit unmodelled seasonal fluctuations. Filtering out the seasonals lets the business cycle parameters fit business cycle features.

Business cycle modellers also frequently abstract from trends, and Hansen and Sargent conjectured that the same rationale would apply to trend filtering. Cogley (2001) studies this conjecture but finds disappointing results. The double-filtering strategy common in business cycle research (which applies the filter to both the data and the model) has no effect on periodic terms in a Gaussian log likelihood, so it is irrelevant for estimation. The seasonal analogy (which filters the data but not the model) also fails, but for a different reason. The key assumption underlying the work of Engle, Sims, and Hansen and Sargent is that specification errors are confined to a narrow frequency band whose location is known a priori. That is true of the seasonal problem but not of the trend problem. Contrary to intuition, trend-specification errors spread throughout the frequency domain and are not quarantined to low frequencies. That difference explains why the promising results on seasonality do not carry over to trend filtering.

Finally, some economists question whether filter-based measures capture an important feature of business cycles. Beveridge and Nelson (1981) believe that trend reversion is a defining characteristic of the business cycle. They say that expected growth should be higher than average at the trough of a recession because agents can look forward to a period of catching up to compensate for past output losses. By the same token, expected growth should be lower than average at the peak of an expansion. Cochrane (1994) confirms that this is a feature of US business cycles by studying a vector autoregression for consumption and GDP.

Cogley and Nason (1995) consider what would happen if x_t were a random walk with drift. For a random walk, expected growth is constant regardless of whether the level is a local maximum or minimum. Because it lacks the catching-up feature, many economists would say that a random walk is acyclical. Nevertheless, when the Hodrick–Prescott filter is applied to a random walk, a large and persistent cycle emerges. Thus the Hodrick–Prescott filter can create a business cycle even if no trend reversion is present in the original data. Cogley and Nason call this a spurious cycle. Furthermore, the problem is not unique to the Hodrick–Prescott filter; Benati (2001); Murray (2003) and Osborn (1995) document similar results for band-pass filters and for other approximations to high-pass filters.

Conclusion

Christiano and Fitzgerald remark that data filters are not for everyone. They are certainly convenient for constructing rough and ready measures of the business cycle, and they produce nice pictures when applied to US data. But some economists worry about the spurious cycle problem, especially in applications to business cycle models where the existence and properties of business cycles are points to be established. In much of that literature, attention has shifted away from replicating properties of filtered data to matching the shape of impulse response functions.

See Also

- ▶ Business Cycle Measurement
- ▶ Seasonal Adjustment
- ▶ Spectral Analysis
- ▶ Structural Vector Autoregressions
- ▶ Trend/Cycle Decomposition

Bibliography

Baxter, M., and R. King. 1999. Measuring business cycles: Approximate band-pass filters for economic time series. *Review of Economics and Statistics* 81: 575–593.

Benati, L. 2001. Band-pass filtering, cointegration, and business cycle analysis. Working Paper No. 142, Bank of England.

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

Blanchard, O., and D. Quah. 1989. The dynamic effects of aggregate demand and supply disturbances. *American Economic Review* 79: 655–673.

Christiano, L., and T. Fitzgerald. 2003. The band pass filter. *International Economic Review* 44: 435–465.

Cochrane, J. 1994. Permanent and transitory components of GNP and stock prices. *Quarterly Journal of Economics* 109: 241–265.

Cogley, T. 2001. Estimating and testing rational expectations models when the trend specification is uncertain. *Journal of Economic Dynamics and Control* 25: 1485–1525.

Cogley, T., and J. Nason. 1995. Effects of the Hodrick–Prescott filter on trend and difference stationary time series: Implications for business cycle research. *Journal of Economic Dynamics and Control* 19: 253–278.

Engle, R. 1974. Band-spectrum regression. *International Economic Review* 15: 1–11.

Fok, D., P. Franses, and R. Paap. 2006. Comparing seasonal adjustment methods. In *Palgrave handbook of econometrics: Volume1; Econometric theory*, ed. T. Mills and K. Patterson. Basingstoke: Palgrave Macmillan.

Granger, C., and P. Newbold. 1986. *Forecasting economic time series*. New York: Academic Press.

Hansen, L., and T. Sargent. 1993. Seasonality and approximation errors in rational expectations models. *Journal of Econometrics* 55: 21–55.

Hodrick, R., and E. Prescott. 1997. Postwar U.S. business cycles: An empirical investigation. *Journal of Money Credit and Banking* 29: 1–16.

Kaiser, R., and A. Maravall. 2001. *Measuring business cycles in economic time series*. New York: Springer.

King, R., and S. Rebelo. 1993. Low-frequency filtering and real business cycles. *Journal of Economic Dynamics and Control* 17: 207–231.

Murray, C. 2003. Cyclical properties of Baxter–King filtered time series. *Review of Economics and Statistics* 85: 472–476.

Osborn, D. 1995. Moving average detrending and the analysis of business cycles. *Oxford Bulletin of Economics and Statistics* 57: 547–558.

Sims, C. 1993. Rational expectations modeling with seasonally adjusted data. *Journal of Econometrics* 55: 9–19.

Data Mining

Clinton A. Greene

Abstract

Data mining is defined by presenting an example contrasting the role of specification search in economics to its role in experimental science. Historical references are provided, along with a short review of contemporary proposals to remedy sources of and problems with data mining.

Keywords

Data mining; Model selection; Regression analysis; Specification problems in econometrics

JEL Classifications

B4

‘Data mining’ and the older word ‘fishing’ are pejorative terms for illusory or distorted statistical inference from an empirical regression model, where the distortion results from explorations of various models in a single sample of data. This process usually involves adding or dropping variables, but may involve exploring a variety of alternative nonlinear functional forms or data subsamples. Data mining properly applies as a derogatory term only when exploratory results are used for inference within the sample used in exploration. But the term is sometimes used to refer to the exploratory process itself, as economists emphasize inference over data exploration, and even use inference to discuss exploratory activities. Some take data mining to be a more serious offence when there is conscious effort to manipulate, although data mining will distort results regardless of intent.

Importance and History

Some economists consider data mining to be pervasive in applied work. But the portion

subscribing to this view is unclear, since those who do so understandably retreat from applied work into economic or econometric theory. Leamer and Leonard (1983, p. 306) give voice to the view that collective data mining renders standard inference meaningless, and hence in general 'statistical analyses are either greatly discounted or completely ignored'. This stance may have reached a peak in the late 1970s, fuelled by an explosion in the volume of regression studies. But contemporary suspicion is still quite common. Kennedy (2003, pp. 82–3) characterizes the 'average economic regression' as perpetrating some of the worst data mining practices.

The issue was known to the originators of econometrics. Ragnar Frisch (1934) advocated methods to deal with the data mining issue which were applied into the 1950s, then neglected for two decades and reincarnated in modern form by Leamer (1983). Because Frisch found that differing but reasonable specifications could yield disparate results, he came to believe attempts at formal inference were illegitimate. Malinvaud (1966, chs. 1 and 2) provides a wonderful exposition of Frisch's methods and of why Frisch's stance was replaced by contemporary textbook assumptions. Even Haavelmo's (1944, ch. 7, sect. 17) founding statement of the contemporary inferential approach discusses data mining.

Econometrics textbooks quite properly warn against data mining, yet it is difficult to avoid and is pervasive in published work. This places the new practitioner in a difficult position. It is helpful to be armed with an understanding of the consequences of data mining and why data mining is difficult to avoid. Econometrics in the contemporary sense began when we decided that economic data could be treated as equivalent to sampling from an uncontrolled experiment (Haavelmo 1944), borrowing from R.A. Fisher's methods for experimental data. The following illustration clarifies these issues.

An Illustration

Suppose two students of the economy live in parallel universes. Both are interested in a variable y, believing the most important determinant of this variable y to be another variable x_1, but also supposing that variables x_2 and x_3 may be relevant. Their initial data-sets are identical, and they propose to model y via a linear regression model. Both start out assuming that the errors of the model (ε) are independent and normally distributed with constant variance. Thus they propose the model $y = b_1x_1 + b_2x_2 + b_3x_3 + \varepsilon$, where the coefficients 'b_i' are to be estimated.

The first student lives in a universe in which he can generate more data via experiments. The second student must wait passively for the passage of time before she can see more data; data generated by events she does not control. Thus, the first student is confident of his science, while the second student is in the actual universe of economics.

Now suppose that in their initial regression results for the coefficient on x_1 they find the sign is the opposite of what they expected. As in standard practice they take this to imply that they have omitted an important variable. After fiddling with their specifications they find that adding a variable x_4 yields a more sensible coefficient estimate for the variable x_1. Suppose also they find that, for the coefficients on x_2 and x_3, the null hypothesis for coefficients of zero would be accepted individually (leaving the other variable coefficient unrestricted, as in a t-test). But suppose they find the joint hypothesis ($b_2 = b_3 = 0$) would be rejected. They find the fit of the regression is penalized least by dropping the variable x_3 and do so. They have used a process of specification search to arrive at a model for y as a function of x_1, x_2 and x_4.

The first student takes the results to his professor. The professor commends the effort to learn from the world, but corrects the student on one point. He notes that, although the estimated standard error for the coefficient on x_3 included zero, it also included (we will suppose) five, and if this coefficient is truly so far from zero then (given expected variation in x_3) the variable x_3 would have appreciable effects. So the professor tells him to run another experiment designed so that the resulting data- set is large enough (and so standard errors of coefficient estimates are small enough) to usefully distinguish

between large and small values of b_3. The student does so, and publishes the results with the statistics and standard critical values treated as valid 'tests'. This is not data mining.

Now the second student takes her results to her professor. This professor says the first regression result (employing x_1, x_2 and x_3) can be treated as possibly generating test statistics drawn from standard distributions. However, in the final model (x_1, x_2 and x_4) some of the t-statistics were created by design. Since one 'fished' or fiddled with variables included in the model until the coefficient on x_1 had the correct sign, the t-statistic was drawn from a distribution such that there was 100 per cent probability it would have the 'correct' sign. Likewise the student explored specifications until the t-statistic for the coefficient on x_2 appeared to be significant. This implies for the final specification that within the interval bounded by the standard critical values (approximately plus or minus 2) the probability of the t-statistic for b_2 falling within this standard range must actually be zero, hardly a standard t-distribution. This process of modifying the model and re-estimating it using the same sample used to suggest those modifications will also affect in an unknown manner the distribution of other test statistics, even those that were not direct objects of exploration and design. These are data mined results.

Note that the two professors agree that something was potentially learned in the exploratory stage. Both students could use data exploration to reveal aspects of the first sample, but the results of exploration over this same sample could not then provide a formal test. As in any legitimate science, the first professor views taking inspiration from observation to be a process separate from confirmation or testing. The second student also hopes to have learned something from the sample, but her professor objects to treating the statistics resulting from this exploration as providing a test. The second student treated each regression as though it was a separate experiment, but regressions and their associated statistics are mere calculations that organize the data. Also note that, when these students took the initial estimate of b_1 as having the 'wrong sign', they were applying strong prior beliefs which led them to place little weight upon this empirical result. Bayesian inference provides a formal treatment of such priors.

The second student continues the consultation with her professor. The professor says these first results are not publishable because economists are interested in inference, and all she has shown is that the first model did not make sense. The professor may advise that she should first have chosen a successful regression model from the empirical literature, modifying it only slightly if at all. If the student is alert, she will notice the data available to her is identical to that in the literature, except for a few more recent observations.

So this alert student will go back to her professor and tell him she already knows the regression results will be the same as those already published, except to the extent the new data observations have some effect when averaged with the old. The test statistics will not have the usual distributions; instead, the distributions are a function of the previous results and the portion of new observations relative to those used in the previously published results. The student has discovered that, to the extent data-sets overlap, taking guidance from the regressions of other researchers is collective data mining, even if one runs only one regression oneself. Thus collective data mining is pervasive, and the meaning of published test statistics is unclear. Only if each data-set is entirely distinct can one learn from the work of others while preserving known statistic distributions.

Contemporary Practice and Remedies

Three partial remedies for data mining are practised in the current literature. One is to insist upon seeing all the possible regression results a reasonable researcher might propose, supplementing imperfect 'tests' with a range of results. This is most associated with Leamer (1983), but we have already mentioned the earlier work of Frisch. Current practice is moving towards this approach, more often presenting multiple specifications.

A second remedy is inspired by noting that it is possible to calculate probabilities for statistics resulting from specification search, if the process begins with a model including a set of variables large enough that the true model is reasonably assumed nested within, and respecification deletes and does not add variables. An example is the general-to-specific approach. This approach is now common when specifying lag-lengths of time-series models, but in other contexts is controversial. The statistical consequences of such an approach fall under the heading of 'pretest' estimators discussed in most econometrics textbooks, but the best introductory discussion is found in Campos et al. (2005, Introduction, sects. 3.3–3.4). Interestingly, Hoover and Perez (1999) show that when pretest distributions are not accounted for this second remedy leads to an acceptable level of distortion.

A third remedy reserves some of the available data for 'out-of-sample' tests. Here one engages in specification search in one portion of the data and then tests in the reserved portion. We place 'out-of-sample' in quotes because this is not confirmation in a new sample. This response cannot avoid collective data mining because it is likely that among many projects the more satisfactory reserved-sample results will be selected for publication, if not by individual authors then through the collective filter of journal referees. But this remedy is useful to the individual researcher.

The first two remedies focus on data exploration, and only the third remedy adds the key scientific step of confirmation in separate data. Followers of the second remedy such as David Hendry and others of the 'London School' are often accused of data mining. Yet they have been the strongest proponents and practitioners of the third remedy, which provides the legitimate test in separate data, even inventing new out-of-sample tests such as for forecast encompassing. A good introduction to the second and third remedies is found in Charemza and Deadman (1997).

As noted in our discussion of the third remedy, universal adoption of these remedies cannot avoid collective data mining. Collective data mining would be avoided if upon accepting a paper the journal offered an explicit or implicit contract to accept a follow-up study. Formal and precise testing would be performed in the subsequent study employing only data not available for the initial paper. This is yet to be practised by any journal, so as a result the methodological issues remain troublesome, leaving room for vague and inconsistent norms across referees and journals. New practitioners must develop their own approaches to navigating these norms and practices, while deciding how to preserve their own sense of integrity.

See Also

- ▶ Bayesian Statistics
- ▶ Extreme Bounds Analysis
- ▶ Model Selection
- ▶ Specification Problems in Econometrics
- ▶ Spurious Regressions

Bibliography

Campos, J., N. Ericsson, and D. Hendry, ed. 2005. *General-to-specific modeling*. Cheltenham: Edward Elgar.

Charemza, W., and D. Deadman. 1997. *New directions in econometric practice*. Cheltenham: Edward Elgar.

Fisher, R. 1925. *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.

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

Haavelmo, T. 1944. The probability approach in econometrics. *Econometrica* 12 (supplement): iii–115.

Hoover, K., and S. Perez. 1999. Data mining reconsidered: Encompassing and the general-to-specific approach to specification search. *The Econometrics Journal* 2: 167–191.

Kennedy, P. 2003. *A guide to econometrics*. Cambridge: MIT Press.

Leamer, E. 1983. Let's take the con out of econometrics. *American Economic Review* 73: 31–43.

Leamer, E., and H. Leonard. 1983. Reporting the fragility of regression estimates. *The Review of Economics and Statistics* 65: 306–317.

Malinvaud, E. 1966. *Statistical methods of econometrics*. Chicago: Rand McNally and Company.

Davanzati, Bernardo (1529–1606)

Peter Groenewegen

Keywords

Barter; Davanzati, B.; Division of labour; Foreign exchange markets; Money; Paradox of value; Quantity theory of money; Scarcity

JEL Classifications

B31

Merchant, classical scholar, translator and economist, Davanzati was born in Florence where, apart from a period of residence in Lyon as a merchant, he worked until his death. His contributions to economics are contained in *Notizia dei cambi* (1582) which explains the operation of the foreign exchanges, and *Lezione delle Monete* (1588), translated into English in 1696 as *A Discourse Upon Coin* presumably because of its relevance to the recoinage controversies. Besides these economic writings, Davanzati produced a history of the English Reformation (1602) and a translation of Tacitus (1637) frequently described as a masterpiece of Italian literature.

Davanzati's observations on the foreign exchanges present a detailed discussion of the origins and practice of this art classified by him as the third type of mercantile transaction, the others being barter (goods for goods) and trade (goods for money). The analysis demonstrates how exchange rates fluctuate between gold points according to the supply and demand of bills, the gold points being determined by a risk premium, transport costs and interest lost while the funds are in transit. His illustration of a foreign exchange transaction by bills of exchange involving six parties residing in Lyon and Florence (1582, pp. 62–8) has been argued by De Roover (1963, p. 113) to be so instructive that, had it been more thoroughly studied by historians and economists, 'fewer blunders in the history of banking' would have been made.

Davanzati's lecture on coin is one of the earliest presentations of the metallist view of the origin and nature of money. He stresses the advantages of money over barter in facilitating both the division of labour and trade of 'superfluities' between cities and nations. In the metallist tradition, money is defined as 'Gold, Silver, or Copper, coin'd by Publick Authority at pleasure, and by the consent of Nations, made the Price and Measure of Things' (1588, p. 12). Non-metallic and nonconvertible money can only be made acceptable to the public through coercion. Money is therefore a human convention and its intrinsic value is small relative to its value as means of exchange. To explain this value, Davanzati presents an early quantity theory which relates the value of stocks of commodities to the world's money stock. Although he is aware of the importance of monetary circulation (he compares it to the importance of the circulation of blood in the animal body), he does not develop a concept of its velocity. The lecture on money concludes with a forceful critique of the practice of debasing the coinage, based on analysing its consequences and illustrated with many examples of the practice. Davanzati argues that this 'evil' can be avoided only by making 'Money pass according to its Intrinsick Value' (1588, p. 24). Davanzati's lecture has also been noted because of its hints at the so-called 'paradox of value' and its references to elements of scarcity and usefulness in the determination of commodity prices. This and other aspects of his work were noted by Galiani (1750). Earlier his views appear to have been well received by Locke who owned, annotated and may even have inspired the Toland translation (Harrison and Laslett 1965, p. 120).

Selected Works

1582. Notizia de'Cambi a M. Giulio del Caccia. In Scrittori classici Italiani di economia politica. Parte Antica vol. 2, ed. Pietro Custodi, Milan: G.G. Destefanis, 1804.

1588. *Lezione delle Monete.* Trans. John Toland as *A Discourse Upon Coin*, London: Awnsham and Churchill, 1969.
1602. Scisma d'Inghilterra sino alla morta della reina Maria ristretto in lingua propria Fiorentina. Milan.
1637. Gli Annali di C. Cornelio Tacito ... con la traduzione in volgar Fiorentino. Florence: Landini.

References

De Roover, R. 1963. *The rise and decline of the Medici Bank*. Cambridge, MA: Harvard University Press.
Galiani, F., ed. 1750. *Della Moneta*. In *Della Moneta e scritti inedite*, ed. A. Merola, Milan: Feltrinelli, 1963.
Harrison, J., and P. Laslett. 1965. *The library of John Locke*. Oxford: Clarendon Press. 1971.

Davenant, Charles (1656–1714)

Peter Groenewegen

Keywords
Davenant, C.; King, G.; King–Davenant law of demand; Mercantilism; Political arithmetic; Recoinage debates; Tax administration

JEL Classifications
B31

Economist and administrator. Born in London, eldest son of William Davenant, the playwright and Poet Laureate, he was educated at Cheam School, Surrey, and entered Balliol College, Oxford, in 1671, going down in 1673 without a degree to take over the management of his father's theatre. In 1675 he wrote a tragedy, *Circe* (Davenant, 1677), but the theatre gained him little financial success. He also obtained an LL.D from Cambridge in 1675 and practised law for a short period. From 1678 to 1689 he was Commissioner of Excise. He sat as MP for St Ives from 1685 to 1688 and represented Great Bedwin in the Tory interest following the elections of 1698 and 1700. The financial consequences of his loss of office as Excise Commissioner in 1689 and unsuccessful attempts in 1692 and 1694 to obtain other positions in the revenue service appear to have inspired a career as pamphleteer, starting in 1695. Until 1702, when he again obtained preferment by being appointed Secretary to the Commission for negotiating the union between England and Scotland, he produced a steady flow of political and economic writings dealing with aspects of taxation, public debt, monetary and trade questions, foreign policy and criticisms of Whig policy in general. In June 1703 he obtained the post of Inspector-General of Exports and Imports in the Customs Office, a position he retained till his death in 1714. Most of his political and commercial writings were collected by C.E. Whitworth (1771) but two manuscript works on money and credit (Davenant, 1695b and 1696) were not published till 1942 (Evans 1942).

Davenant's position in the history of economics rests on a variety of contributions. Initially, his work was largely depicted as typically that of an 'adherent of the mercantile theory' (Hughes 1894, p. 483), but 'Tory free trader' (Ashley 1900, p. 269) better describes his pronouncements on foreign trade policy as he particularly advocated the removal of trade restrictions, such as those affecting woollen exports, which benefited the landed interest by raising land values (Davenant 1695a, pp. 16–17; 1697, pp. 98–104). His free trade position is not unambiguous. Although Davenant's remark that 'Trade is by its nature free, finds its own channel, and best directeth its own course.' (1697, p. 98) is often quoted, the contradictory view that 'it is the prudence of a state to see that [its] industry, and stock, be not diverted from things profitable to the whole, and turned upon objects unprofitable, and perhaps dangerous to the public' (1697,

p. 107) is less frequently noticed. Schumpeter's (1954, p. 196, n.4, and p. 242) depiction of Davenant's work as 'comprehensive quasi-system' emphasizing the interdependence of economic activity is also rather difficult to sustain, though it is possible to quote isolated remarks from Davenant's works in support. For example, Davenant's statement that 'all trades have a mutual dependence one upon the other, and one begets another, and the loss of one frequently loses half the rest' (1697, p. 97) cannot really be described as the general theoretical proposition it appears to be. Its only use is to provide a basis for some special pleading on behalf of the East India trade. Waddell's conclusion (1958, p. 288) that Davenant was a person neither of 'exceptional ability, nor of any great strength of character' and 'a competent publicist' rather than 'an original thinker' or 'practical man of affairs' seems a more appropriate assessment from an examination of his economic writings.

Davenant's plea for the importance of 'political arithmetic' or 'the art of reasoning by figures, upon things relating to government' (1698, p. 128) provides a further claim to fame, partly because it made more readily available the fairly sophisticated national income and expenditure estimates of his friend Gregory King (1696). Most of Davenant's political arithmetic application relates to taxation and estimating the gains from trade in terms of bullion, but he himself also made a useful contribution to the collection of international trade data as part of his duties as Inspector-General of Exports and Imports.

The precise details of Davenant's association with Gregory King are not fully known, but their names are also linked in another famous 'statistical' exercise, the so-called King–Davenant law of demand, first noted by Thornton (1802) and Lauderdale (1966), and later extensively discussed by Jevons (1871, pp. 154–8), who on the evidence available to him cautiously attributed to Davenant the data on which the law is based (but see Barnett 1936, pp. 6–7). However, apart from providing these data, Davenant himself characteristically drew no such analytical conclusions from this information (1698, Part II, pp. 224–5; see Creedy, 1986, for a detailed discussion).

Davenant's contributions to the recoinage debates (1695b; 1696) are less well known because they were not included in Whitworth (1771). Full recoinage was not necessary in Davenant's view when the inferior (because clipped or worn) coins were still usefully employed in small retail transactions. In addition, the detrimental effects on the exchange rate and commodity prices of the deteriorating currency were greatly exaggerated. The rise in prices, Davenant argues, could be attributed to a great many other causes; the depreciated exchange rate was more easily explained by the substantial overseas remittances induced by the European war and was therefore better remedied by floating a public loan in Holland. Although in these essays, Davenant's exposition is not always complete, Evans (1942, p. vi) regards them as containing 'all the essential elements of the analysis of money and credit' and integrating 'the entire problem of currency and public finance'. Finally, Davenant's contributions to tax administration need to be recognized. They have been described as ' translating into principles, and trying to provide a reasonable justification for the practices that the more methodical and innovating officials (such as Pepys at the Navy Office and Admiralty, and Downing and Lowndes at the Treasury ...) were adopting and enforcing' and that in these matters of administrative thinking, unlike his economics, 'Davenant's viewpoint steadily became [dominant] in the course of the next century or so' (Hume 1974, p.477). His writings also remain a useful source for much information on trade and finance over the final decades of the Stuart monarchy.

See Also

► King, Gregory (1648–1712)

Selected Works

1677. *Circe,* A Tragedy. As it is acted at His Royal Highness the Duke of York's Theatre. *London: Richard Tonson.*

1695a. An essay on ways and means of supplying the war. *In Whitworth (1771, vol. 1).*

1695b. A memorial concerning the Coyn of England. *Reprinted in Evans (1942).*

1696. *A Memoriall Concerning Creditt.* Reprinted in Evans (1942).

1697. *An Essay on the East-India Trade.* Reprinted in Whitworth (1771, vol. 1).

1698. Discourses on the Public Revenues, and on the Trade of England in Two Parts. *Reprinted in Whitworth (1771, Part I in vol. 1; Part II in vol. 2).*

Bibliography

Ashley, W.J. 1900. The tory origin of free trade policy. In *Surveys historic and economic*, ed. W.J. Ashley. London: Longmans.

Barnett, G.E. 1936. *Two tracts by gregory king*. Baltimore: Johns Hopkins Press.

Creedy, J. 1986. On the King–Davenant 'law' of demand. *Scottish Journal of Political Economy* 33 (3): 193–212.

Evans, G.H. 1942. *Two manuscripts by charles davenant*. Baltimore: Johns Hopkins Reprints of Economic Tracts.

Hughes, D. 1894. Charles D'Avenant (1686–1714). In *Dictionary of political economy*, ed. R.H.I. Palgrave, vol. 1. London: Macmillan.

Hume, L.J. 1974. Charles Davenant on financial administration. *History of Political Economy* 6: 463–477.

Jevons, W.S. 1871. *Theory of political economy.* 4th ed, 1911. London: Macmillan.

King, G. (1696). *Natural and political observations and conclusions upon the state and condition of England.* Reprinted in Barnett.

Lauderdale, J.M., Eighth Earl of. 1804. *An inquiry into the nature and origin or public wealth.* Edinburgh. Repr. with an introduction and revisions from the 2nd edn, New York: Kelley, 1966.

Schumpeter, J.A. 1954. *History of economic analysis*, 1959. London: Allen & Unwin.

Thornton, H. 1802. *An inquiry into the nature and effects of the paper credit of Great Britain.* London.

Waddell, D. 1958. Charles Davenant (1656–1714), a biographical sketch. *Economic History Review* 11: 279–288.

Whitworth, Sir C.E. 1771. *The political and commercial works of that celebrated writer Charles D'Avenant LL. D.* London.

Davenport, Herbert Joseph (1861–1931)

Warren J. Samuels

Keywords

Absolutist value theory; American Economic Association; Davenport, H. J.; Positive economics; Relativistic price theory; Supply and demand; Veblen, T

JEL Classifications

B31

Davenport was born on 10 August 1861, in Wilmington, Vermont, and died on 16 June 1931, in New York City. He commenced a professorial career at the age of 41 after having been a land speculator (initially successful, but wiped out in the Panic of 1893) and high school teacher and principal. His academic work was at the University of South Dakota, Harvard Law School, Leipzig, Paris and Chicago (Ph.D., 1898). He taught at Chicago (1902–8), Missouri (1908–16) and Cornell (1916–29). He was President of the American Economic Association in 1920.

A leading, albeit somewhat iconoclastic, economic theorist of his day, he contributed to the reformulation of microeconomics from absolutist value theory to relativistic price theory. He stressed that, while there were real forces at work in the economy, identifying them as human desires and productive capacities, price itself reflected nothing more fundamental than a temporary equation of demand and supply. Prices are not determined *by* the margins but *at* the margins. Recognizing the limits imposed by a resultant superficiality and simultaneity of determination, he felt that economists qua economists need not inquire into the formation of desires or institutions but should study the pecuniary logic of phenomena from the standpoint of price in a society dominated by the private and acquisitive point of view. His economics focused on entrepreneurial

opportunity-cost adjustments and encompassed a non-normative distribution theory based directly on price theory.

While differing from his close friend Thorstein Veblen on certain substantive issues, Davenport's work nonetheless reflected the impact of Veblen's critiques of traditional theory and of the actual market economy. Emphasizing positive economics and rejecting apologetics (economic theory was not to be the monopoly of reactionaries), Davenport was willing to recognize that the search for private gain did not always conduce to social welfare, but this conclusion was not to be considered a part of economic science per se.

Selected Works

1896. *Outlines of economic theory.* New York: Macmillan.
1897. *Outlines of elementary economics.* New York: Macmillan.
1908. *Value and distribution.* Chicago: University of Chicago Press.
1913. *Economics of enterprise.* New York: Macmillan.
1935. *The economics of Alfred Marshall.* Ithaca: Cornell University Press.

Davidson, David (1854–1942)

Carl G. Uhr

Keywords

Böhm-Bawerk, E. von; Capital gains taxation; Capital theory; Consumption taxation; Davidson, D.; Progressive taxation; Taxation of income; Taxation of wealth; Wicksell, J. G. K

JEL Classifications

B31

Born into a Jewish merchant family in Stockholm, Davidson studied law and economics at Uppsala University from 1871, became a docent in 1878, professor extraordinarius from 1880 to 1889, and then professor ordinarius for 30 years until he retired in 1919. Frequently called on to serve on parliamentary committees from 1891 to 1931, Davidson's influence was strongly felt on Sweden's monetary and tax policies, for instance the 'gold exclusion policy' of 1916–1924.

In 1899 Davidson launched Sweden's first economic journal, *Ekonomisk Tidskrift,* to which he contributed almost all his work over 40 years as its owner and editor (in 1965 it was renamed *The Swedish Journal of Economics* and issued in English). This journal greatly stimulated economic research in Sweden with numerous contributions from, among others, Wicksell, Cassel, Lindahl, Myrdal and Ohlin.

Unlike Wicksell and Cassel, who published their works in German (later translated into English), all of Davidson's writings are in Swedish, none of them translated. This, and the fact that his work – five tracts 1878–1989, over 200 articles in his journal on a variety of subjects, plus chapters in several government reports – was never systematized in treatise form, accounts for his contributions to economics having been known, until recently, only to Scandinavian academics.

In his dissertation, *Bidrag till läran om de ekonomiska lagarna för kapitalbildningen* (A Contribution to the Theory of Capital Formation), Davidson anticipated Böhm-Bawerk's *Positive Theory of Capital* (1884). To Davidson, capital was generated in the main by the unequal distribution of income. To the wealthy, increases in present goods have small and declining utility relative to that of future goods. The latter are obtained in greater quantity, variety and value by investing savings for a return – interest – in production of capital goods which, indirectly, increase productivity. This perspective inverts the first of Böhm-Bawerk's famous 'three grounds' for interest, and transforms the third to a marginal productivity theory of waiting. In his later work Davidson adopted the substance of Wicksell's amendments and reconstruction of Böhm-Bawerk's capital theory.

Davidson's monetary theory is best understood from his response in articles of 1908–1925 to his friend Wicksell's path-breaking work in this area. *Inter alia,* Davidson criticized Wicksell's monetary norm of price level stability as inappropriate in conditions of 'commodity shortage'. Eventually, by 1925 Wicksell was moved to amend his norm to accommodate Davidson's critique (Uhr 1960, chs. 10 and 11).

In his early tract *Om beskattningsnormen vid inkomstskatten* (A Taxation Norm for the Income Tax, 1889), Davidson urged the replacement of Sweden's several property taxes and most of its excises by a progressive income tax with a uniquely broad base. It base was to include 'the citizen's potential consumption power' by levying the tax (*a*) on any increment in his net worth accrued (*whether realized or not*) between the end and the beginning of the tax year; and (*b*) also on his actual consumption spending during the year. Net worth increments accrue to a person as the value of his assets increases over that of his liabilities, due to savings, capital gains, bequests, and so on. Such gains confer potential consumption power, which should be taxed along with actual consumption spending out of income.

Over the years, aware of difficulties his proposed tax base would encounter as it called for annual balance sheet and income–consumption statements, Davidson conceded some simplifications on the tax declarations, and to taxing capital gains only when realized by the sale of value-appreciated assets. He also agreed that the tax rates levied on net worth increments would have to be lower than the rates levied on consumption expenditures.

These concessions notwithstanding, Sweden's parliament in its first comprehensive income tax of 1910 adopted only one part of Davidson's proposal. It passed a progressive tax on income as usually defined (rather than on consumption spending as such), and added to it a second title, a tax on net worth increments at rates substantially lower than on income. Largely due to Davidson, this combination of an income and a net worth increments tax has remained a standard feature in Sweden's tax system since 1910.

Selected Works

1878. *Bidrag till läran om de ekonomiska lagarna för kapitalbildningen* [A contribution to the theory of capital formation]. Uppsala.

1889. *Om beskattningsnormen vid inkomstskatten* [A taxation norm for the income tax]. Uppsala.

Bibliography

von Böhm-Bawerk, E. 1884. *The Positive Theory of Capital*. Trans. W. Smart. New York: G.E. Stechert & Co., 1930.

Heckscher, E.F. 1952. David Davidson. *International Economic Papers* 2: 111–135.

Uhr, C.G. 1960. *Economic doctrines of Knut Wicksell*. Berkeley: University of California Press.

Uhr, C.G. 1975. *Economic doctrines of David Davidson*. Uppsala: Studia Oeconomica Upsaliensis.

De Finetti, Bruno (1906–1985)

Giancarlo Gandolfo

De Finetti was born in Innsbruck, Austria, and died in Rome. After a degree in mathematics at Milan University, he chose practical activities rather than an academic career, and worked at the Istituto Centrale di Statistica (1927–31) and then at the Assicurazioni Generali (1931–46). Only later did he turn to an academic career and win a chair in Financial Mathematics at Trieste University (1939); from 1954 to 1961 he held the chair in the same subject at the University of Rome and from 1961 to 1976 the chair of Calculus of Probabilities at the same university. He was a member of the Accademia Nazionale dei Lincei and Fellow of the International Institute of Mathematical Statistics.

De Finetti's fame rests on his contributions to probability and to decision theory, but he also worked in descriptive statistics, mathematics and economics.

Together with Ramsey and Savage, de Finetti is one of the founders of the subjectivist approach to probability theory. The first illustrations

(in non-technical terms) of his conception are in (1930a) and (1931b). He considers probability as a purely subjective entity 'as it is conceived by all of us in everyday life'. The probability that a person attributes to the occurrence of an event is nothing more or less than the measure of the person's degree of confidence (hope, fear, . . .) in this event actually taking place. This can be interpreted as the amount (say, 0.72) that the person deems it fair to pay (or receive) in order to receive (or pay) the amount 1 if the event in question occurs. The mathematical theory was presented in his 1935 lectures at the Institut Poincaré (1937); see also (1970) and (1972).

De Finetti also introduced the important concept of *exchangeability* in probability (1929, 1930b, 1937, 1938) and proved the theorem on exchangeable variables named after him. Exchangeability is a weaker concept than independence and has been receiving increasing attention in probability theory (in fact, the natural assumption for a Bayesian is not independence, but exchangeability). In his 1935 Poincaré lectures (1937) he also treated the relations between the subjectivist point of view and the concept of exchangeability, which in his vision are at the basis of sound inductive reasoning and behaviour and, hence, of (statistical) decision theory (1959, 1961). It goes without saying that his position on the subject of statistical inference is fundamentally Bayesian.

In descriptive statistics he adhered to the functional concept according to which a statistic is an index selected on the basis of the single case (the aspects that one wants to stress, the aim of the statistical investigation, etc.); in (1931a) he stressed the importance of means which have the property of being associative.

Among his mathematical contributions the (1949) paper is especially interesting for economists. Here de Finetti investigates the conditions under which a concave function can be associated with a given 'convex stratification' (i.e. a one-parameter family of convex sets, one interior to the other as the parameter varies). The author also discusses the conditions for a quasi-concave function to be transformed into a concave one by means of an increasing function. This paper started the literature on the 'concavification' of quasi-concave functions. As the author pointed out, these investigations also bear on consumer theory – where the convex stratification is the indifference map and the associated function is the utility function.

De Finetti also wrote on economic problems, where he stressed the importance of rigorous reasoning and verification, and emphasized the idea that the scope of economics, freed from the tangle of individual and corporative interests, should always and only be that of realizing a collective optimum (in Pareto's sense) inspired by criteria of equity (1969). An important initiative of his for the diffusion and correct application of mathematical and econometric methods in economics was the annual CIME (Centro Internazionale Matematico Estivo) seminar that he organized from 1965 to 1975; this enabled young Italian economists to benefit from courses given by Frisch, Koopmans, Malinvaud, Morishima, Zellner, to mention only a few of the lecturers.

See Also

- ▶ Bayesian Inference
- ▶ Convexity
- ▶ Savage, Leonard J. (Jimmie) (1917–1971)
- ▶ Subjective Probability

Selected Works

A full bibliography of de Finetti's works up to 1980 is contained in B. de Finetti, *Scritti (1926–1930),* ed. L. Daboni et al. Padua: Cedam, 1981, with an autobiographical note.

1929. Funzione caratteristica di un fenomeno aleatorio. In *Atti del Congresso Internazionale dei Matematici* (1928). Bologna: Zanichelli, 179–190.

1930a. Fondamenti logici del ragionamento probabilistico. *Bolettino dell'Unione Matematica Italiana* 9:258–261.

1930b. Funzione caratteristica di un fenomeno aleatorio. *Memorie della Reale Academia dei Lincei,* Classe di scienze fisiche, matematiche e naturali, vol. IV, fasc. 5.

1931a. Sul concetto di media. *Giornale dell'Istituto Italiano degli Attuari* 2:369–396.
1931b. *Probabilismo. Saggio critico sulla teoria delle probabilità e sul valore della scienza.* Naples: Perrella; also in *Logos,* 1931, 163–219.
1937. La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré,* vol. VII, fasc. I. Trans. as 'Foresight: its logical laws, its subjective sources', in *Studies in subjective probability,* ed. H.E. Kyburg Jr. and H.E. Smokler. New York: Wiley, 1964.
1938. Sur la condition de 'équivalence partielle'. (Conférence au Colloque consacré à la théorie des probabilités, University of Geneva, 1937.) In *Actualités Scientifiques et Industrielles,* no. 739. Paris: Herman.
1949. Sulle stratificazioni convesse. *Annali di matematica pura e applicata* XXX:173–183, Series IV.
1959. La probabilità e la statistica nei rapporti con l'induzione, secondo i diversi punti di vista. In *Atti corso CIME su Induzione e Statistica* (Varenna). Rome: Cremonese, 1–115.
1961. Dans quel sens la théorie de la décision est-elle et doit-elle être 'normative'. In *Colloques internationaux du Centre National de la Récherche Scientifique.* Paris: CNRS, 159–169.
1969. *Un matematico e l'economia.* Milan: F. Angeli (anthology of previously published papers).
1970. *Teoria delle probabilità. Sintesi introducttiva con appendice critica*, 2 vols. Turin: Einaudi. Trans. as *Theory of probability,* 2 vols. New York: Wiley, 1974–5.
1972. *Probability, induction and statistics.* New York: Wiley (anthology of writings).

De Moivre, Abraham (1667–1754)

A. W. F. Edwards

De Moivre was born in Vitry-le-François on 26 May 1667, of French Protestant stock. Following the revocation of the Edict of Nantes in 1685 he fled to London, where he earned a precarious living as a mathematical author and tutor until his death there on 27 November 1754.

De Moivre was the most important writer on probability of his day, building on the work of Pascal, Fermat, Huygens and James Bernoulli. His *De mensura sortis* (On the measurement of lots) appeared in the *Philosophical Transactions* for 1711 and in ever-expanding form in English as the *Doctrine of Chances* (1718, 1738, 1756). It contained the first publication of the expression for the binomial distribution for general chances. The second edition (1738) included an English translation of the privately-circulated Latin pamphlet of 1733 in which De Moivre gave his celebrated Normal approximation to the binomial distribution '*A method of approximating the Sum of the Terms of the Binomial* $(a + b)^n$ *expanded into a Series from whence are deduced some practical Rules to estimate the Degree of Assent which is to be given to Experiments*.' De Moivre was fully seized of the importance of Bernoulli's limit theorem and its application to the problem of estimating a binomial parameter; this work replaced Bernoulli's 'very wide limits' by an approximation.

De Moivre also made important contributions to the 'Gambler's Ruin' problem, involving the question of the duration of play, to the use of generating functions, and to the study of annuities.

De Quincey, Thomas (1785–1859)

F. Y. Edgeworth

The son of a prosperous merchant, De Quincey was born in 1785, and, after a brilliant literary career, died in 1859. That a genius of so high an order of imagination found the abstract reasoning of political economy 'Not harsh and crabbed as dull fools suppose' is instructive. The fascination which the severer aspect of the science had for De Quincey is expressed in that passage of the *Confessions of an Opium Eater* where the writer

describes how he was aroused from lethargy by the study of Ricardo's *Political Economy* (1818). The fruit of that study appeared in the *Dialogues of Three Templars* (1824), a brilliant exposition and defence of the Ricardian theory of value. The paradox, for so De Quincey admits it to be in a good sense, that real value is measured by quantity of labour, that

> a million men may produce double or treble the amount of riches, of 'necessaries, conveniences, and amusements', in one state of society that they could in another, but will not on that account add anything to value (Ricardo, Political Economy; chapter on 'Value and Riches')

is expounded by the disciple even more fearlessly than by the master.

> 'My thesis,' says X, the Socrates of the dialogues, who represents the author's views, 'is that no such connection subsists between the two [the quantity obtained and the value obtaining] as warrants any inference that the real value is great because the quantity it buys is great, or small because the quantity it buys is small.' 'I have a barouche,' says the objector, 'which is worth about 600 guineas at this moment. Now, if I should keep this barouche unused in my coach-house for five years, and at the end of this term it should happen from any cause that carriages had doubled in value, my understanding would lead me to expect double the quantity of any commodity for which I might then exchange it, whether that were money, sugar, besoms, or anything whatsoever. But you tell me no.' . . . 'You are in the right,' replies X, 'I do tell you so . . . If A double its value, it will not therefore command double the former quantity of B' [B representing any assignable thing] (Fourth Dialogue).

The intelligent Bailey might well be stirred by these startling deductions to attempt a reply (preface to *Critical Dissertation*). In the later dialogues Ricardo's theory of value is defended against Malthus. This controversy had been commenced in the 'Measure of Value', published in the *London Magazine* for December 1823. An article on 'Malthus' in an earlier number of the same journal contains a mild attack on the theory of population. Some of the points are elucidated in a letter to Hazlitt which appeared in the *London Magazine*, December 1823. To the same period belongs a sort of *éloge* of Ricardo, which De Quincey, shortly after the death of his revered master, contributed to the *London Magazine*, March 1824.

De Quincey's latest and greatest economical work is the *Logic of Political Economy* (1844). The more original portion of this book may be described as a vindication of the part played by utility in the determination of value. The cause is just and the reasoning ingenious; yet the censure with which J.S. Mill tempers his copious citation from this discourse seems deserved (*Political Economy*, bk. iii, chapter ii, §1, and §3 *end*). Certainly De Quincey's illustrations are perfect. The rhinoceros which in the reign of Charles II was sold for a figure far above the cost of importation; the *Valdarfer* copy of Boccaccio which Lord Blandford bought for £2240 and afterwards, when in pecuniary embarrassments, was sold by auction and purchased for £750 by Lord Spencer, whom he outbid at the first sale; Popish reliques which had a high value, but no cost of production (p. 60 *et seq*., 1844 edn); these and other 'shining instances' throw light upon an obscure subject. The 'dry light' of logic is intensified by a coruscation of wit. Sometimes, however, the doubt occurs whether the writer was as competent to point a moral as to adorn a tale. Thus, in the case of the pearl-market, and the vividly pictured slave-market (ibid. p. 77 *et seq*.) is it correctly stated that for 'the *plebs* amongst the slaves', and the 'ordinary pearls', value is determined by cost of production, while 'the natural aristocracy amongst the slaves, like the rarer pearls, will be valued on other principles'?

Even the famous parable of the musical snuff-box (cited by Mill, *Political Economy*, bk. iii, chapter ii, § 1) is not rightly interpreted by its author. It is not in general true of a bargain between two isolated individuals that the price will be 'racked up to U' (ibid. pp. 25–27)–the measure of the 'intrinsic worth of the article in your individual estimate for your individual purposes'; in other words its *total utility* to the purchaser (cp. Mill, §1 *end*). The following passage seems more correct.

> The purpose which any article answers and the cost which it imposes must eternally form the two limits within which the tennis-ball of price flies backwards and forwards. Five guineas being, upon the particular article X, the maximum of teleologic price, the utmost sacrifice to which you would ever submit, under the fullest appreciation of the natural purposes which X can fulfil, and then only

D

> under the known alternative of losing it if you refuse the five guineas, this constitutes one pole, the aphelion, or remotest point to which the price for you could ever ascend.

The other limit is fixed by the cost of reproduction. These are 'the two limits between which the price must always be held potentially to oscillate' (ibid., pp. 105, 106). But even here it is not clearly stated that, in the absence of competition, the terms are indeterminate; the 'tennis-ball' may fall anywhere between the extreme limits. It is nowhere stated that in the presence of competition the upper limit is formed, not by *total*, but *final degree of utility*. De Quincey is far removed from the recent theorists to whom he bears a superficial resemblance by his not having attended to final utility and cognate conceptions. The connection between demand and value is denied by him on the strength of exceptional though striking instances (ibid., p. 331, quoted by Mill, bk. iii, chapter iii, § 2). 'A crazy maxim,' he says, 'has got possession of the whole world: viz. that price is, or can be, determined by the relation between supply and demand.' This imperfect conception of supply and demand is the special object of Mill's severe remarks on De Quincey. Mill's censure is endorsed by Sir Leslie Stephen in his article on De Quincey in the *Fortnightly Review* (1871). Mr. Shadworth Hodgson in one of his *Outcast Essays* has traversed this unfavourable verdict.

Whatever be the fate of De Quincey's cardinal tenets, it is certain that his occasional suggestions, the minor pearls of his discourse, enhanced as they are by a setting of consummate literary perfection, will preserve a lasting worth. Some important corrections of Ricardo's expressions deserve particular notice. De Quincey perceived, just as clearly as more recent critics, that 'the current rate of profits, as a thing settled and defined, must be a chimera'. He exposes

> the puerility of that little receipt current among economists, viz. unlimited competition for keeping down profits to one uniform level. . . . Everybody must see that it is a very elaborate problem to ascertain even for one year, still more for a fair average of years, what has been the rate of profits upon the capital employed in any one trade (ibid., p. 237 et seq.).

What more could Cliffe Leslie say? De Quincey complains much that Ricardo, while insisting on the tendency towards the degradation of soils (the law of diminishing returns) has not sufficiently emphasized the counter-tendency towards improvement in the arts of cultivation. 'The land is travelling downwards, but always the productive management of land is travelling upwards' (ibid., p. 239). De Quincey discerns what a handle is afforded by Ricardo's partial statement to 'the systematic enemies of property' . . . 'the policy of gloomy disorganising Jacobinism'. Rent is referred by De Quincey not to the 'indestructible', but the *differential* powers of the soil. Rent is defined as '*that portion of the produce from the soil* (or from any agency of production) *which is paid to the landlord for the use of its differential powers as measured by comparison with those of similar agencies operating on the same market*.'

The parenthesis exemplifies the pregnancy of De Quincey's occasional suggestions. In presenting the theory of rent, De Quincey employs an admirable geometrical construction. As in the construction which Prof. A. Marshall has made familiar (*Economics of Industry*, bk. ii, ch. iii), the *ordinate* in De Quincey's diagram represents produce. But the *abscissa* represents not doses of capital but qualities of soil. The two constructions have been combined by the present writer in an illustration of the *abstract theory of rent*, contributed to the British Association (Report, 1886). Referring to the use of diagrams, De Quincey well says:

> A construction (i.e. a geometrical exhibition) of any elaborate truth is not often practicable; but, wherever it is so, prudence will not allow it to be neglected. What is called evidentia, that sort of demonstration which shows out . . . is by a natural necessity more convincing to the learner. And, had Ricardo relied on this constructive mode of illustration his chapters upon rent and upon wages, they would not have tried the patience of his students in the way they have done.

Had De Quincey pursued his mathematical studies further, and applied the conceptions of the infinitesimal calculus to the theory of value, he would have escaped his capital error of having

confused integral (or total), with differential (or final) utility. If he had worked with *d*U, instead of U, he might have anticipated Jevons.

Selected Works

1844. *The logic of political economy.* Edinburgh/London: William Blackwood & Sons.
1889–1890. In *The collected writings of Thomas De Quincey*, 14 vols., ed. David Masson. Edinburgh: Adam & Charles Black.
1890. In *The uncollected writings of Thomas De Quincey*, 2 vols., ed. J. Hogg. London: Sonnenschein & Co.

Dear Money

Susan Howson

The obverse of cheap money, 'dear money' is also used to denote episodes in which central banks have raised (short-term) interest rates deliberately to bring about a contraction of money or credit, often in order to preserve a fixed exchange rate. The historical episodes are memorable for their effects on economic activity and on subsequent monetary theory and policy.

The major financial crises of the 19th century were accompanied by the Bank of England's raising of its discount rate (Bank rate) to at least 5% (the maximum permitted under the usury laws until 1833) in order to protect the gold reserve from an internal or external drain. The tradition as it developed after the Bank Charter Act of 1844 was for the Bank to act as a lender of last resort even when that involved an expansion of the fixed fiduciary note issue imposed by the Act, but at a penal rate. Hence Bank rate went to 8% in 1847, 10% in 1857 and again in 1866, 9% in 1873, but only 6% in the Baring crisis of 1890, the smooth handling of which was seen as a success for the Bank's methods (Hawtrey 1938, chs 1 and 3; Morgan 1943, chs 7–9; Clapham 1944, Vol. 2, ch. 6; Sayers 1976, pp. 1–3). In the early 20th century the events of the crisis of 1907 seemed to confirm the utility of central banks in general and the efficacy of Bank rate in particular. When the American stock exchange boom broke, Bank rate was quickly raised to 7% in response to gold outflows from London. The outflows were swiftly reversed while a banking panic in the US turned into a severe though short-lived slump. The outcome in the US was the establishment of the National Monetary Commission in 1908 and the Federal Reserve System which it recommended, in 1914. In Britain, belief in the power of interest rates to influence economic activity was reinforced, and lasted for a generation (Hawtrey 1938, pp. 115–18; Friedman and Schwartz 1963, pp. 156–74; Sayers 1957, pp. 62–4; Sayers 1976, pp. 54–60; Keynes 1930, Vol. I, ch. 13).

After World War I dear money was applied again, vigorously but after some hesitation, in both Britain and America to curb the postwar boom: Bank rate went to 6% in November 1919, 7% in April 1920, the Federal Reserve Bank of New York rediscount rate to 6% in January 1920. In both countries the rises came too late and were too strong: the restocking boom was already breaking and the subsequent slump was severe and (in the UK) prolonged (Friedman and Schwartz 1963, pp. 221–39; Howson 1974, 1975, ch. 2). The Federal System continued to experiment in the 1920s with the use of interest rates to control the domestic economy (Chandler 1958; Friedman and Schwartz 1963, ch. 6), but elsewhere, with many countries struggling to return to or maintain the international gold standard, dear money, in the sense of high (short-term) interest rates was frequently and widely used for balance of payments reasons (Clarke 1967; Moggridge 1972). It was with considerable relief that countries falling off the gold standard in the 1930s took advantage of their new-found monetary independence to promote cheap money. The revival of monetary policy on both sides of the Atlantic after 1951 did not involve the use of dear money in traditional

ways: concern with price stability was initially tempered by the objective of 'full employment' and in Britain at least interest rate rises for the sake of external balance were usually employed only as one element in 'packages' of deflationary measures; by the time the reduction of inflation became an important objective dear money as a target or as an indicator of monetary policy had been replaced by the rate of growth of the money supply (Dow 1964, ch. 3; OECD 1974; Blackaby 1978, chs 5 and 6).

See Also

- ▶ Bank Rate
- ▶ Cheap Money

References

Blackaby, F.T. (ed.). 1978. *British economic policy 1960–74*. Cambridge: Cambridge University Press.

Chandler, L.V. 1958. *Benjamin Strong: Central banker*. Washington, DC: Brookings Institution.

Clapham, Sir John. 1944. *The bank of England*. Cambridge: Cambridge University Press.

Clarke, S.V.O. 1967. *Central bank cooperation 1924–31*. New York: Federal Reserve Bank of New York.

Dow, J.C.R. 1964. *The management of the British economy 1945–60*. Cambridge: Cambridge University Press.

Friedman, M., and A.J. Schwartz. 1963. *A monetary history of the United States 1867–1960*. Princeton: Princeton University Press.

Hawtrey, R.G. 1938. *A century of bank rate*. London: Longmans Green & Co.

Howson, S. 1974. The origins of dear money, 1919–20. *Economic History Review* 27(1): 88–107.

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

Keynes, J.M. 1930. *A treatise on money*. London: Macmillan for the Royal Economic Society, 1971.

Moggridge, D.E. 1972. *British monetary policy 1924–1931*. Cambridge: Cambridge University Press.

Morgan, E.V. 1943. *The theory and practice of central banking 1797–1913*. Cambridge: Cambridge University Press.

OECD. 1974. *Monetary policy in the United States*. Paris: OECD.

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

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

Débouchés, Théorie des

Henry Higgs

Generally regarded as the main original contribution of J.B. Say to economic science, this *theory of outlets* or *of vent* affirms that a general glut or general over-production is impossible. If all products could be had for nothing, men would everywhere spring into existence to consume them. Products are bought with other products. Therefore each product is more in demand as other products increase and bid against it. In other words, as the same product constitutes the producer's demand and the consumer's supply, a general excess of supply over the general demand is absurd. Moreover, human desires expand indefinitely. So long as these are unsatisfied there can be no over-production except from lack of purchasing power arising from under-production on the part of the would-be purchasers.

Hence it is concluded that to maximize production is the interest of all; that industry is *solidaire;* and that cosmopolitanism in commerce is true wisdom, imports stimulating the sale of indigenous products. This theory, Say predicted, 'will change the politics of the world' (*Traité,* 5th edn, 1826, I. ciii).

The theory was resisted by Malthus and Sismondi, but was supported by James Mill and Ricardo, whose friendship grew out of this agreement, as we learn from J.S. Mill (*Principles,* 1875 edn, III. xiv). The last-mentioned writer's examination of the theory, though enforcing the strength of the main position, leaves still something to be desired. Arguments are used which take no account of the relativity of demand to price, the imperfection of the world market, or the element of time necessary to create new habits of production or consumption or to raise up a new generation of consumers. The case is, however, conclusive against those whose view involves the fallacy of a general fall of values, or who mistake the phenomenon of a commercial crisis, in times of

contracting credit, for over-production. The remedy, says J.S. Mill, for 'what may be indiscriminately called a glut of commodities or a dearth of money, is not a diminution of supply, but the restoration of confidence'.

Reprinted from *Palgrave's Dictionary of Political Economy.*

See Also

► Say's Law

Debreu, Gerard (1921–2004)

Lawrence E. Blume

Abstract

This article surveys the life and work of Gerard Debreu. Although his research was largely confined to general equilibrium theory and welfare economics, the influence of his work can be seen throughout contemporary economics.

Keywords

Allais, M.; Arrow, K.J.; Competitive equilibrium; Contingent commodities; Contract curve; Convexity; Core; Deadweight loss; Debreu, G.; Edgeworth, F.Y.; Excess demand; Existence of competitive equilibrium; General equilibrium; Incomplete markets; Inverse function theorem; Marginal rates of substitution; Mathematical economics; Models; Multiple equilibria; Natural preference models; Optimality; Risk aversion; Sard's theorem; Savage's subjective expected utility model; Separating hyperplane theorem; Social equilibrium; Value theory; Welfare economics

JEL Classifications

B31

Life

Gerard Debreu, the son of a Calais lace manufacturer, was born on 4 July 1921. He took his baccalauréat in 1939, just before the outbreak of the Second World War. Instead of entering university, he then began an improvised mathematics curriculum in Ambert and, later, in Grenoble. In 1941 he was admitted to the École normale supérieure, where he studied with Henri Cartan and the Bourbaki group. After D-Day he enlisted in the French Army, and served in Algeria and Germany. Returning to his studies, he completed the agrégation de mathématiques in early 1946. While pursuing his mathematical studies in Paris, he was captivated by Maurice Allais's (1943) exposition of the Walrasian general equilibrium analysis, which became the central pillar of his research programme. It was the flip of a coin which determined that he, rather than Edmond Malinvaud, would receive a travelling fellowship from the Rockefeller Foundation. This funded a year at Harvard, Berkeley and the Cowles Commission at Chicago, followed by studies at Uppsala and, with Ragnar Frisch, in Oslo. Debreu returned to Chicago and the Cowles Commission, and moved with it to Yale in 1955 with his wife of ten years and his nine- and five-year-old daughters. A year at the Center for Advanced Study in the Behavioral Sciences at Stanford gave the Debreu family a taste for California, and in 1962 Debreu accepted a position at the University of California at Berkeley. There he remained until his retirement. Debreu became a US citizen in 1975, having been deeply moved by America's response to the Watergate affair.

Gerard Debreu received numerous honours and awards. He was a Fellow of the American Academy of Arts and Sciences (1970), vice president and president of the Econometric Society (1970, 1971), a Chevalier de la Légion d'honneur (1976), a member of the National Academy of Sciences (1977), a Distinguished Fellow of the American Economic Association (1982) and its president in 1990, a Foreign Associate of the French Académie des sciences (1984) and a Fellow of the American Association for the

Advancement of Science (1984). He was awarded honorary degrees from, among many, the University of Bonn, Université de Lausanne, Northwestern University, Université des sciences sociales de Toulouse, and Yale University. Most prominent of all, in 1983 he was the recipient of the Bank of Sweden Prize in Economic Sciences in Memory of Alfred Nobel.

The elegance of Gerard Debreu's work was reflected in his personal style. He was also a competitive bridge player, and perhaps his first publication was a monograph on the game. In contrast to his revealed preference for the spare prose and clean, elegant arguments of the *Theory of Value* (1959) was his love of *A La Recherche du Temps Perdu*. 'My appreciation of Proust', he said in a 1983 *New York Times* interview, 'is in his style, subtlety and taste. I prize conciseness very much, and that is certainly something that you cannot accuse Proust of. His compulsion, as you know, eventually killed him. I'll try to escape that fate.' Debreu was reserved in person, but displayed a quick and subtle wit. I remember his beginning a lecture on the computation of economic equilibrium with the observation that the existence of equilibrium had been established and that now Herbert Scarf has taught us how to compute the zeros of the excess demand function. It only remains, he said, for the econometricians to estimate it, and we would be done. Gerard Debreu died in Paris on New Year's Eve 2004. His ashes were placed in a niche in the Père Lachaise cemetery, the final resting place of many of Frances's most eminent artists and intellectuals, including Marcel Proust.

Work

The influence of Gerard Debreu's work can be seen throughout contemporary economics, but his research output was largely confined to general equilibrium theory and its requirements.

The Existence of Competitive Equilibrium

Gerard Debreu's broad fame in the economics community is due to his work on the existence of competitive equilibrium. The complexity of simultaneous price and quantity determination in multiple markets of related and unrelated goods stands in stark contrast to the cutting power of the simple Marshallian scissors of supply and demand in a market with a single good. It is certainly not obvious that a multi-market equilibrium should exist. The existence problem, open since the publication of Léon Walras's *Éléments D'économie Politique Pure* (1874), was first given a broad and general treatment by Arrow and Debreu (1954a). As Arrow tells the story, in earlier work on the problem, he and Debreu had each made a mistake for which the other had a solution. It was suggested that they collude, and the outcome was displayed at the remarkable 1952 Winter Meeting of the Econometric Society in Chicago where both the Arrow and Debreu's paper (1954a) and McKenzie's (1954) paper were presented. The Arrow and Debreu 'private ownership economy' is today the standard reference for a general competitive model. McKenzie's treatment of technology is somewhat more special, although the two models are not directly comparable. The method of proof is to introduce a fictitious agent, a Walrasian auctioneer, whose role is to choose prices. Then the entire problem sets up like a non-cooperative game, with the added wrinkle that feasible strategies for one player may depend upon the choices of the others. Fortunately, Debreu (1952) had already established the existence of a kind of Nash equilibrium for these games, which he called a 'social equilibrium'. This approach to the existence of equilibrium is quite different from the approach through the excess demand correspondence, which was already developed in 1954 and appears in Debreu's (1959) essential masterwork, the *Theory of Value*. The social equilibrium approach is particularly well-suited to economies in which it is difficult to get one's hands on excess demand directly, such as economies with externalities, public sector decision-making, non-convexities, and incomplete and intransitive preferences.

Welfare Economics

The central question of economic analysis, the workings of the invisible hand, is formulated today as the achievement (or not) of an optimal

allocation of resources. The characterization of optimality by means of marginal rates of substitution was first completed by Oscar Lange (1942). This characterization, however, is unsatisfactory for several reasons, including the facts that marginal rates of substitution may fail to exist for otherwise unremarkable preference orders, the treatment of corners is complicated, and the corresponding second-order conditions are sufficient only for local optimality. At about the same time on two different American coasts, Kenneth Arrow (1952) and Gerard Debreu (1951) proposed an alternative analysis of the relationship between equilibrium and optimality, making use of convexity assumptions and, in particular, the separating hyperplane theorem instead of the calculus. Debreu (1954b) extended his geometric analysis from finite dimensional vector spaces to linear topological vector spaces, that is, from finite to an infinite number of commodities. This advance is important for such diverse topics as financial markets, uncertainty, dynamic modelling and commodity differentiation. The first half of Debreu (1951) establishes the classical welfare theorems, relying only on convexity and topological assumptions on preferences. The second half of the paper introduces the coefficient of resource utilization, a measure of deadweight loss. Debreu (1954a) applied this measure to the deadweight loss associated with tax-subsidy schemes, a measure that has been implemented empirically by Farrell (1957) and Whalley (1976) to study productive efficiency and the deadweight loss of alternative tax schemes. A comparison of the Debreu coefficient with other measures of deadweight loss, including that of his contemporary M. Boiteux at the École normale supérieure, can be found in Diewert (1981).

The Theory of Value

Debreu's *Theory of Value* (1959) is not simply about the existence and optimality of equilibrium. It is a statement of method that has profoundly changed the way economics is practised. For this alone it is among the most original books of 20th-century economic thought. Most economists identify Debreu with mathematics, manipulating formulas and proving theorems. But for Debreu this, although pleasurable, was the easy part of economic theory. He once told me that it was harder to be an economist than a mathematician. A mathematician had to be correct and elegant; but an economist had to be all that and also interesting. The power of a model lies in the economist's ability to interpret with it, and this is the point of all the 'elegance' and clarity in Debreu's exposition. In the preface, he writes (1959, p. x), 'Allegiance to rigor dictates the axiomatic form of the analysis where the theory, in the strict sense, is logically entirely disconnected from its interpretations. ... Such a dichotomy reveals all the assumptions and the logical structure of the analysis.' Debreu taught that the separation of logical analysis from interpretation is crucial to good theory. The logic of market equilibrium is independent of what commodities actually are, except in so far as what they are may suggest additional structure on the primitives of the equilibrium model. This is most clearly demonstrated in Chapter 7. Here Debreu reinterprets the model by appending to the description of commodities the state of nature in which it is available. The use of Arrow's (1953) contingent commodities 'allows one to obtain a theory of uncertainty free from any probability concept and formally identical with the theory of certainty developed in the preceding chapters' (1959, p. 98). Three pages later, Debreu observes that the convexity assumptions required by the theoretical analysis could be understood as risk aversion. And although Debreu stops here, it is not a big step to observe that natural preference models, like Savage's subjective expected utility model, lead to an additive structure for preferences that may have implications for the nature of equilibrium.

Large Economies and the Core

Competitive equilibrium requires prices, and prices in turn already require a sophisticated set of market institutions. Nonetheless, 'general' is a key word in the phrase general competitive equilibrium. The principle behind the abstract treatment of market equilibrium is that the workings of supply and demand are more or less the same whether the market under discussion is a modern financial market in London or New York or a

village market of farmers and petty traders in India or East Africa. This is quite a claim. Support for this idea comes from the fact that the Walrasian outcome from markets with quoted prices can also be supported by a seemingly more fundamental equilibrium concept that makes no mention of prices at all: the core.

The core comes from F.Y. Edgeworth's *Mathematical Psychics* (1881), in which the contract curve is first introduced, and which, remarkably, undertakes a limit analysis of the economy with two types of traders and two goods. Edgeworth showed that the set of core allocations shrinks to the set of competitive equilibria as the number of agents becomes large. Debreu and Scarf (1963) pick up this question and quickly dispatch it for replica economies, which are generalizations of the large population structures Edgeworth studied. Immediately thereafter came Aumann's (1964) equivalence theorem for the core and equilibrium set of an economy with a continuum of agents, which, among other things, launched the subject of economies described by a measure space of agents. These developments are important because perfect competition is most naturally expressed as a large economy (large number of agents) phenomenon, and because empirical descriptions of large markets may be best described by distributions on the space of agent characteristics.

Smooth Economies

It is often said that Gerard Debreu took the calculus out of economics with his topological equilibrium analysis of the 1950s and early 1960s. If so, it returned with a vengeance in his 1970 and 1972 papers on economies with differentiable excess demand. It has been clear since the Edgeworth box that economies with multiple equilibria are inescapable, a fundamental indeterminacy of the analysis. One can easily construct exchange economies with a continuum of equilibria. But how far does it extend? Is this the norm or are these economies pathological? In a path-breaking series of papers Debreu drew the line between normal and bizarre. He demonstrated that if individual demand is differentiable, then the 'generic' case is one in which there are only a finite number of isolated equilibria; that is, equilibria are locally unique. 'Economies with a Finite Set of Equilibria', his 1970 paper, is particularly striking in its simplicity. Once it is determined that an economy is regular, the main result follows from the inverse function theorem – surely a result known to anyone who has taken a multivariate calculus course. Only the deeper fact that regularity is generic requires more advanced tools such as Sard's theorem. Again, Debreu's intuition was geometric. In lectures this was explained with a simple diagram. Subsequent work has used the tools of differential topology to uncover the deeper structure of the equilibrium manifold, the graph of the equilibrium correspondence. These tools are also of fundamental importance for economies with incomplete markets. With incomplete markets and financial assets rather than real assets, indeterminacy is no longer unusual, and this is of critical importance for applications to macroeconomics and finance. Some of this work is surveyed in the monographs of Balasko (1988) and Mas-Colell (1985).

Excess Demand

It is important to ask of any theory, 'what can it say?' That is, what kinds of predictions will the theory make, and what patterns in data will contradict the theory? In general equilibrium theory this question was first asked by Sonnenschein (1972) in the following way: in exchange economies, the market excess demand function satisfies the restrictions of continuity, homogeneity and Walras's Law. This and a boundary condition is enough to prove the existence of equilibrium prices. Sonnenschein asked if excess demand functions had any additional structure beyond these three requirements. Sonnenschein (1972), Mantel (1974) and Debreu (1974), with an important extension by (Mas-Colell 1977), showed that the answer is 'no'. Any function defined for strictly positive prices and satisfying these three conditions is identical up to boundary behaviour with an excess demand function for an exchange economy containing no more agents than goods, each agent with continuous, strictly convex and monotonic preferences. Thus the hypothesis of utility maximization in exchange

economies, with no additional assumptions about agents' characteristics, will place few restrictions on comparative static results or on the nature of the equilibrium price set.

These results are often incorrectly interpreted to mean that general equilibrium theory is empty, that it predicts nothing. This is entirely incorrect. General equilibrium theory is not so much a theory as a theoretical framework within which theories can be built by making explicit assumptions about the nature of tastes, technologies and endowments. To say that the framework does not limit market behaviour without any assumptions about its primitive objects is to say that the framework is maximally expressive. Its power to predict market behaviour comes from assumptions about the population of agents participating in the market. The so-called 'anything goes' theorems simply imply that more results will require more assumptions about the preferences and endowments of agents. It had been Debreu's hope that restrictions on the distributions of agents' characteristics would lead to interesting conclusions: but progress has been slow.

Other Contributions

Debreu has produced seminal papers in areas of economic theory other than general equilibrium analysis. Which preference orders have a continuous utility representation? This question is answered by (1954c). Which preferences have additive separable representations? Debreu's (1958) answer to this very difficult question is topological in nature, and quite distinct from the algebraic answers found in the mathematical psychology literature.

Debreu was exceptional in the classroom and in seminar. His lectures were crystalline, elegantly shaped, and parsimonious. Often they were too clear; we students left the class convinced we understood, only to discover on problem sets how subtle were the arguments that had seemed so obvious on the blackboard. Debreu's expository writings, especially his Nobel Address (1984), are required for everyone with a serious interest in contemporary economics.

Conclusion

It is impossible to imagine modern economics without the scholarship of Gerard Debreu. Debreu, Kenneth Arrow and a few others who solved the big open questions of general equilibrium theory in the 1950s had an impact that reached far beyond the confines of formal competitive analysis. They were responsible for making formal modelling a requirement for serious economic analysis of any kind. Formal modelling is not merely a theoretical discourse; the availability of formal models requires a means for the models to confront data. Modern econometrics is inconceivable without the idea of formal modelling as a strategy of enquiry. It is not by accident that, just as the general equilibrium theory was taking off at the Cowles Commission in the 1950s, so too was modern econometrics. The contributions of the 'mathematical economists' launched a revolution that has touched on every area of economic practice.

See Also

- ▶ Core Convergence
- ▶ Cores
- ▶ General Equilibrium
- ▶ Welfare Economics

Selected Works

1951. The coefficient of resource utilization. *Econometrica* 19: 273–292.

1952. A social equilibrium existence theorem. *Proceedings of the National Academy of Sciences* 38: 597–607.

1954a. (With Arrow, K.J.) Existence of an equilibrium for a competitive economy. *Econometrica* 22: 265–290.

1954b. Valuation equilibrium and Pareto optimum. *Proceedings of the National Academy of Sciences* 40: 584–592.

1954c. A classical tax-subsidy problem. *Econometrica* 22: 14–22.

1954d. Representation of a preference ordering by a numerical function. In *Decision*

processes, ed. R.M. Thrall, C.H. Coombs, and R.L. Davis. New York: Wiley.
1958. Stochastic choice and cardinal utility. *Econometrica* 26: 440–444.
1959. *Theory of value*. New York: Wiley. Repr. New Haven: Yale University Press, 1971.
1963. (With Scarf, H.) A limit theorem on the core of an economy. *International Economic Review* 4: 235–246.
1970. Economies with a finite set of equilibria. *Econometrica* 38: 387–392.
1972. Smooth preferences. *Econometrica* 40: 603–615.
1974. Excess demand functions. *Journal of Mathematical Economics* 1: 15–21.
1984. Economic theory in the mathematical mode. *American Economic Review* 74: 267–278.

Bibliography

Arrow, K.J. 1952. An extension of the basic theorems of classical welfare economics. In *Proceedings of the second berkeley symposium on mathematical statistics and probability*, ed. J. Neyman. Berkeley: University of California Press.
Arrow, K.J. 1953. Le rôle des valeurs boursières pour la répartition la meilleure des risques. *Econométrie* 40: 41–48. Cahiers du CNRS.
Aumann, R.J. 1964. Markets with a continuum of traders. *Econometrica* 32: 39–50.
Balasko, Y. 1988. *Foundations of the theory of general equilibrium*. Boston: Academic.
Diewert, W.E. 1981. The measurement of deadweight loss revisited. *Econometrica* 49: 1225–1244.
Edgeworth, F.Y. 1881. *Mathematical psychics: An essay on the application of mathematics to the moral sciences*. London: C. Kegan Paul.
Farrell, M.J. 1957. The measurement of productive efficiency. *Journal of the Royal Statistical Society* 120: 253–291.
Lange, O. 1942. The foundations of welfare economics. *Econometrica* 10: 215–228.
Mantel, R. 1974. On the characterization of aggregate excess demand. *Journal of Economic Theory* 7: 348–353.
Mas-Colell, A. 1977. On the equilibrium price set of an exchange economy. *Journal of Mathematical Economics* 4: 117–126.
Mas-Colell, A. 1985. *The theory of general economic equilibrium: A differentiable approach*. Cambridge: Cambridge University Press.
McKenzie, L. 1954. On equilibrium in Graham's model of world trade and other competitive systems. *Econometrica* 22: 147–161.
Sonnenschein, H. 1972. Market excess demand functions. *Econometrica* 40: 549–563.
Walras, M.E.L. 1974. *Éléments d'économie politique pure, ou théorie de la richesse sociale*. Lausanne: Corbaz.
Whalley, J. 1976. Some general equilibrium analysis applied to fiscal harmonization in the European community. *European Economic Review* 8: 290–312.

Debt Mutualisation in the Ongoing Eurozone Crisis – A Tale of the 'North' and the 'South'

Ansgar Belke

This article builds upon a highly stylised but widespread definition of the 'Southern' and 'Northern' views on debt mutualisation. It explains both positions in the ongoing Eurozone crisis and what both sides hope to achieve in reshaping the governance of the euro. Both sides agree on many things, such as the current threat to the survival of the euro. But the 'South' sees the main threat to the Eurozone as coming from the fear and panic that can suddenly increase borrowing costs and push countries into insolvency. The 'North', on the contrary, reckons that the principal menace stems from removing this market pressure too quickly, dampening the need to reform. Both speak of the political backlash that the crisis creates. For the 'South' it is excessive austerity in debtor nations that should be resisted; for the 'North' it is excessive liabilities in creditor states that can cause resentment. The article concludes that the debate about mutualisation of debt is not just about the future of monetary union, but also about the political future of the European Union. Any successful deal must come up with a recipe of how to (re-)create trust between European citizens and their elected governments.

Introduction

The European summit that ended on 29 June 2012 declared that it was 'imperative to break the

vicious circle between banks and sovereigns'. Markets revived in the hope that the political leaders were finally ready to act to deal with the threat to the euro, and then soon lost heart amid the cacophony of rival interpretations about what had been agreed. Still, the leaders had identified the right issue: weak banks and weak sovereigns are like two bad swimmers that are pulling each other under water (Pisany-Ferry 2012).

But which one should be saved first? Proponents of the 'Southern view', like, for instance, Paul de Grauwe (2012) say we should start with the sovereigns, by throwing them the lifejacket of joint-issued debt. In effect, richer countries would guarantee at least part of the debt of weaker ones.

Representatives of the 'Northern', and let's say especially the 'German' view, reckon instead that it is better to start by saving the banks. This would be done through stronger central supervision and the mutualisation of some liabilities in the banking sector, for instance through a joint fund to wind up failing banks and provide a Europe-wide guarantee of bank deposits. In effect depositors in solid banks would be guaranteeing the savings of those in more fragile ones.

This article builds upon a highly stylised but widespread definition of the 'Southern' and 'Northern' views. The former is usually held by countries like Greece, Italy, Portugal and Spain and, since François Hollande has taken office, also France. The latter is often used synonymous with the 'German' view and also includes countries like Austria, Finland and the Netherlands and, for some periods under French President Nicolas Sarkozy, also France (Merler and Pisany-Ferry 2012; see also a recent statement by the French Minister of Finance who points to the need for common debt instruments: http://www.reuters.com/article/2012/10/30/Eurozonefrance-germany-idUSL5E8LU44020121030). Since the exact characteristics of both views may still remain unclear, the remainder of this article examines them more deeply.

Both sides – the 'North' and the 'South' – agree on many things, such as the current threat to the survival of the euro. They both recognise the danger that debt mutualisation could bring moral hazard and higher costs for creditor countries. For representatives of the 'Northern view' there is no getting around these problems. For the 'South', though, these risks can be removed, or at least mitigated through careful design of the system. For instance, the Eurozone could impose conditions on countries seeking the benefit of jointly issued debt.

The 'South' sees the main threat to the Eurozone as coming from the fear and panic that can suddenly increase borrowing costs and push countries into insolvency (Pisany-Ferry 2012). The 'North', on the contrary, reckons that the principal menace stems from removing this market pressure too quickly, dampening the need to reform (Sinn and Wollmershäuser 2012).

Both speak of the political backlash that the crisis creates. For the 'South' it is *excessive austerity* in debtor nations that should be resisted; for the 'North' it is *excessive liabilities* in creditor states that can cause resentment. In some ways, though, they are not so far apart. The 'North' concedes that it is necessary to have some mutualisation of debt, if only to recapitalise banks (Belke 2012a). The 'South' accepts that debt mutualisation must be limited to avoid moral hazard (de Grauwe 2012).

Contrasting the 'Southern' and 'Northern' Views

In the following, the basic ingredients of the 'Southern' and the 'Northern' view are contrasted.

The 'Southern' View: Some Basics

The main argument of the 'South' runs as follows: since the 1970s economists have warned that a budgetary union would be a necessity for a sustainable monetary union. But the founders of the Eurozone had no ears for this warning. It is now patently clear that they were mistaken and that the governments of the euro area member countries face the following hard choice today: either they fix this design failure and move to a budgetary union; or they do not fix it, which means that the euro will have to be abandoned (Pisany-Ferry 2012). Although analysts such as Paul de Grauwe were sceptical about the desirability of a monetary

union during the 1990s (contrary to Gros and Thygesen 1998), the same author now takes the view that we cannot properly manage a deconstruction of the Eurozone (de Grauwe 2012). A disintegration of the Eurozone would produce huge economic, social and political upheavals in Europe. If the euro area governments want to avoid these, they have to look for strategies that move us closer towards a budgetary union.

A budgetary union, such as the US one, appears to be so far off that there is no reasonable prospect of achieving this in the Eurozone 'during our lifetimes' (Henning and Kessler 2012). Does that imply that the idea of establishing a budgetary union and thus a 'genuine' EMU is a pure chimera? De Grauwe (2012) argues that this drastic assessment is not at all valid and that there is a strategy of taking small steps that lead us in the right direction. But before this strategy can be outlined it is – according to the 'Southern' view – important to understand one of the main design failures of the Eurozone, which will inform the debate about what exactly has to be fixed.

The 'Southern' argument starts with the basic insight that Eurozone governments issue debt in euros, which is a currency they cannot control. As a result, and in contrast to 'standalone' countries like the UK, they endow bondholders with a guarantee that the cash to pay them at maturity will always be available (Belke and Burghof 2010).

The fact that governments of the Eurozone are not able to deliver such a guarantee to bondholders makes them vulnerable to upsurges of distrust and fear in the bond markets. This can trigger liquidity crises that in a self-fulfilling way can drive countries towards default, forcing them to apply austerity programmes that lead to deep recessions and ultimately also to banking crises (Claessens et al. 2012; de Grauwe 2011, 2012). This is not to say that countries that have overspent in the past do not have to apply austerity – they will have to (Pisany-Ferry 2012). It is rather that financial markets, when they are driven by panic, force austerity on these countries with an intensity that can trigger major social and political backlashes that policymakers may not be able to control. The effects are there to see in a number of Southern European countries (de Grauwe 2011, 2012; Freedman et al. 2009): namely Greece, Italy, Spain and Portugal.

Their previous diagnosis of a design failure of the Eurozone leads proponents of the 'Southern' view to the idea that some form of pooling of government debt is necessary to overcome this failure (Pisany-Ferry 2012). By *pooling government debt*, the weakest in the union are shielded from the destructive upsurges of fear and panic that regularly arise in the financial markets of a monetary union and that can hit any country. 'Those that are strong today may become weak tomorrow, and vice versa' (de Grauwe 2012).

Representatives of the 'South' see the 'moral hazard' risk that those that profit from the creditworthiness of the strong countries will exploit this and lessen their efforts to reduce debts and deficits. This moral hazard risk is the main obstacle to pooling debt in the Eurozone. The second obstacle is that inevitably the strongest countries will pay a higher interest rate on their debts as they become jointly liable for the debts of governments with lower creditworthiness. Thus debt pooling must be designed in such a way as to overcome these obstacles (Claessens et al. 2012; Pisany-Ferry 2012).

Moderate proponents of the 'Southern' view agree, apparently in line with theMerkel government that there are *three principles* that should be followed in designing the right type of debt pooling (Claessens et al. 2012; de Grauwe 2012; Pisany-Ferry 2012). First, it should be *partial* – that is, a significant part of the debt must remain the responsibility of the national governments, so as to give them an ongoing incentive to reduce debts and deficits. Several proposals have been made to achieve this (among them Delpla and Weizsäcker 2011, and German Council of Economic Advisors 2012). Second, an *internal transfer mechanism* between the members of the pool must ensure that the less creditworthy countries compensate (at least partially) the more creditworthy ones (de Grauwe 2012). Third, a tight *control mechanism* on the progress of national governments in achieving sustainable debt levels must be an essential part of debt pooling. The Padoa-Schioppa group has recently proposed a gradual loss of control over

their national budgetary process for the breakers of budgetary rules (Padoa-Schioppa Group 2012).

Proponents of the 'Southern' view acknowledge that the Eurozone is in the midst of an existential crisis that is slowly but inexorably destroying its foundations. They immediately conclude that the only way to stop this is to convince the financial markets that the Eurozone is here to stay (de Grauwe 2012; Pisany-Ferry 2012). Their main argument is that debt pooling, which satisfies the principles outlined above, would give a signal to the markets that the members of the Eurozone are serious in their intention to stick together. Without this signal, the markets will not calm down and an end to the euro is inevitable (Aizenman 2012; de Grauwe 2012). In the words of the German Chancellor Angela Merkel: these policies are *without alternative*.

Materially, the 'Northern' view sketched in the following represents the accumulation of a multitude of reactions of the 'North' to these much more activist 'Southern' proposals of several kinds of debt mutualisation which have been frequently put forward since the onset of the euro crisis (Claessens et al. 2012).

The 'Northern' View: Important Facets

One of the main priorities of the 'Northern' view is that the mutualisation of the Eurozone's debt to bring about the convergence of interest rates, as proposed within building block 2 of the Interim Report, will not in the long run tackle the root of the problems. Instead it has the potential to sow the seeds of an even larger crisis in the future (Sinn and Wollmershäuser 2012; Weidmann 2012). They allude to what happened in the early years of the euro, when interest rates largely converged. Paradoxically, perhaps, this paved the way for a greater divergence of national fiscal policies. A reckless lack of discipline in countries such as Greece and Portugal – be they more (Greece) or less (Portugal) insolvent – was matched by the build-up of asset bubbles in other member countries, such as Spain and Ireland, deemed merely illiquid. Structural reforms were delayed, while wages outstripped productivity growth. The representatives of the 'Northern' view stress that the consequence was a huge loss of competitiveness in the periphery, which will by definition not be resolved by the mutualisation of debt (Belke 2012a).

Debt mutualisation can take different forms (Aizenman 2012). One is to mutualise newsovereign debt through Eurobonds (Delpla and von Weizsaecker 2010, develop one ofmore than seven variants; Pisany-Ferry 2012). Another is to merge part of the old debt, as advocated by the German Council of Economic Advisors (2012), with its proposal for a partly gold-backed European Redemption Fund (Belke 2012b). A third means is to activate the Eurozone's 'firewall' by using the rescue funds (either the temporary European Financial Stability Facility or the permanent European Stability Mechanism) to buy sovereign bonds on the secondary (or even primary) market, or to inject capital directly into distressed banks. Indeed, the ECB is already engaged in a hidden form of mutualisation – of risk if not (yet) of actual debt – through its programmes of sovereign bond purchases (the Securities Market Programme, SMP, and the announced conditional Outright Monetary Transactions,OMTs) and its long-term refinancing operations for banks.

The view of the 'North' is that almost all these are bound to fail, either for economic or political reasons, or both. The governments of even financially strong countries cannot agree to open-ended commitments that could endanger their own financial stability or, given that they are the main guarantors, of the bailout funds. And the danger of moral hazard is ever-present (Belke 2012a).

Proponents of the 'Northern' view point to the fact that any form of mutualisation involves an element of subsidy, which severely weakens fiscal discipline: the interest rate premium on bonds of fiscally weaker countries declines and the premium for stronger countries increases. Fiscally solid countries are punished and less solid ones, in turn, are rewarded for their lack of fiscal discipline and excess private and public consumption.

If yields are too low there is no incentive for private investors to buy sovereign bonds. The countries risk becoming decoupled from the capital markets permanently and the debt problems become increasingly structural (Belke 2012b).

This is true also for the ECB's bond-buying announcements and activities. The credit risk is thus just rolled over from the bonds of the weaker countries to those of the stronger ones (depending on the buyback price), and the ECB is made responsible for its liability. Over time, the ECB's measures might even be inflationary. Having the rescue funds buy bonds is little different, except that they lack the lending capacity to be credible. If they are given a banking licence, as demanded by the 'South' (for instance, by French President Hollande) it would be no different from having the ECB buy bonds directly (Belke 2012b).

What about the European Redemption Fund (ERP) from the 'Northern' perspective? This type of fund could be of particular help to Italy, which could unload half of its debt. But its partners could not force Italy to tax its citizens to ensure that it pays back the dormant debt. And with the assumption of debt, the credit rating of Germany might drop, due to the increase of the German interest burden. The pressure on Italy and Spain to consolidate their budgets sustainably would be reduced. The problems of Greece, Ireland and Portugal would not be resolved, since these countries are unlikely to qualify for the ERP.

In addition to moral hazard, there are political obstacles, which would be most acute in the case of Eurobonds. Germany demands political union before Eurobonds can be considered. But this is sometimes said to put the cart before the horse: a political union would be created simply to justify Eurobonds (Gros 2011). Advocates from the Merkel government, like Finance Minister Wolfgang Schäuble, say treaty changes and high-level political agreements would be sufficient to make sure that euro area member countries comply with all decisions taken at the euro area level. This became clear when Wolfgang Schäuble came up with a plan a plan drawn to bolster the power of the EU's economic and monetary affairs commissioner (*Daily Telegraph* 2012). Even Mario Draghi, President of the European Central Bank, has supported this German scheme to allow the EU to intervene in countries' budgets and propose changes before they are agreed in parliaments. But the experience with Greece's adjustment casts severe doubt on the optimism expressed by such a proposal.

Even a quick glance at the World Bank's databank of 'governance indicators' shows that differences between Eurozone members, on everything from respect for the rule of law to administrative capacity, are so great that political union is unlikely to work, at least in the next couple of years. It follows from the perspective of the 'North' that the basis for Eurobonds is extremely thin.

According to the 'Northern' or 'German' view, the introduction of Eurobonds would in principle have to be backed by tight oversight of national fiscal and economic policies. But this view neglects that there is no true enforcement as long as the individual Eurozone members remain sovereign.

Intervening directly in the fiscal sovereignty of member states would require a functioning pan-European democratic legitimacy (Claessens et al. 2012), but we are far from that. Voters in Southern countries can reject the strong conditionality demanded by Brussels at any time, while those of Northern countries can refuse to keep paying for the south. And either can choose to exit the Eurozone (Gros 2011).

The emphasis on pushing through a fiscal union as a precondition for debt mutualisation means the debate, at least in Germany, has become a question of 'all or nothing': either deeper political union or *deep chaos* (Belke 2012a; de Grauwe 2011). This unnecessarily narrows the strategic options for the players and causes the permanent 'North—South' divide described in this section, which is severely hampering the realisation of a 'genuine' monetary and economic union (President of the European Council 2012).

However, I argue that there is in fact an alternative option to the notion of cooperative fiscal federalism involving bailouts and debt mutualisation: competition-based fiscal federalism, of the sort successfully operating in the USA, Canada and Switzerland, among others. These countries have largely avoided serious and sustained public debt in their component states. The sub-federal entities, faced with insolvency, have a great incentive to take early corrective action – without having to force the member states into a corset of centralised

fiscal policy coordination (von Hagen 1993). This approach seems to be a good compromise between the 'Southern' and 'Northern' views.

To achieve this sort of federalism, it is necessary to separate the fate of the banks from that of the sovereigns. What is needed is not a fiscal union *in first instance*, but a banking union. It should be based on four elements: a European bank with far-reaching powers to intervene; reformed banking regulation with significantly higher equity capital standards; a banking resolution fund; and a European deposit insurance scheme. At least the first ingredients have also been recognised and acknowledged by the Merkel government.

A banking union – a less comprehensive, more clearly delineated and rather technical task – should be much more acceptable for the 'North' than the Europeanisation of fiscal policy as a whole. This is exactly because it touches upon only a small fraction of the fiscal policy areas which have to be subordinated to central control in a fiscal union.

Obviously, a central resolution authority has to be endowed with the resources to wind up large cross-border banks. Where does the money for this come from? In the long run, the existence of a resolution authority goes along with a deposit insurance scheme for cross-border banks. This should be – according to the 'German' view – funded partly by the banking industry. But there should also be a backstop by the euro area governments provided through the EFSF or the ESM in order to cope with situations of systemic bank failure (Gros and Schoenmaker 2012).

As a temporary transition measure, however, limited debt mutualisation may then be necessary – but only to recapitalise banks that cannot be sustained by their sovereigns. The amounts required are much smaller than for, say Eurobonds (Gros and Schoenmaker 2012).

With the banking system and the debt crisis thus disentangled, banking sector losses will no longer threaten to destroy the solvency of solid sovereigns such as Ireland and Spain. Eurobonds will then not be needed, and neither will the bail-out of sovereigns. The debt of over-indebted states could be restructured, which means that the capital market could exert stronger discipline on borrowers (Belke 2012a).

There are at least two questions left which have yet to be covered in this article and which will be answered in the next sections. If the banking sector is really to be stabilised, a solution will surely have to deal with the devalued sovereign debt that some are holding. Would the banks not be better off holding at least some Eurobonds instead of, say, Greek or Spanish bonds? That said, 'Southern' economists who advocate Eurobonds need to find a way of making them politically acceptable. And how much political union is feasible, or even desirable, just for the sake of a single currency that many never loved? And also, where does the burden end up?

Rebuttal – Banking Union and Other Issues

For 'Northern' governments like the German one, mutualisation of debt is just another form of subsidy and bail-out that the markets clamour for, be it the overt help given to Greece or the more discreet liquidity provided by the European Central Bank.

The fact that there is a loud chorus demanding subsidies does not, in Germany's view, make it right (Belke 2012a). The Merkel government argues that assistance does not help countries make the necessary macroeconomic adjustment in either public or private borrowing. Safeguards and conditions as standalone measures will not work. Anything that puts off the rebalancing of the current account deficit only builds up the forces for the disintegration of the Eurozone. Watching the 'South' borrow and spend themselves into bankruptcy and then bailing them out is called both immoral and irresponsible.

In their rebuttal, 'Southern' governments target what they regard as the contradiction in the 'North's' position, rejecting debt mutualisation while supporting a joint Eurozone backstop for the banking sector (de Grauwe 2012). Are banks any more trustworthy than sovereigns?

The 'South' usually argues, moreover, that mutualisation of banking liabilities will inevitably be followed by the pooling of debt. Banking union on its own, for instance, de Grauwe (2012) notes,

would protect the sovereigns from banking crises. But it would not protect banks from sovereign debt crises. If banking union must be followed by the fiscal sort, it would be best to do it at the same time, the 'South' argues.

Many questions remain unresolved. Some German politicians identify the tendency of the single currency to push the economies of its members apart (Belke 2012a). If each country is to fend for itself, as some proponents of the 'German' view say, would they not be better off restoring their own national currencies so that macroeconomic adjustment can take place more painlessly? As a blogger in The Economist Online put it, 'The south will end up having to leave the euro to save what's left of its economy'. (https://www.economist.com/users/turbatothomas/comments?page=1).

Closing – Debt Mutualisation Versus Fiscal Federalism

'South': A Monetary Union Cannot Last Without Debt Mutualisation to Avoid Deflation

The key issue is this: can a monetary union last without some form of fiscal union? Economists have been debating this issue for decades. It seems, at least to the 'South', that the consensus among them is that a monetary union without some form of fiscal union will not last.

What kind of fiscal union is necessary to sustain a monetary union? 'Southern' governments tend to argue that such a fiscal union must have two components. First, it must have some insurance component, i.e. there must be some transfer mechanism from regions (countries) that experience good economic times to regions (countries) that experience bad times. (The Interim Report alternatively proposes a central budget with a similar function; see President of the European Council 2012). According to the 'South', the USA is often seen as a successful monetary union, partly because the federal government's budget performs this role of insurance (Henning and Kessler 2012). Also 'Southern' governments are eager to point out that the opponents will not cease to stress that such an insurance mechanism creates moral hazard issues. But that is the case with all insurance mechanisms. Representatives of the 'Southern' view argue as an analogy that one generally also does not conclude that people should not have fire insurance because such insurance creates moral hazard, i.e. it will lead to more fires.

The second component of a fiscal union is some degree of debt pooling. Economists defending the 'Southern' view have argued that this is necessary because, in becoming members of a monetary union, countries have to issue debt in a 'foreign' currency and therefore become more vulnerable to upsurges of distrust and fear in financial markets. These can in a self-fulfilling way push countries into a bad equilibrium that makes it more difficult for them to adjust to imbalances (de Grauwe 2012). Of course, debt pooling does not solve these fundamental problems (as 'Northern' governments suggest that the 'South' believes), but it avoids pushing countries, like Spain today, into a deflationary spiral that makes their debt problems worse, not better.

Thus monetary union and fiscal union (including some degree of debt mutualisation) are the opposite sides of the same coin. As has become clear in the previous sections, the proponents of the 'Northern' view like to refer to history. The 'Southern' economists do this also. According to them, there are no successful monetary unions that are not embedded in a fiscal union that includes debt mutualisation.

Some economists, especially in Northern Europe, continue to argue that one can have a monetary union without a fiscal union. Paul de Grauwe (2012), for instance, reduces the 'Northern' view to something like 'all we need is discipline (a fiscal compact?), including a credible no-bail-out clause. If we allow governments to default, financial markets will do their work in disciplining these governments'. According to the 'South' and the Interim Report by the President of the European Council (2012) as well, this view can certainly not be taken seriously any more (de Grauwe 2012). This is because financial markets are entirely incapable of applying the right discipline on governments. When markets are

euphoric, as they were during the 10 years before the crisis, they intensify indiscipline by giving incentives to borrowers and lenders alike to create excessive debt and credit. Since the crisis erupted, financial markets have been in a continuous state of fear and panic, leading them to apply excessive discipline that has improved nothing and could not prevent increasing debt burdens (de Grauwe 2012).

When this debate will have been settled it will – according to the 'Southern' view – be clear that the greatest obstacle to debt mutualisation and to the continuing existence of the Eurozone is a lack of trust (Belke 2012c; de Grauwe 2011). Northern European countries distrust southern European countries and have propagated the myth that the North is morally superior compared with the corrupt regimes in the South. In Northern mythology, southern European countries are seen as completely incapable of setting their house in order. Lending money to these countries is pouring the 'hard-earned money of virtuous German savers' into a bottomless pit.

'North': Towards a Concept of Competition-Based Fiscal Federalism in the Eurozone

The most important components of a competition-based fiscal federalism that would make Eurobonds unnecessary were set out earlier in this article. This is not because banking union is equivalent to Eurobonds (as claimed by de Grauwe 2012) but because it would disentangle *a banking and a sovereign-debt crisis*. With a solid banking system in place, banking sector losses would no longer threaten the solvency of solid sovereigns (such as Ireland and Spain), and the bail-out of less reliable sovereigns would no longer be necessary. That means there would be a lower chance that fundamentally sound sovereigns would suffer from a confidence crisis and rocketing risk premiums.

Proponents of the 'Northern' view do not accept the argument of the 'South', coined for instance by de Grauwe (2012), that a banking union does not protect the banks from sovereign failures. In a banking union, the capital market could exert its disciplining influence more effectively than it does now. Debt restructuring for insolvent states would become more probable. The debtor state would lose its strongest asset (the claim that default would cause huge damage to the entire financial system) and creditors could not rely on taxpayers to get their money back. This, in turn, would put governments with unsound finances under pressure to curb their deficits.

Instead they hint at a wide array of econometric studies showing a systematic relationship of sovereign bond yields and the anticipated sustainability of a country's public debt – at least in the medium term. They leave it to the Banca d'Italia's research department to come up with convertibility risk (measured by google-nomics, counting google searches for 'euro area breakup') as an explanatory variable of Southern sovereign bond yield spreads over the German one (Di Cesare et al. 2012). Only recently, the spread on Spanish bonds moved up after Mariano Rajoy, the Spanish prime minister, announced that he intended to relax Spain's deficit-adjustment path; the same was true when Italy decelerated its pace of reforms. Hence proponents of the 'Northern' view can sleep quite well with the idea that 'capital markets will take care of the rest'.

To get rid of the fragility of the banking system, we need to establish a temporary or even permanent European Resolution Authority (ERA), whose task would be to sort out fragile banks across Europe, regardless of size. Weaker banks would receive a one-time injection of capital or be wound down, wholly or partly. This body should have the power to turn bank debt into equity capital. Creditors of ailing banks – but not the taxpayers, as de Grauwe (2012) assumes – should as far as possible be made liable for their risky investments. In contrast with Eurobonds, which tend to cover a lot of bad risks, a European deposit scheme based on funding from the banks themselves (in order to avoid the taxpayers bearing the risk) would in the end embrace only stronger banks (Gros and Schoenmaker 2012).

The 'North' admits to the 'South' that is right to argue that the lack of a budgetary union, akin to the American system, is a design failure of the Eurozone. Proponents of the 'Northern' view also

strongly support the 'South's' view that a proper application of the American system would prevent a costly disintegration – but most probably for different reasons. Since the US system prevents central bank loans from being more attractive than market loans, it avoids permanent balance-of-payment imbalances between member states. In America, neither the individual state nor the private sector has access to the printing press to finance itself and can also default. If the inhabitants of a state need to finance their current account deficits, they have to offer attractive interest rates and provide sufficient collateral to private lenders from other American states (Henning and Kessler 2012; Belke 2012a).

Yet the 'South' argues, essentially, that the main problem of Eurozone countries is that they do not have direct access to the printing press (de Grauwe 2012). According to the 'North', it is thus following the strange behaviour of rating agencies, which penalise members of the Eurozone simply for being part of the single currency. For too long the agencies rated countries too generously, pricing in a potential bail-out rather than basing ratings purely on macroeconomic fundamentals. This pattern made possible riskless profits from riskless speculation against sometimes hopelessly noncompetitive member states. The 'South' reinterprets this as a question of 'panicked financial markets' in its mother of all arguments for debt pooling (de Grauwe 2012).

Especially according to the 'Northern' view, the members of the Eurozone are intentionally kept away from the ECB to avoid them activating the inflation tax to finance themselves. The scope for an individual country to incur government debt is simply lower within a currency union than outside. This scope cannot be extended through debt pooling without risking the disintegration of the Eurozone (Belke 2012a).

But the 'Northern' view contains a lot more. As a rule, the burden on bank balance sheets should be borne by the country of domicile and not – as in the case of Eurobonds – be passed on to other countries. However, it is not clear whether and to what extent over-indebted countries will be capable of doing this. Using the rescue funds would make sense as a fiscal backstop. Subject to negotiation, a temporary debt mutualisation to cover the cost of bank recapitalisation would make sense, to avoid a larger and permanent mutualisation of sovereign debt. But only after a proper pan-European banking oversight has been worked out and implemented (Belke 2012a).

Remarkably, the "South" on some occasions outlines hard budget constraints to accompany debt pooling (Pisany-Ferry 2012; de Grauwe 2012). Its representatives propose binding mechanisms of compensating the more creditworthy countries and controlling the behaviour of those that are less so. But, according to the "Northern" view, historical experience gives reason to doubt that this will work – for several reasons.

One is that, for instance, Spanish foreign debt is currently among the greatest risks for the euro zone, and it is essentially private. As long as the private sector has access to the ECB system at interest rates that are below the market rate, the correction of external imbalances through real internal devaluations will not take place or if it does, at least not in sufficient quantities. The "South's" approach would require not only public debt limits but also private debt barriers to bring about such a correction, the "North" claims, but that would be an absurd endeavour.

According to the "North", the "South" should draw some lessons from the current conduct of monetary policy. The latter already uses debt pooling, of a sort. The quality of the collateral that the ECB accepts varies considerably from country to country. In the case of the ECB's lending to Greek banks, it consists of doubtful private Greek assets and Greek government debt whose value depends on election results, as has been recently observed. Thus the ECB acts as a central counterparty for crossborder lending which incurs risks along national lines (Gros et al. 2012). Risk mutualisation could well, if things go wrong, turn into full debt mutualisation, and lead to conflicts between member states. It provides an advance warning of how debt pooling could lead to the disintegration of the eurozone.

Conclusion – The Pre-eminent Role of Trust

Throughout the Eurozone's debt crisis, many Europeans have looked across the Atlantic for lessons on how to run a successful monetary union. The European Commission boasts that, taken together, the Eurozone's fiscal deficit and debt are lower than America's. Yet the euro faces an existential crisis while the dollar, despite the troubles of the American economy, still remains a shelter.

So, how much banking and fiscal integration does the Eurozone need in order to restore stability? And how much political unity does it need to maintain checks and balances, and democratic legitimacy? Looking at the USA, 'Southern' and 'Northern' economists and politicians more or less agree on the need for some kind of federalised system to recapitalise, restructure or wind down ailing banks. That is where the 'North' thinks integration should stop – in contrast to the Interim Report (President of the EU Council 2012). The key lesson from the USA is, in its view, that it pays to enhance market discipline on the states: as long as the banking system is stabilised at minimal cost to the taxpayer, over-indebted states can be allowed to go bust (Henning and Kessler 2012). But proponents of the 'South' think that this deals with only half of the vicious circle between weak banks and weak sovereigns. So it cannot work in the long run. What makes America and other monetary unions stable is a system of joint bonds and other forms of mutual insurance, and internal transfers to redress economic imbalances (de Grauwe 2012).

So the debate about mutualisation of debt is not just about the future of monetary union, but also about the political future of the European Union. Leaders usually try to avoid such questions about the end point, known as the *finalité politique*. Any successful deal must come up with a recipe of how to (re-)create trust between European citizens and their elected governments.

See Also

- ▶ Credit Crunch Chronology: April 2007–September 2009
- ▶ European Monetary Integration
- ▶ European Monetary Union
- ▶ European Union Budget
- ▶ Euro Zone Crisis 2010
- ▶ European Union (EU) Trade Policy
- ▶ Greek Crisis in Perspective: Origins, Effects and Ways-Out

Acknowledgments I am grateful for valuable comments to Alison Howson and Iain Begg.

Bibliography

Aizenman, J. 2012. *The euro and the global crisis: Finding the balance between short term stabilization and forward looking reforms*, NBER working paper 18138. Cambridge, MA: National Bureau of Economic Research.

Belke, A. 2012a. Euro debt: Should the Eurozone's debt be mutualised? Economist Debates, *The Economist Online*, 11–23 July. Available at: http://www.economist.com/debate/days/view/856%20to%20858/print. Accessed 10 Feb 2013.

Belke, A. 2012b. *A more effective Eurozone monetary policy tool – Gold-backed Sovereign Debt*. Briefing paper prepared for presentation at the Committee on Economic and Monetary Affairs of the European Parliament for the quarterly dialogue with the President of the European Central Bank, September, Brussels.

Belke, A. 2012c. *Towards a genuine economic and monetary union – Comments on a roadmap*. Briefing paper prepared for presentation at the Committee on Economic and Monetary Affairs of the European Parliament for the quarterly dialogue with the President of the European Central Bank, September, December, Brussels.

Belke, A. and H.-P. Burghof. 2010. Stand-alone ratings for countries – Remedial action in case of market failure. *EuroIntelligence*, 24 November. Available at: http://www.eurointelligence.com/eurointelligence-news/archive/single-view/article/stand-alone-ratings-for-countries-remedial-action-in-case-of-market-failure.html. Accessed 10 Feb 2013.

Claessens, S., A. Mody, and S. Valleé. 2012. *Paths to eurobonds*, Bruegel working paper 2012/10. Brussels: Bruegel.

Daily Telegraph. 2012. *Germany shocks EU with fiscal overlord demand*, 16 October. Available at: http://www.telegraph.co.uk/finance/financialcrisis/9613384/German

D

y-shocks-EU-with-fiscal-overlord-demand.html. Accessed 9 Feb 2013. Debt mutualisation in the ongoing Eurozone crisis 11.
De Grauwe, P. 2011. *A less punishing, more forgiving approach to the debt crisis in the Eurozone*, CEPS policy brief 230. Brussels: Centre for European Policy Studies.
De Grauwe, P. 2012. Euro debt: should the Eurozone's debt be mutualised? Economist debates, *The Economist Online*, 11–23 July. Available at: http://www.economist.com/debate/days/view/856%20to%20858/print. Accessed 10 Feb.
Delpla, J. and J. von Weizsäcker. 2011. *The blue bond proposal*. Breugel.org, 11 May. Available at: http://www.bruegel.org/download/parent/403-the-blue-bond-proposal/file/885-the-blue-bond-proposal-english/. Accessed 10 Feb 2013.
Di Cesare, A., G. Grande, M. Manna, and M. Taboga. 2012. *Recent estimates of sovereign risk premia for euro-area countries*, Occasional papers 128. Rome: Banca d'Italia.
Freedman, C., M. Kumhof, L. Douglas, and J. Lee. 2009. *The case for global stimulus. IMF staff position note*. Washington DC: International Monetary Fund.
German Council of Economic Advisors. 2012. *A redemption pact for Europe: Time to act now*. VoxEU. Available at: http://voxeu.org/article/redemption-pact-europetime-act-now. Accessed 10 Feb 2013.
Gros, D. 2011. *Eurobonds: Wrong solution for legal, political, and economic reasons*. VoxEU. Available at: http://www.voxeu.org/article/eurobonds-arewrong-solution. Accessed 10 Feb 2013.
Gros, D., and D. Schoenmaker. 2012. *A European deposit insurance and resolution fund – An update*. CEPS policy brief. Brussels: Centre for European Policy Studies, 11 September.
Gros, D., and N. Thygesen. 1998. *European monetary integration*, 2nd ed. London: Longman.
Gros, D., C. Alcidi, and A. Giovannini. 2012. *Central banks in times of crisis: The FED vs. the ECB. CEPS Policy Briefs,* 11 July 2012.
Henning, R., and M. Kessler. 2012. *Fiscal federalism: US history for architects of Europe's fiscal union*, Working paper 12.1. Washington, DC: Peterson Institute for International Economics.
Merler, S., and J. Pisany-Ferry. 2012. *The simple macroeconomics of north and south in EMU*, Bruegel working paper 2012/12. Brussels: Bruegel.
Padoa-Schioppa Group. 2012. *Completing the Euro. A road map towards fiscal union in Europe*. Available at: http://www.eng.notre-europe.eu/011-3024-Tommaso-Padoa-Schioppa-Group.html. Accessed 10 Feb 2013.
Pisany-Ferry, J. 2012. *The known unknowns and unknown unknowns of EMU*. Bruegel policy contribution, 2012/18. Brussels: Bruegel, October.
President of the European Council. 2012. *Towards a genuine economic and monetary union. Interim Report, 12 October, Brussels*. Available at: http://www.consilium.europa.eu/uedocs/cms_data/docs/pressdata/en/ec/132809.pdf. Accessed 10 Feb 2013.
Sinn, H.-W., and T. Wollmershäuser. 2012. Target loans, current account and capital flows: The ECB's rescue facility. *International Tax and Public Finance* 19: 468–508.
Von Hagen, J. 1993. Monetary union and fiscal union: A perspective from fiscal federalism. In *Policy issues in the operation of currency unions*, ed. P.R. Masson and M.P. Taylor. Cambridge: Cambridge University Press.
Weidmann, J. 2012. *There is light at the end of the tunnel*. Available at: http://www.bundesbank.de/Redaktion/EN/Interviews/2012_12_22_weidmann_wirtschaftswoche.html. 22 December.

Decentralization

E. Malinvaud

Abstract

Central planning is inefficient because it lacks incentives and is poorly informed. Complete decentralization risks being inequitable and also inefficient because markets are incomplete and public goods may be neglected. Intermediate systems can overcome these difficulties to the extent that planning mechanisms can mimic the market system while avoiding its deficiencies, public goods can be successfully delivered at the local level, and incentives to report and behave faithfully and to avoid free riding can be secured.

Keywords

Autonomy; Central planning; Competitive equilibrium; Decentralization; Decomposability; Free rider problem; Hayek, F.; Incentive compatibility; Incentives; Information; Non-economic motivation; Lange, O.; Malinvaud, E.; Misreporting; Pareto efficiency; Principal and agent; Private information; Prospective indices; Public goods; Resource allocation mechanisms; Social choice; Socialism; Mises, L. von

JEL Classifications

P0

The main question to be answered by the theory of resource allocation, or by the theory of economic organization, concerns the performances of alternative systems characterized by different degrees of centralization of decision taking. A fully centralized system runs the risk of being inefficient because it does not create proper economic incentives and the centre is poorly informed. A pure market system with its high degree of decentralization runs the risk of bringing inequitable results and being inefficient because markets can never be complete, externalities exist and public wants tend to be neglected. Can these risks be avoided within the two opposite extremes of pure centralization or full decentralization? Can intermediate systems better resolve the difficulties? And if so, how?

Basic to the discussion are two features: the nature of the *information* held by various agents, and the *incentives* that should lead them to behave in conformity with collective requirements. These features and the issue of decentralization do not only appear for full economic systems, which this entry will consider, but also for the internal organization of firms or communities. They are stylized in the principal–agent problem: which rules should determine how to share the proceeds of an activity between the principal owner and his better-informed agent? (Ross 1973; Grossman and Hart 1983).

For the clarification of the complex issues involved, theory starts from a model of the conditions of economic activity. It makes assumptions such that, independently of economic organization, there exists a best outcome, or at least a set of 'optimal' outcomes. It then asks how well alternative forms of organization succeed in finding, implementing or at least approaching this best outcome or set of optimal outcomes.

By so doing, the theory discussed here neglects two related questions: how to determine what should be considered as 'the best' outcome in a society with many individuals, and which non-economic considerations interfere with the issue of decentralization? The theory of social choice shows the fundamental difficulty of the first question (Arrow 1951), which is avoided when optimality is identified with Pareto efficiency. As for the second, philosophers may find in human nature or in the aims pursued by human societies reasons that favour some organization, beyond its economic performance; in particular, the right of individuals to autonomy appears fundamental in Western culture and is an important justification of decentralization, and even of the market system for such economists as Hayek (1944).

D

Formal Concepts and Preliminaries

The following conceptual apparatus, although not yet common, is well suited to the purpose (see Hurwicz 1960; Mount and Reiter 1974).

An *economic environment* is defined by a set of commodities and their possible uses, by a list of agents and their characteristics (technology, endowments, preferences, and so on), and by an initial information structure (what each agent knows). The feasible set of economic environments defines 'the economy'.

An important property of an economy is its higher or lower degree of *decomposability*, which concerns agents' characteristics and the information structure. The highest decomposability is assumed in competitive equilibrium theory, where all consumption is private, no external effect exists and a *private information structure* prevails (each agent perfectly knows its own characteristics and the situation on all markets, but nothing else). But models with public goods, for instance, usually admit some decomposability, which matters for the validity of the results.

An *optimality correspondence* P: $E \to A$ defines which vectors of actions simultaneously taken by the various agents are optimal when the economic environment is e, i.e. optimal vectors belong to $P(e)$ (clearly, E is the set of feasible e, that is 'the economy', while A is the set of feasible vectors a, each one of them defining the actions taken by all the agents). For instance $P(e)$ may be the set of Pareto efficient vectors. But in the theory discussed here, it is often more narrowly defined so as to take equity considerations into account: a social utility function may have to be maximized or a rule on the consumers 'income distribution' satisfied.

A *resource allocation mechanism* f: $E \to A$ should select one $a = f(e)$ for each environment

e (in some cases f may be multivalued, i.e. become a correspondence). The best formalized mechanism is the competitive equilibrium of a 'private ownership economy'. A study of decentralization requires a careful specification of the mechanism, which is typically viewed as operating in two stages: first, an iterative exchange of messages, usually between the agents and a centre, resulting in a message correspondence g: $E \rightarrow M$ (the message $m = g(e)$ specifies what information about e has been collected at the centre), second an outcome function h: $M ! \rightarrow$ A. For instance, the competitive mechanism is often specified as resulting from the tâtonnement process, in which an auctioneer learns which demands and supplies are announced at various proposed vectors of prices, and searches for the equilibrium prices; once these prices are found, the outcome function gives the equilibrium exchanges, hence productions and consumptions.

The performances of alternative mechanisms of course concern the final result: one must know whether the outcome $f(e)$ belongs to the optimal set $P(e)$ for all environments in E, or at least for a precise subset of E, and how close it is to $P(e)$ otherwise. But interesting performances also concern intermediate features of the mechanism, which usually is iterative. At step t the previously collected message m_{t-1} is enriched according to $m_t = g_t\ (m_{t-1},\ e)$ and, if necessary, the process could end by $a = h_t\ (m_t)$. In a *finite* procedure it does end at T with $m = m_T$ and $h(m) = h_T(m_T)$; but most mechanisms assume an infinite sequence of m_t for $t = 1, 2 \ldots$ ad infinitum. One must then know whether and how $h_t(m_t)$ approaches P(e), monotonically or otherwise. Since the transmission of information is costly, the nature and size of the message space M_t to which m_t belongs are also important characteristics (Mount and Reiter 1974).

The Planning Problem

Early in this century many economists objected to socialist planning programmes that could not be implemented, because they unrealistically assumed that a central administration could have the knowledge and computing power required for an efficient control of economic activity. The leading figure was L. von Mises (1920 in particular); but Hayek (1935) was first to emphasize the problems raised by the decentralization of information. Socialist economists answered that decentralized mechanisms could operate, either mimicking the market system while being free of its deficiencies (Lange 1936) or using different well conceived modes of information gathering (Taylor 1929). The debate was, in the interwar years, the subject of the *'economic theory of socialism'*. (For a well-documented survey, see Bergson 1948).

The problem was again taken up during the 1960s, in particular because the logic of efficient planning was discussed in Eastern and Western Europe (Arrow and Hurwicz 1960; Kornai 1967; Malinvaud 1967; Heal 1973). Many planning procedures were rigorously studied as resource allocation mechanisms. Their definition implied an iterative exchange of information between a Central Planning Board and firms, sometimes also representative consumers. The additional messages provided by the function gt at step t then consisted of *prospective indices* announced by the Board, for instance prices for the various commodities, and replies called *proposals* sent to the Board by firms and other agents, for instance preferred techniques of production and their input requirements, or supplies and demands.

In this discussion it is common to distinguish between price-guided procedures, in which the Board announces price vectors, and other procedures, in which quantity indices or targets worked out at the centre play a more or less important role. The nature and properties of the environment are then found to be crucial for the determination of the relative performances of alternative procedures, in particular of price-guided against quantity-guided procedures (Weitzman 1974).

The analytical study of various procedures usually assumes that decentralized agents exactly follow specified rules for the determination of their proposals and so faithfully reveal part of their private information. Some procedures are then found to be efficient and to permit achievement of distributive objectives. But efficiency is typically easier precisely in those environments that are also

favourable to the efficiency of free competition. Besides the possibility of incorrect reporting, the main difficulty concerning the relevance of this literature is to know whether its models provide an approximate representation of procedures that are actually used, or at least administratively feasible. Manove (1976) has made this claim for his representation of Soviet planning.

The Public Good Problem

The most relevant field of application may very well be the theory of public goods. Decisions concerning the provision of public services and their financing cannot be fully decentralized; but the knowledge required is dispersed and must be gathered in a proper way. Hence even the positive theory of public goods was often formulated along lines that look like those of planning procedures (Malinvaud 1971). The same remark applies to decisions concerning public projects with large fixed costs, even if their output is privately consumed.

Considered as a planning procedure, the search for the best decision is often viewed as involving 'prospective indices' that define amounts of service to be provided, ask for corresponding individual marginal utilities and look whether the sum of the latter would cover the cost of additional service. This is compatible with the dual arrangement for private goods, prices being announced, supplies and demands being the replies. The procedure is then quantity-guided for public goods and price-guided for private goods (Drèze and Vallée Poussin 1971).

The collective consumption of many types of public goods is not really national but limited to local communities (primary education, city transports, and so on). Administrative science sees the decentralization issue as being to know at which level should decisions be taken: at the national level, so as to distribute fairly these services among communities, or at the local level, so as to permit better adaptation to local needs and wishes. Economists do not seem to have contributed to this issue; their discussion of local public goods assumes full administrative decentralization (Tiebout 1956).

Incentive Compatibility

The study of a decentralized system has to consider whether the actual reports and behaviour of individual agents do not deviate from what they are supposed to report and do; in case of deviations, how are the performances of the system affected? The problem is serious: once the rules of organization and decisions are known, individual agents may benefit from misreporting their private information or from behaving in a way that, although deviant, does not clearly appear to be so. In other words, they may act as players in a game, rather than as members of a team, and this may be more or less detrimental for the optimality of the final result.

The problem has long been known for organizations in which some agents do not individually benefit from what is achieved and therefore lack the incentive to do their best. Monopolistic or other non-competitive behaviour is often interpreted as a breach of the normal rules of resource allocation. In the theory of public good the 'free rider problem' occurs as soon as some individuals, having a high marginal utility for the public good, would benefit from hiding this fact so as to contribute little to the financing of the good.

Study of the problem has been active during the past two decades (Green and Laffont 1979). The fundamental difficulty has been exhibited by such results as the following one: in the classical model of an exchange economy with a finite number of consumers, no procedure can be found that would necessarily lead to a Pareto efficient result in which individuals, acting as players in a non-cooperative game, would faithfully report (Hurwicz 1972). However, misreporting may not prevent a procedure from eventually leading to an optimum, as was proved in a number of cases.

Experiments moreover show that the game-theoretic approach to the incentive problem may be misleading because it neglects non-economic motivations that individuals may find for accepting a team-like behaviour and therefore for faithfully reporting (Smith 1980).

Bibliography

Arrow, K. 1951. *Social choice and individual values*. New York: Wiley.

Arrow, K., and L. Hurwicz. 1960. Decentralization and computation in resource allocation. In *Essays in economics and econometrics in honour of Harold Hotelling*, ed. R. Pfouts. Chapel Hill: University of North Carolina Press.

Bergson, A. 1948. Socialist economics. In *A survey of contemporary economics*, ed. H. Ellis. Philadelphia: Blakiston.

Drèze, J., and D. de la Vallée Poussin. 1971. A tâtonnement process for public goods. *Review of Economic Studies* 38: 133–150.

Green, J., and J.-J. Laffont. 1979. *Incentives in public decision-making*. Amsterdam: North-Holland.

Grossman, S., and O. Hart. 1983. An analysis of the principal-agent problem. *Econometrica* 51 (1): 7–45.

Hayek, F. 1935. Socialist calculation: The state of the debate. In *Collectivist economic planning*, ed. F. Hayek. London: G. Routledge & Sons.

Hayek, F. 1944. *The road to Serfdom*. Chicago: University of Chicago Press.

Heal, G. 1973. *The theory of economic planning*. Amsterdam: North-Holland.

Hurwicz, L. 1960. Optimality and information efficiency in resource allocation processes. In *Mathematical methods in the social sciences*, ed. K.J. Arrow, S. Karlin, and P. Suppes. Stanford: Stanford University Press.

Hurwicz, L. 1972. On informationally decentralized systems. In *Decision and organization*, ed. R. Radner and C. McGuire. Amsterdam: North-Holland.

Kornai, J. 1967. *Mathematical planning of structural decisions*. Amsterdam: North-Holland.

Lange, O. 1936. On the economic theory of socialism. *Review of Economic Studies* 4 (53–71): 123–142.

Malinvaud, E. 1967. Decentralized procedures for planning. In *Activity analysis in the theory of growth and planning*, ed. E. Malinvaud and M. Bacharach. London: Macmillan.

Malinvaud, E. 1971. A planning approach to the public good problem. *The Swedish Journal of Economics* 11: 96–112.

Manove, M. 1976. Soviet pricing, profit and technological choice. *Review of Economic Studies* 43: 413–421.

Mount, K., and S. Reiter. 1974. The informational size of message spaces. *Journal of Economic Theory* 8 (2): 161–192.

Ross, S. 1973. The economic theory of agency: The principal's problem. *American Economic Review* 63 (2): 134–139.

Smith, V. 1980. Experiments with a decentralized mechanism for public good decisions. *American Economic Review* 70: 584–599.

Taylor, F.M. 1929. The guidance of production in a socialist state. *American Economic Review* 19: 1–8.

Tiebout, C.M. 1956. A pure theory of local expenditures. *Journal of Political Economy* 64: 416–424.

von Mises, L. 1920. Economic calculation in the socialist commonwealth. First published in German in *Archiv für Sozialwissenshaft*, April; English translation in *Collectivist Economic Planning,* ed. F. Hayek. London: G. Routledge & Sons, 1935.

Weitzman, M. 1974. Prices versus quantities. *Review of Economic Studies* 41: 477–491.

Decision Theory

H. M. Polemarchakis

To decide is to choose from sets of alternatives. Decision theory is concerned with *rationality* in choice.

1. An individual faces a set of alternatives, C. Over the set of alternatives, the individual has preferences described by a binary relation R: c is preferred or indifferent to (at least as good as) c' if and only if cRc'. The following postulate then characterizes rationality.

Postulate

The preference relation R is complete and transitive.

If cRc' and $c''Rc'$; $c''Rc$ for any c' and c, either $c'Rc$ or cRc'. Rationality in this simple framework is thus a consistency requirement.

From $A \subseteq C$ the individual chooses $d(A) \subseteq A$, not necessarily a singleton, the set of elements of A which are maximal for R: $d(A) = \{c \in A: cRc' \text{ for all } c' \in A\}$; the definition of the choice correspondence d is sometimes considered an additional aspect of rationality.

Note as an example that, if C is the consumption set, R is the individual's preference relation over commodity bundles, prices and income generate budget sets A, and d is the demand correspondence.

We are often interested in the following aspects:

Representability

A function u defined on the set of alternatives C represents the preference relations on R, if $u(c) \geq u(c')$ if and only if cRc'. When such a function exists, it is the objective function of the individual; in the special case of the consumer, it is referred to as the utility function. Note that if a function u represents the relation R, so does any monotonically increasing transformation of u; the representation is thus ordinal. The classical theorem on representability is due to Debreu (1954): A transitive and complete preference relation R on a set C is representable by a (continuous) objective function as long as the set of alternatives C is a connected, separable topological space, and the relation R is continuous: for all $c \in C$, the sets $\{c' \in C: c'Rc\}$ and $\{c' \in C: cRc'\}$ are closed. Debreu also gave an example of a relation which fails to be representable: Let C be the non-negative orthant of two-dimensional Euclidean space, and let the relation R be defined as follows: $c = (c_1, c_2)Rc' = (c_1', c_2')$ if $c_1 > c_1'$ or $(c_1 = c_1'$ and $c_2 > c_2')$; this is known as the lexicographic relation. A straight-forward argument shows that the representability of R (not necessarily by a continuous function) would imply that the set of real numbers is countable, a contradiction. Representability is thus a strictly stronger requirement than rationality. Beyond representability and rationality, one may investigate the correspondence between qualitive properties of the relation R and the functional form of some representation u. Additive separability turns out to be of interest, as we shall see when we impose more structure; and the main result is due again to Debreu (1959): Let $N = \{1,\ldots, n\}$, $n > 2$, and let $C = \prod_{j \in N} C_j$ be a connected and separable topological space. A transitive, complete and continuous relation R on C has an additively separable representation $u(c) = \sum_{j \in N} u_j(c_j)$ if and only if for every $J \subseteq N$ and $c_J = (c_j)_{j \in J}$ the induced relation R_{cJ} on $C_{N/J} = \prod_{j \in N/J} C_j$ is independent of c_J, and at least three factors $j \in N$ are essential: a factor $j \in N$ is essential if not all elements of $C_{N/j}$ are mutually (preferred or) indifferent under the induced relation R_j. The case of only two essential factors can be treated separately.

Observability and Recoverability

The preference relation R is an unobservable characteristic of the individual. What is, in principle at least, observable is the choice correspondence d on a class A of subsets of C. Samuelson (1938) first, in the context of consumer theory, gave a definition of rationality in terms of the observable characteristics of the consumer. Elaborating on Samuelson, Richter (1966) defined a consumption bundle, c, to be directly revealed preferred to another bundle, c', cVc', if, for some budget set A, $c \in$ d(A) while $c' \in A$. A bundle c is indirectly revealed preferred to another, c', cWc', if there exists a finite sequence of bundles, $c^1, \ldots, c^n$, such that $c = c^1Vc^2,\ldots,Vc^{n-1}Vc^n = c'$, The consumer is congruous if, for all c, $c' \in C$ and all budget sets A, whenever $c \in d(A)$, $c' \in A$ and, cWc', $c' \in d(A)$. Richter proceeded to show that congruence, which characterizes the observable choice correspondence, is equivalent to the earlier definition of rationality: A consumer satisfies the congruence axiom if and only if he is rational. Sen (1971) extended the argument beyond consumer theory, to general choice situations. Closely related to observability is the issue of recoverability. Even when the choice correspondence is known to be generated from the maximization of some underlying complete and transitive preference relation, knowledge of d need not suffice to identify unambiguously and recover R; the generating binary relation need not be unique. Mas-Colell (1977), in the context of consumer theory, showed that recoverability is indeed possible under mild regularity assumptions. Questions of prediction require recoverability based on the observation of the choice correspondence on a restricted domain, for which further qualitative assumptions on the underlying binary relation are necessary.

D

Existence and Computability

For an arbitrary $A \subseteq C$ no maximal element for the relation R needs to exist. The choice correspondence is then defined on a restricted class A of subsets of the set of alternatives. Even if maximal elements can be shown to exist for $A \subseteq C$, there remains the issue of computability.

2. Under uncertainty, the objects of choice are not what ultimately determines the welfare of the individual. We follow the formalization of Savage (1954). States of the world are $s \in S$; a state of the world is an exhaustive and exclusive description of the environment. Consequences are $c \in C$; a consequence is what ultimately determines the welfare of the individual. Acts are $f \in F$, an act is a function f: $S \to C$, which associates consequences to states. The set F is the set of all possible acts; elements of the set F are the objects of choice of the individual. An event is a subset $B \subseteq S$: an event is said to occur if it contains the true or actual state of the world; for an event B, its complement is $B = S/B$. Certainty is the limiting case in which S is a singleton and the sets of acts. F, and of consequences, C, coincide.

A series of postulates which characterize rationality under uncertainty imply that the individual's preferences over acts, described by the preference relation R, have an expected utility representation $E_p u$, where E is the expectation operator, p is a probability measure on the set of states of the world S, and u is a cardinal utility index on the set of consequences C, unique up to monotonically increasing, linear transformations. Note that the existence of such a probability measure is not taken for granted.

Postulate (I)

The preference relation R over acts is transitive and complete.

This is the exact analogue of the postulate of rationality under certainty. Thus, under the additional technical assumptions of the representation theorem of Debreu (1954), the preference relation R is representable by an objective function $v : v(f) \geq v(f')$ if and only if fRf'.

The set of consequences C can be identified with the subset of constant acts, the acts which yield the same consequence at all states. It follows that implicit in the preference relation over acts is a preference relation over consequences.

Postulate (II)

For facts f, f', g and g' and an event B, of $f = f'$ and $g = g'$ on B, while $f = g$ and $f' = g'$ on B, f Rg if and only if f' Rg'.

Preferences over acts do not depend on the consequences they yield on states at which their consequences coincide; this is known as the sure thing principle and it corresponds, when probabilistic beliefs are taken as given, to the strong independence axiom. The sure thing principle guarantees the additive separability of the objective function across states; up to technical conditions, additive separability follows from the theorem of Debreu (1959) on additively separable representations. The sure thing principle is tenable as an aspect of rationality as long as states are exhaustive and exclusive descriptions of the environment. It has been challenged, however, on the ground that it is frequently violated in experimental set ups. The most famous such refutation is due to Allais (1953):

Let $S = \{s^1, s^2, s^3\}$ and $C = [0, \infty)$, and consider the following acts:

$$f = \begin{bmatrix} s^1 \to 1 \\ s^2 \to 1; \\ s^3 \to 1 \end{bmatrix} \quad g = \begin{bmatrix} s^1 \to 0 \\ s^2 \to 5; \\ s^3 \to 1 \end{bmatrix}$$

$$f' = \begin{bmatrix} s^1 \to 0 \\ s^2 \to 1; \\ s^3 \to 1 \end{bmatrix} \quad g' = \begin{bmatrix} s^1 \to 0 \\ s^2 \to 5. \\ s^3 \to 0 \end{bmatrix}$$

According to the sure thing principle ($B = \{s_1, s_2\}$, $\sim B = \{s_3\}$) $f\,Rg$ if and only if f' Rg'. It is most often the case, however, that with payoffs (consequences) dominated in units of $1,000,000 and the probability of occurrence of the states

known to be (0.01, 0.1, 0.89), individuals state their preferences as $f\,Rg$ and $g'\,Rf'$. Machina (1982) has argued that the sure thing principle can be understood as characteristic of an approximation to a general preference relation.

With postulate (ii), conditional preferences are well defined: For an event B, $f\,R_Bg$ if and only if there exist acts f' and g' such that $f'\,Rg'$ and f coincides with g on B while, on B, f' coincides with g'. Knowledge of the restriction of f and g on B determines unambiguously the individual's preferences between f and g conditional on B; it suffices to complete f_B and g_B so that they coincide on B.

An event B is null if and only if $f\,R_Bg$ for any acts $f, g \in F$.

Postulate (III)

For any constant acts f and g and any non-null event B, $f\,R_Bg$ if and only if $f\,Rg$.

This excludes state-dependent preferences. For any $s, s' \in S$ the representations v_s and $v_{s'}$ of the conditional preferences R_s and $R_{s'}$, must be ordinally equivalent. It may seem that one can introduce state dependence by replacing the set of consequences C by the product $C^* = C \times S$; this allows for state dependence since states of the world are now part of the specification of the consequences of an act. This may, however, be just empty formalism; the construction would oblige the individual to contemplate acts assigning to, say, states s the consequence (c, s'), while s and s' are mutually exclusive states of the world. For postulate (iii) to be tenable it is necessary to keep clear the distinction between acts and consequences.

Postulate (IV)

For consequences c, c', d and $d' \in C$, acts f, f', g and $g' \in$, and events A and B

$$c'Rc, f = \begin{bmatrix} c & \text{on } A \\ c' & \text{on } \sim A' \end{bmatrix} \quad g = \begin{bmatrix} c & \text{on } B \\ d' & \text{on } \sim B' \end{bmatrix}$$

and

$$d'Rd, f' = \begin{bmatrix} d & \text{on } A \\ d' & \text{on } \sim A' \end{bmatrix} \quad g' = \begin{bmatrix} d & \text{on } B \\ d' & \text{on } \sim B \end{bmatrix}$$

then $f\,Rg$ if and only if $f'\,Rg'$.

The individual has consistent probability beliefs. In addition to yielding a probability measure on the set of states, S, postulate (iv) will imply that the conditional objective functions $v_s, s \in S$, are not simply ordinally equivalent, but differ only by a monotonically increasing linear transformation; $v_s = p_s\,u$.

With postulate (iv), it makes sense to speak of one event B being at least as probable as another event B': $BR^*\,B'$ if there exist acts f and g and consequences c and c' such that

$$c'Rc, f = \begin{bmatrix} c & \text{on } B \\ c' & \text{on } \sim B' \end{bmatrix} \quad g = \begin{bmatrix} c & \text{on } B' \\ c' & \text{on } \sim B' \end{bmatrix}$$

and gRf.

Postulate (V)

There exists at least a pair of consequences c and c' such that cRc' but not $c'Rc$.

This is simply to avoid the case of a preference relation which leaves the individual indifferent between any two acts. Such a preference relation could not be used to elicit the individual's probability beliefs which, by definition, must assign higher probability to some events than others.

We now proceed to outline the argument first for the derivation of a probability measure and then for the expected utility representation.

It is straightforward to check that the binary relation R^* over events is indeed a qualitative probability; that is, a complete and transitive relation which in addition satisfies the conditions that BR^*B' if and only if $(B \cup B'')R^*(B' \cup B'')$ whenever $(B \cap B'') = \phi, BR^*\phi$, and $SR^*\phi$ but not ϕR^*S. A probability measure p on S is a positive function such that $p(B \cup B') = p(B) + p(B)$ whenever $B \cap B' = \phi$ and $p(S) = 1$. If S carries a probability measure p and a qualitative probability R^* such that $p(B) \geq p(B')$ if and only if $BR^*\,B$, p agrees with (represents) R^*. Even for finite S, however, there exist qualitative

probabilities for which no agreeing probability measure can be found. A probability measure p which agrees with R^* exists as long as an additional continuity condition is satisfied. We shall assume that this condition holds; thus p exists and is unique. Note that the definition of probability requires finite and not countable additivity; thus we avoid the need to specify the σ on which the measure is defined. The continuity assumption which we employ to guarantee that a probability measure exists implies that this probability measure satisfies a certain non-atomicity property; it excludes finite and even countable state spaces; to relax the condition is, however, cumbersome. Finally, observe that the representation of qualitative probability by a probability measure extends to conditional probability; indeed, we obtain Bayes' rule for every non-null event $B'p(B') = p(B \cap B')$.

With the probability measure p on S, every action $f \in F$ induces a probability measure μ_f on the set of consequences C: for $A \subseteq C, \mu_f(A) = p\{s \in S : f(s) \in A\}$. A probability measure is simple if it has finite support: acts which induce simple measures are gambles. Let M^* be the subset of the set M of all probability measures on C of simple measures. Observe that M^* is a mixture set: to each $\alpha \in [0,1]$ and each pair of elements $\mu, \mu' \in M^*$ there corresponds unambiguously an element $\alpha\mu + (1-\alpha)\mu'$ such that $1\mu + 0\mu' = \mu, \alpha\mu + (1-\alpha)\mu' = (1-\alpha)\mu' + \alpha\mu$ and $\alpha[\alpha'\mu + (1-\alpha')\mu'] + (1-\alpha)\mu' = \alpha\alpha'\mu + (1-\alpha\alpha')\mu'$. It follows from the postulates, and this is the key step in the construction, that acts are evaluated by the individual only with respect to the measures which they induce on the set of preferences. Equivalently, the preference relation R on F induces unambiguously a complete and transitive binary relation on M^*, which we also denote by R : $\mu R\mu'$ if $\mu = \mu_f, \mu' = \mu_f$, and fRf'. Furthermore, for $\mu, \mu', \mu'' \in M^*$ and $\alpha \in [0,1]$, $\mu R\mu'$ if and only if $[\alpha\mu + (1-\alpha)\mu'']R[\alpha\mu' + (1-\alpha)\mu'']$, while for $\mu, \mu', \mu'' \in M^*$ with $\mu R\mu' R\mu''$, there exists a unique $\alpha(\mu'; \mu, \mu'') \in [0,1]$ such that $[\alpha\mu + (1-\alpha)\mu'']R\mu'$ and $\mu' R[\alpha\mu + (1-\alpha)\mu'']$. The cardinal utility index u on C such that, restricted to the subset $F^* \subseteq F$ of gambles, $v = E_p u$ is constructed as follows: For a given pair of consequences $\overline{c}$ and $\underline{c}$ with $\overline{c}R\underline{c}$, let $u(\overline{c}) = 1$, $u(\underline{c}) = 0$; for c such that $\overline{c}RcR\underline{c}$, let $u(c) = \alpha(\mu_c; u_{\overline{c}}, u_{\underline{c}})$; the extension of u to all of C is straightforward. Evidently, the cardinal utility index is unique up to monotonically increasing linear transformations. Under additional technical restrictions which involve the boundedness of the cardinal utility index (or, equivalently, the continuity of the preference relation with respect to the appropriate topology on the set of probability measures over consequences) the expected utility representation can be extended to acts which are not necessarily gambles. The Savage postulates do not allow for state dependence of the cardinal utility index u. Additional structure is required, as in Dreze (1984), for state dependence to be introduced and for probability beliefs to be distinguished from state dependence.

3. Choice may occur *sequentially*. We revert to a framework of certainty. The set of alternatives over which the individual has preferences and among which he chooses is C: for simplicity, we take it to be finite. The individual is characterized by his preference relation R and C, which is transitive and complete. Let C be the power set of all subsets, A, of C. The preference relation R on C induces unambiguously a preference relation R on C, which inherits its transitivity and completeness: ARA' if and only if $cR\underline{c}'$ for some $c \in A$ and all $c' \in A'$. Note that the definition of R embodies the principle of backward induction: Faced with the choice between sets of alternatives A and A', the individual prefers A if a subsequent choice among the elements of A is at least as good as any possible choice among the elements of A'.

   The problem of (time) *consistency* arises, as Strotz (1955–56) has noticed, when the individual's preferences over C change between the point at which he chooses among sets of alternatives and the subsequent point at which he chooses among alternatives

in the set he chose earlier: Let R^2 be the final preference relation over C and let R^1 be the preference relation over C when he chooses over C. Two preference relations on C can be induced by the pair (R^1, R^2). The naive preference: $AR^N A'$ if and only if cRc' for some $c \in A$ and all $c' \in A'$; in this case the individual ignores the subsequent change of preferences which he may be able to foresee. The sophisticated preference: $AR^s A'$ if and only if cR^1c' for some $c \in A$ and some $c' \in A'$ such that cR^2c'' and $c' R^2C''$ for all $c''' \in A$ and all $c''' \in A'$, respectively; in this case, the individual foresees his subsequent change of preferences and attempts to commit himself to the extent that the choice out of A which will be made according to R^2 is as good as possible according to the current ranking, R^1. Again, the use of backward induction is evident. The individual is consistent if R^1 and R^2 and hence R^N and R^S coincide.

The resolution of uncertainty may occur sequentially. States of the world, acts, and consequences are as before. The individual's preferences over acts are represented by the objective function v. Let S be a partition of S. For $S^0 \in S$, let F^0 be the set of all acts $f^0 : S^0 \to C$. The question follows whether there exists an objective function on F^0 which is naturally induced by v. When the objective function v has an expected utility representation, the answer is straightforward: it suffices to replace the probability measure p by the conditional probability measure $p^0 = p|S^0$, thus obtaining $v^0 = E_{p0}u$. Formally, the domain of v^0 is F, not F^0;yet no ambiguity arises, since, for any act $f \in Fv^0$ (f) depends only on the restriction of f *to* S^0. Suppose C and hence F are well as linear spaces (addition and scalar multiplication are well defined). If the individual has taken act $\overline{f} \in F$ before S^0 is realized, he ranks elements $f^0 \in F^0$ according to $v^0(\overline{f} + f_0)$, where f_0 is the unambiguous extension of f^0 to F which takes the value zero on S/S^0. When the objective function v does not have an expected utility representation, the argument breaks down. It is formally possible to ignore the resolution of uncertainty and rank acts $f^0 \in F^0$ according to $v^0(\overline{f} + f_0) = v(\overline{f} + f_0)$. But this is contrived: it amounts to considering as 'occurring' states of the world $s \in S/S^0$ when they are known not to have occurred.

4. We have concentrated on individual behaviour. Alternatively, it may be only aggregate behaviour which is observable or of interest. Suppose that the set of alternatives is a linear space: Let $h = 1, \ldots, H$ be a collection of individuals, and let $Q = (\ldots, Q^h, \ldots)$ be a distribution scheme: to any subset $A \subseteq C$ in a class A, it assigns a vector of subsets $(\ldots, A^h \subseteq C, \ldots)$ such that $A^1 + \ldots + A^h + \ldots A^H = A$. Let d be the aggregate choice correspondence restricted to A. Two questions follow: Under what conditions do there exist individual preference relations $R^1, \ldots, R^h, \ldots, R^H$ such that the aggregate correspondence coincide with d? Note that the question is well posed only with reference to a distribution scheme Q. Alternatively, under what conditions on the individual preference relations $R^1, \ldots, R^h, \ldots, R^H$ and the distribution scheme Q can the aggregate choice correspondence be derived from the optimization of a representative preference relation? In the context of consumer theory both questions have been studied extensively. Sonnenschein (1972) first suggested that as long as the number of individuals is large relative to the number of commodities, the income distribution scheme is derived from an arbitrary but fixed distribution of initial endowments, and only the excess demand of individuals is observed as prices vary, homogeneity with respect to prices and the budget constraint (Walras' Law) are the only constraints which aggregate behaviour must display; individual rationality fails to have observable implications in the aggregate. Alternatively, as Gorman (1953) has shown, if individual preference are identical and homothetic (cRc' if and only if (λc) $R(\lambda c'M)$, $(\lambda > 0)$ as well, the aggregate behaves like a single, rational individual.

D

Note that a qualitative restriction on the preference relation is employed for individual rationality to have observable implications in the aggregate.

5. Throughout, the alternative which obtained was determined unambiguously by the decision of the individual and the resolution of exogenous uncertainty. We have ignored issues of feasibility, equilibrium and strategy.

See Also

- ▶ Arrow–Debreu Model of General Equilibrium
- ▶ Exchange
- ▶ Organization Theory
- ▶ Statistical Decision Theory
- ▶ Uncertainty

Bibliography

Allais, M. 1953. Le comportement de l'homme rationel devant le risque; critique des postulats et axiomes de l'école Americaine. *Econometrica* 21: 503–546.

Debreu, G. 1954. Representation of a preference ordering by a numerical function. In *Decision processes*, ed. R.M. Thrall, C.H. Coombs, and R.L. Davis, 159–165. New York: Wiley.

Debreu, G. 1959. Topological methods in cardinal utility. In *Mathematical methods in the social sciences*, ed. K.J. Arrow, S. Karlin, and P. Suppes, 16–26. Stanford: Stanford University Press.

Gorman, W.M. 1953. Community preference fields. *Econometrica* 21(1): 63–80.

Green, J.R., L.J. Lau, and H.M. Polemarchakis. 1979. Identifiability of the von Neumann–Morgenstern utility function from asset demands. In *General equilibrium, growth and trade*, ed. J.R. Green and J. Scheinkman. New York: Academic Press.

Machina, M. 1982. 'Expected utility' analysis without the independence axiom. *Econometrica* 50(2): 277–323.

Mas-Colell, A. 1977. The recoverability of consumers' preferences from market demand functions. *Econometrica* 45(6): 1409–1430.

Richter, M.K. 1966. Revealed preference theory. *Econometrica* 34(3): 635–645.

Samuelson, P.A. 1938. A note on the pure theory of consumers' behaviour. *Economica* 5(1): 61–71.

Savage, L.J. 1954. *The foundations of statistics*. New York: Wiley.

Sen, A.K. 1971. Choice functions and revealed preference. *Review of Economic Studies* 38(2): 307–317.

Sonnenschein, H. 1972. Market excess demand functions. *Econometrica* 40(3): 549–563.

Strotz, R.H. 1955–56. Myopia and inconsistency in dynamic utility maximization. *Review of Economic Studies* 23(3): 165–180.

Decision Theory in Econometrics

Keisuke Hirano

Abstract

The decision-theoretic approach to statistics and econometrics explicitly specifies a set of models under consideration, a set of actions that can be taken, and a loss function that quantifies the value to the decision-maker of applying a particular action when a particular model holds. Decision rules, or procedures, map data into actions, and can be ordered according to their Bayes, minmax, or minmax regret risks. Large sample approximations can be used to approximate complicated decision problems with simpler ones that are easier to solve. Some examples of applications of decision theory in econometrics are discussed.

Keywords

Admissibility criterion; Auction models; Bayes risk; Bayes rule; Computational methods; Decision rules; Decision theory in econometrics; Instrumental variables; Local asymptotic normality (LAN); Markov chain Monte Carlo methods; Maximum likelihood; Minmax principle; Minmax-regret principle; Nonparametric density estimation; Nonparametric models; Nonparametric regression; Point estimators; Portfolio choice; Savage, L. J.; Search models; Semiparametric models; Statistical decision theory; Time series models; Treatment assignment; White noise

JEL Classifications

C44

The decision-theoretic approach to statistics and econometrics explicitly specifies a set of models under consideration, a set of actions available to the analyst, and a loss function (or, equivalently, a utility function) that quantifies the value to the decision-maker of applying a particular action when a particular model holds. Decision rules, or procedures, map data into actions, and can be evaluated on the basis of their expected loss.

Abraham Wald, in a series of papers beginning with Wald (1939) and culminating in the monograph (Wald 1950), developed statistical decision theory as an extension of the Neyman–Pearson theory of testing. It has since played a major role in statistical theory for point estimation, hypothesis testing, and forecasting, especially in the construction of 'optimal' procedures. Some textbooks such as Ferguson (1967) and Berger (1985) emphasize statistical decision theory as a foundation for statistics. But the decision theory framework is sufficiently flexible that it can be used for many empirical applications that do not fit neatly into the usual statistical set-ups. Some examples are discussed below.

Like the Neyman–Pearson theory, Wald's approach emphasizes evaluating the performance of a decision rule under various possible parameter values. There does not always exist a single rule that dominates all others uniformly over the parameter space, just as there does not always exist a uniformly most powerful test in the special case of hypothesis testing. Wald, who also made contributions to game theory, proposed to evaluate a procedure by its minmax risk – the worst-case expected loss over the parameter space. Savage (1951) discusses the minmax principle and suggests an alternative, the minmax-regret principle. Alternatively, one can place a probability measure on the parameter space, and evaluate rules by their weighted average (Bayes) risk.

Basic Framework

In Wald's basic framework, we start with a set of actions $\mathscr{A}$, and a parameter space Θ, which characterizes the set of models under consideration. A loss function $L(\theta, a)$ gives the loss or disutility suffered from taking action $a \in \mathscr{A}$ when the parameter is $\theta \in \Theta$. The decision maker observes some random variable Z, distributed according to a probability measure P_θ when θ is the 'true' parameter. Here, the parameter space Θ could be finite-dimensional (corresponding to a parametric family of distributions) or infinite-dimensional (corresponding to semiparametric and nonparametric models). The observed random variable Z could be a vector, as for example in the situation of observing a random sample of size n from some distribution. Often, the set of possible probability measures $\{P_\theta : \theta \in \Theta\}$ is called a *statistical experiment*.

A decision rule or procedure $d(z)$ maps observations on Z into actions. In some cases, it is useful to allow for randomization over the actions. A randomized decision rule is a mapping from observations into probability measures over the action space. A simpler, usually equivalent formulation is to consider rules $\delta(z, u)$ which are allowed to depend on the observed value z and the value u of a random variable U, distributed standard uniform independently of Z. The *risk*, or expected loss, of a decision rule δ under θ is defined as

$$\begin{aligned} R(\theta, \delta) &= E_\theta[L(\theta, \delta(Z, U))] \\ &= \int_0^1 \int L(\theta, \delta(z, u)) dP_\theta(z) du. \end{aligned}$$

A rule δ is *admissible* if there exists no other rule δ' with

$$R(\theta, \delta') \leq R(\theta, \delta), \forall \theta \in \Theta,$$

And

$$R(\theta, \delta') < (\theta, \delta) \text{ for some } \theta.$$

Ordering Decision Rules

In general, there are many admissible decision rules, which may do well in different parts of the

parameter space. Thus, while the admissibility criterion eliminates obviously inferior rules, it may not provide concrete guidance on how to 'solve' the decision problem. Additional criteria can help by providing a sharper partial ordering of decision rules.

One way to rank decision rules is to average their risk over the parameter space. Let Π be a probability measure on Θ. The *Bayes risk* of a decision rule δ is

$$r(\Pi, \delta) = \int R(\theta, \delta) d\Pi(\theta).$$

A rule is a *Bayes rule* if it minimizes this weighted average risk. Let the probabilities P_θ have densities p_θ with respect to some dominating measure, and let the prior Π have density π. Typically, a Bayes rule can be implemented by choosing, for any given observed data z, the action that minimizes the posterior expected loss

$$\int L(\theta, a) d\Pi(\theta | z),$$

where $\Pi(\theta|z)$ is the posterior distribution with density

$$\pi(\theta | z) = \frac{\pi(\theta) p_\theta(z)}{\int p_\theta(z) d\Pi(\theta)}.$$

There is a close connection between the admissible rules and the Bayes rules. If the parameter set is finite, a Bayes rule for a prior that places positive probability on every element of Θ is admissible. Furthermore, 'complete class theorems' give results in the opposite direction. In particular, if the parameter set is finite, any admissible rule is Bayes for some prior distribution. If Θ is not finite, some care needs to be taken to make a precise statement of the relationship between the admissible and Bayes rules; see for example Ferguson (1967).

An alternative ordering is based on the worst-case risk $\sup_{\theta \in \Theta} R(\theta, \delta)$. A *minmax* rule δm satisfies

$$\sup_{\theta \in \Theta} R(\theta, \delta_m) = \inf_{\delta} \sup_{\theta \in \Theta} R(\theta, \delta).$$

In general, a minmax rule need not be admissible.

A closely related criterion is the *minmax regret* criterion. The regret loss of a rule is the difference between its loss and the loss of the best possible action under θ:

$$L_r(\theta, a) = L(\theta, a) - \inf_{a \in \mathcal{A}} L(\theta, a).$$

We can then define regret risk as $Rr(\theta, \delta) = E_\theta(Lr(\theta, \delta(Z, U))$. The *minmax regret rule* minimizes the worst-case regret risk. This rule was suggested by Savage (1951) as an alternative to the minmax criterion. He argued that in cases where the minmax criterion is unduly conservative, minmax regret rules can be reasonable.

Savage (1954) showed that a decision-maker who satisfied certain axioms of coherent behaviour would act as if she placed a prior on the parameter space and minimized posterior expected loss. Gilboa and Schmeidler (1989) showed that, under a different set of axioms, a decision-maker would follow the minmax principle.

Calculation of Bayes and minmax rules can be difficult in many applications. Bayesian posterior distributions can be calculated directly when the prior and likelihood have a conjugate form. One way to solve for a minmax rule is to guess the form of a 'least favourable' prior and solve for the associated Bayes rule. If the risk function of the Bayes rule is everywhere less than the Bayes risk, then the rule is minmax. A related method is to construct a least favourable sequence of prior distributions, and calculate the limit of the Bayes risks. If a particular rule has worst-case risk lower than the limit of Bayes risks, then the rule is minmax. Another useful technique for obtaining minmax rules makes use of invariance properties of the decision problem. If the model and loss are invariant with respect to a group of transformations, and that group satisfies a condition called amenability, then the best equivariant procedure is minmax by the Hunt–Stein theorem. These

techniques are discussed in Ferguson (1967) and Berger (1985).

If Bayes and minmax rules cannot be obtained analytically, computational methods can sometimes be useful. Recently developed simulation methods such as Markov chain Monte Carlo have greatly expanded the range of settings where Bayes rules can be numerically computed. Chamberlain (2000) develops algorithms for computing minmax rules, and applies them to an estimation problem for a dynamic panel data model.

Asymptotic Statistical Decision Theory

Despite advances in computational methods, many statistical decision problems remain intractable. In such cases, large-sample approximations may be used to show that certain rules are approximately optimal. Le Cam (1972, 1986) proposed to approximate complex statistical decision problems by simpler ones, in which optimal decision rules can be calculated relatively easily. One then finds sequences of rules in the original problem that approach the optimal rule in the limiting version of the problem.

As an example, suppose we observe n i.i.d. draws from a distribution P_θ, where $\theta \in \Theta \subset \mathbb{R}^k$ and the probability measures $\{P_\theta\}$ satisfy conventional regularity conditions with non-singular Fisher information I_θ. We can think of this as defining a sequence of experiments, where the nth experiment consists of observing an $n$dimensional random vector distributed according to P_θ^n, the n-fold product of P_θ. Since, in the limit, θ can be determined exactly, we fix a centring value θ_0, and reparametrize the model in terms of local alternatives $\theta_0 + h/\sqrt{n}$, for $h \in \mathbb{R}^k$. This sequence of experiments has as its 'limit experiment' the experiment consisting of observing a single draw $Z \sim N\left(h, I_{\theta_0}^{-1}\right)$, and we say that the original sequence of experiments satisfies local asymptotic normality (LAN). More precisely, according to an asymptotic representation theorem (see van der Vaart 1991), for any sequence of procedures δn in the original experiments that converge in distribution under every local parameter h, these limit distributions are matched by the distributions associated with some randomized procedure $\delta(Z)$ in the limit experiment. Thus, the limit experiment characterizes the set of attainable limit distributions of procedures in the original sequence of experiments. Solving the decision problem in the limit experiment leads to bounds on the best possible asymptotic behaviour of procedures in the original problem, and often suggests the form of asymptotically optimal procedures.

Le Cam's theory underlies the classic result that in regular parametric models, Bayes and maximum likelihood point estimators of θ are 'asymptotically efficient'. In the LAN limit experiment $Z \sim N\left(h, I_{\theta_0}^{-1}\right)$, a natural estimator for the parameter h is $\delta(Z) = Z$. This can be shown to be minmax and best equivariant for 'bowl-shaped' loss functions. Both the Bayes and MLE estimators in the original problem are matched asymptotically by this optimal estimator, so they are locally asymptotically minmax and best equivariant. The ideas have been extended to models with an infinite-dimensional parameter space (see Bickel et al. (1993) and van der Vaart 1991, among others), to obtain semiparametric efficiency bounds for finite-dimensional subparameters. More recently, a body of work has developed limit experiment theory for nonparametric problems such as nonparametric regression and nonparametric density estimation (see Brown and Low 1996, and Nussbaum 1996, among others). These results show that nonparametric regression and density estimation are asymptotically equivalent to a white-noise model with drift, for which a number of optimality results are available.

Applications in Economics

Portfolio Choice

A number of authors have used statistical decision theory to study portfolio allocation when the distribution of returns is uncertain. Some examples

include Klein and Bawa (1976), Kandel and Stambaugh (1996), and Barberis (2000), who develop Bayes rules for portfolio choice problems.

Treatment Choice

Another econometric application of statistical decision theory is to treatment assignment problems, in which a social planner wishes to assign individuals to different treatments (for example, different job training programmes) to maximize some measure of social welfare. Manski (2004) develops minmax-regret results for the treatment assignment problem, Dehejia (2005) develops Bayesian rules, and Hirano and Porter (2005) obtain asymptotic minmax regret-risk bounds and show that certain simple rules are optimal according to this criterion.

Model Uncertainty and Macroeconomic Policy

Brainard (1967) studied a macroeconomic policy problem, in which a parameter describing the effect of a policy instrument on a macroeconomic outcome is not known with certainty but is given a distribution. The policymaker has a utility function over outcomes and chooses the policy that makes expected utility. More recently, a number of authors have continued this line of work, extending the analysis to more general forms of model uncertainty and developing both Bayesian and minmax solutions. Some examples include Hansen and Sargent (2001), Rudebusch (2001), Onatski and Stock (2002), Giannoni (2002), and Brock, Durlauf and West (2003).

Instrumental Variables Models

Decision-theoretic ideas underlie recent work on the linear instrumental variables model in econometrics. Chamberlain (2005) develops minmax optimal point estimators in the IV model using invariance arguments. Andrews, Moreira and Stock (2004) have developed tests in the IV model that are optimal under an invariance restriction, and Chioda and Jansson (2004) have developed optimal conditional tests.

Time Series Models

Asymptotic statistical decision theory has been useful in studying certain time series models which do not satisfy standard regularity conditions. Jeganathan (1995) shows that a number of models for econometric time series have limit experiments that are not of the standard LAN form, but are locally asymptotically mixed normal (LAMN) or locally asymptotically quadratic (LAQ). Ploberger (2004) obtains a complete class theorem for hypothesis tests in the LAQ case, which nests the LAMN and LAN cases.

Auction and Search Models

Some parametric auction and search models, in which the support of the data depends on some of the model parameters, do not satisfy the LAN regularity conditions. For such models, Hirano and Porter (2003) showed that the maximum likelihood point estimator is not generally optimal in the local asymptotic minmax sense, but that Bayes estimators are asymptotically efficient.

See Also

- ▶ Bayesian Econometrics
- ▶ Markov Chain Monte Carlo Methods
- ▶ Maximum Likelihood
- ▶ Savage, Leonard J. (Jimmie) (1917–1971)
- ▶ Wald, Abraham (1902–1950)

Bibliography

Andrews, D.W.K., Moreira, M.M. and Stock, J.H. 2004. Optimal invariant similar tests for instrumental variables regression. Discussion Paper No. 1476. Cowles Foundation, Yale University.

Barberis, N.C. 2000. Investing for the long run when returns are predictable. *Journal of Finance* 55: 225–264.

Berger, J.O. 1985. *Statistical decision theory and Bayesian analysis*. New York: Springer.

Bickel, P.J., C.A. Klaasen, Y. Ritov, and J.A. Wellner. 1993. *Efficient and adaptive estimation for semiparametric models*. New York: Springer.

Brainard, W.C. 1967. Uncertainty and the effectiveness of policy. *American Economic Review* 57: 411–425.

Brock, W.A., S.N. Durlauf, and K.D. West. 2003. Policy evaluation in uncertain economic environments. *Brookings Papers on Economic Activity* 2003(1): 235–322.
Brown, L.D., and M.G. Low. 1996. Asymptotic equivalence of nonparametric regression and white noise. *Annals of Statistics* 24: 2384–2398.
Chamberlain, G. 2000. Econometric applications of maxmin expected utility. *Journal of Applied Econometrics* 15: 625–644.
Chamberlain, G. 2005. Decision theory applied to an instrumental variables model. Working paper, Harvard University.
Chioda, L. and Jansson, M. 2004. Optimal conditional inference for instrumental variables regression. Working paper, UC Berkeley.
Dehejia, R.H. 2005. Program evaluation as a decision problem. *Journal of Econometrics* 125: 141–173.
Ferguson, T.S. 1967. *Mathematical statistics: A decision theoretic approach*. New York: Academic Press.
Giannoni, M.P. 2002. Does model uncertainty justify caution? Robust optimal monetary policy in a forward-looking model. *Macroeconomic Dynamics* 6(1): 111–144.
Gilboa, I., and D. Schmeidler. 1989. Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics* 18: 141–153.
Hansen, L.P., and T.J. Sargent. 2001. Acknowledging misspecification in macroeconomic theory. *Review of Economic Dynamics* 4: 519–535.
Hirano, K., and J. Porter. 2003. Asymptotic efficiency in parametric structural models with parameter-dependent support. *Econometrica* 71: 1307–1338.
Hirano, K. and Porter, J. 2005. Asymptotics for statistical treatment rules. Working paper, University of Arizona.
Jeganathan, P. 1995. Some aspects of asymptotic theory with applications to time series models. *Econometric Theory* 11: 818–887.
Kandel, S., and R.F. Stambaugh. 1996. On the predictability of stock returns: An asset-allocation perspective. *Journal of Finance* 51: 385–424.
Klein, R.W., and V.S. Bawa. 1976. The effect of estimation risk on optimal portfolio choice. *Journal of Financial Economics* 3: 215–231.
Le Cam, L. 1972. Limits of experiments. *Proceedings of the Sixth Berkeley Symposium of Mathematical Statistics* 1: 245–261.
Le Cam, L. 1986. *Asymptotic methods in statistical decision theory*. New York: Springer.
Manski, C.F. 2004. Statistical treatment rules for heterogeneous populations. *Econometrica* 72: 1221–1246.
Nussbaum, M. 1996. Asymptotic equivalence of density estimation and Gaussian white noise. *Annals of Statistics* 24: 2399–2430.
Onatski, A., and J.H. Stock. 2002. Robust monetary policy under model uncertainty in a small model of the U.S. economy. *Macroeconomic Dynamics* 6(1): 85–110.
Ploberger, W. 2004. A complete class of tests when the likelihood is locally asymptotically quadratic. *Journal of Econometrics* 118: 67–94.
Rudebusch, G.D. 2001. Is the fed too timid? Monetary policy in an uncertain world. *Review of Economics and Statistics* 83: 203–217.
Savage, L.J. 1951. The theory of statistical decision. *Journal of the American Statistical Association* 46: 55–67.
Savage, L.J. 1954. *The foundations of statistics*. New York: Wiley.
van der Vaart, A.W. 1991. An asymptotic representation theorem. *International Statistical Review* 59: 97–121.
Wald, A. 1939. Contributions to the theory of statistical estimation and testing hypotheses. *Annals of Mathematical Statistics* 10: 299–326.
Wald, A. 1950. *Statistical decision functions*. New York: Wiley.

Declining Industries

Lester C. Thurow

Logically, there are two meanings to the term declining industries. Industries can decline because their products have been replaced by new and better products, or industries can decline because what used to be most cheaply produced in country A is now most cheaply produced in country B and exported to country A. In the first case, the word processor replaces the typewriter. In the second case, steel production moves from the United States to Brazil and American needs are met with imports from Brazil.

In economic discussions the term declining industries is almost always used in conjunction with the shift of industries from one country to another. This occurs because there is little public controversy about the first type of decline and much public controversy about the second.

With a shift from one product to another it is immediately obvious to everyone that to prevent such declines is to hold one's standard of living below where it otherwise would be. New products and the better jobs that go with them have to be held back to maintain a market for old products

and old jobs. To do so is to retard progress and no one seriously proposes such actions.

It is equally true that to prevent the second type of decline is to hold one's standard of living below where it might otherwise be, but this conclusion is not as immediately obvious. Everyone can see in the first type of decline that additional new jobs serve as a counterbalance to the loss of old jobs and that the consumer get a better product. In the second type of decline the lost jobs are politically visible at home and the new jobs are politically invisible abroad. The home gain in real income comes via lower costs for consumers who replace expensive domestic products with cheap foreign products.

Most often the producers who lose their jobs suffer large immediate reductions in their incomes but are small in number, while the consumers are large in number but reap only small gains in their real incomes. The aggregate gains exceed the aggregate loss but the losses are highly visible while the gains are so small on a per capita basis as to be almost invisible politically. Combine this with a world where producer interests almost always have more political clout than consumer interests, and you have the political ingredients for policies to protect declining industries despite the fact that a country lowers its rate of growth by so doing.

Almost all countries protect their declining industries to some extent. Steel, for example, benefits from various forms of protection in Europe, the United States and Japan since none of them is today the low cost producer for basic steel products. The more extensive the protection, however, the more harm a country does to its economic future.

The pattern of events is well known. Given protection in the home market, cheap foreign producers first drive the home industry out of its unprotected export markets. After World War II, the American steel industry first lost its export markets. Without those export markets home production falls. The home producers of unsophisticated metal products then find that they cannot compete against foreign producers who can buy cheap foreign steel while they have to buy expensive domestic steel. Products such as nails and wire start to be produced abroad and imported into the United States. Home production again falls. Eventually, foreign producers of sophisticated metal-using products such as cars find that their lower cost of materials is one of their advantages in competing against the American auto industry with its high material costs. The steel that is not exported as steel is exported as cars. As the case of steel indicates, protection can serve to slow down the rate of decline, but it is almost never possible to stop it.

To protect a declining industry is to weaken related industries and set in motion spreading waves of decline and protection. As a result, protecting declining industries is much like poking a balloon: for every successful indentation there is an equal expansion somewhere else.

While it is clear that a country should not seek to delay declines in industries where comparative advantage has shifted abroad, it is often not clear as to whether comparative advantage really has shifted. This occurs since currency values have not moved smoothly to maintain national balances between exports and imports as they should have done if they had operated as expected from textbook models. They have often in the past 15 years given very misleading signals – and very rapidly changing signals – as to where a country's real comparative advantage lies.

Thus in February 1985 the value of the dollar was so high that foreign wheat could be sold for less in the United States than American wheat; yet it is clear that the United States still has a comparative advantage in the production of wheat. It just does not seem to be so because of the temporarily high value of the dollar and the markets, such as the Common Market, that have rules and regulations essentially closing them to American exports.

Since the transition costs of closing an industry when the value of the dollar is high and reopening the industry when the dollar falls are very large, it may not make sense to allow the market to operate as it would without government interference. The question then becomes one of whether the right solution is protection or subsidies for the affected domestic industries, or international actions to moderate the movements between major

currencies and to open closed foreign markets. Given that protection once in place is difficult to remove politically, international actions to moderate currency movements and open markets would seem to be the preferable solution.

When one analyses a declining industry, one seldom finds an industry in total decline without competitively viable parts. In the steel industry, for example, there are parts – mini-steel mills using electric furnaces and low cost scrap iron, speciality high-tech alloy steels – that could be competitively operated in the United States given a value of the dollar that would balance exports and imports. To say that an industry is a declining industry is not to say that it will disappear.

A declining industry also need not lead to declining firms. While it is certainly true that modern industrial economies need less steel per unit of GNP produced, it is also true that there is a new growing high-tech industry in new materials – powdered metals, composites, pressed graphite – that is the new steel industry of tomorrow. Today's declining steel firms could be tomorrow's expanding new material firms. But most often they are not.

If one asks why not, it is clear that firms find it very difficult to develop new products that will destroy large old markets that they dominate. The firm has a large vested interest in the old markets and entrenched forces within the firm make it very difficult for it to move into these new areas quickly. Thus IBM, the dominant force in the office typewriter business, was slow to develop a word processor despite the fact that it was the world's leader in computers. At General Electric, the dominant vacuum tube division sat on the transistor and prevented General Electric from becoming a leader in transistors. The classic example is of course the railroads, which saw themselves as railroads rather than as transportation companies.

While decline is the flip side of progress, real costs are involved. Most of these costs come in the form of human resources that are not easily transferred to new areas. An unemployed 55-year-old Pennsylvania steel worker is not apt to be retrained to be a California computer assembler. Such an individual faces a large cut in expected income over the remainder of his working life and society may well find itself burdened with higher social welfare costs.

Economic theory has little to say about these transition problems and costs since it assumes that mobility is easy and that transition costs either do not exist or are very marginal. With its concept of equilibrium, wage workers forced out of work in old industries quickly find jobs in new industries with closely comparable wages. In contrast, those who have actually followed workers forced out of work in declining industries in the United States find that most of them find work only with a long time lag and then only with much lower wages. The losses in real incomes are not the marginal ones assumed by economic theory.

As a result there is a real issue in how a nation manages decline. A nation cannot and should not prevent declining industries from shrinking, but it still has to face the issue of how it manages the transition of human resources from old sunset industries to new sunrise industries and what it does about those human resources that are essentially junked in the transition.

See Also

- ▶ Manufacturing and De-industrialization
- ▶ Verdoorn's Law

Bibliography

Borrus, M, Millstein, J. and Zysman, J. 1982. U.S. Japanese competition in the semiconductor industry. *Policy Papers in International Affairs,* No. 17. Berkeley: Institute of International Studies/University of California.

Eckstein, O., C. Caton, R. Brinner, and P. Duprey. 1984. *The DRI report on U.S. manufacturing industries*. New York: McGraw-Hill.

Hatsopoulos, G.N. 1983. *High cost of capital: Handicap of American industry.* Waltham: American Business Conference/Thermo Electron Corp.

Konaga, K. 1983. Industrial policy: The Japanese version of a universal trend. *Journal of Japanese Trade and Industry* 4: 21.

Krist, W.K. 1984. The U.S. response to foreign industrial policies. In *High technology public policies for the 1980s*. Washington, DC: A National Journal Issues Book.

Krugman, P. 1984. The United States response to foreign industrial targeting. *Brookings Papers on Economic Activity* 1: 77–121.
Labor-Industry Coalition for International Trade. 1983. *International trade, industrial policies, and the future of American industry*. Washington, DC: Labor-Industry Coalition for International Trade, 40f. Lawrence, R.A. 1984. *Can America compete?*. Washington, DC: Brookings.
Magaziner, I. 1983. New policies for wealth creation in the United States. In *Growth with fairness*. Washington, DC: Institute on Taxation and Economic Policy.
McKenna, R., M. Borrus, and S. Cohen. 1984. Industrial policy and international competition in high technology – Part I: Blocking capital formation. *California Management Review* 26(6): 15–32.
Melman, S. 1984. The high-tech dream won't come true. *INC Magazine*.
Office of Technology Assessment. 1981. *US industrial competitiveness: A comparison of steel, electronics and automobiles*. Washington, DC: Congress of the US/ Office of Technology Assessment.
Phillips, K. 1984. *Staying on top: The business case for a national industrial strategy*. New York: Random House.
Piore, M.J. 1982. American labor and the industrial crisis. *Challenge* 25: 5–11.
Reich, R.B. 1982. Why the United States needs an industrial policy. *Harvard Business Review* 60(1): 74–80.
Schultze, C. 1983. Industrial policy: A dissent. *The Brookings Review* 2: 3–12.
Zysman, J. 1983. *Governments, markets, and growth*. Ithaca: Cornell University Press.

Declining Population

Robin Barlow

Population decline is much less common than population growth. Looking at the geographical areas occupied by present-day nations, or by their administrative subdivisions, one sees that over the last millenium the number of years when the human population declined is almost always much exceeded by the number when it grew. Reflecting this fact, economics has devoted much more attention to the growth of population than to its decline. The preoccupation with growth, however, may be ending as more countries experience lengthy periods of reduced fertility.

Many of the economists writing on the growth of population, from Malthus to the Club of Rome, are notorious for their bleak view of the future. If population growth is a bad thing, one might be excused for thinking that its decline might be beneficial. But much of the writing on decline is equally alarmist. Does this indicate a general tendency towards pessimism in demographic commentary? Or is the model used for analysing the consequences of population change genuinely asymmetrical, in the sense that increases and decreases of population do not produce opposite effects? Or are different models being used for growth and decline?

Before considering the consequences of population decline, it is desirable first to consider the causes, because in many respects the consequences are conditioned by the causes. The population of a given geographical area can decrease because of a reduction in fertility, an increase in mortality or an increase in net emigration. Of these three factors, fertility reduction has had the least importance as a historical cause of depopulation. Most areas in the world have indeed experienced prolonged periods of fertility decline, particularly within the past 200 years, but these declines have normally been accompanied by significant reductions in mortality, and indeed many would argue that the fall in fertility has been partly a consequence of the fall in mortality, particularly infant mortality. The result has been that populations have continued to grow even when the total fertility rate (the number of live childbirths per woman during the childbearing period, assuming age-specific birth rates to stay at their current levels) has been reduced by as much as 75 per cent, from a 'traditional' level of about eight to a 'modern' level of about two.

Of course, when the total fertility rate falls below the long-run replacement level, which in modern conditions of mortality is about 2.1 children per woman, the population must eventually diminish, in the absence of net immigration. However, some decades may elapse between the decline of the total fertility rate below this critical level and the subsequent decline of the population, because a pyramidal age-structure inherited from earlier regimes of high fertility can sustain

the absolute number of births at a high level for several years even while age-specific birth rates are falling. In the United States, for example, the total fertility rate has been below 2.1 since 1972, but in 1984 the annual number of births was still 80 per cent greater than the number of deaths.

Increases in death rates, on the other hand, have often been so extreme and abrupt as to produce an immediate decline in population. Historically there have been three main causes of sudden increases in mortality: famine, disease and war. The three causes are not unrelated to each other. War has often caused famine, for example, and famine has caused disease. Famine, besides sometimes resulting from war and other political disorder, has been the product of natural disasters like drought and floods. Cases when disease has caused sudden increases in mortality include epidemics, like the bubonic plague in medieval Europe, and the importing of new infections into populations without immunity. A classic example of the latter is the decline of American Indian populations after their encounter with the measles, influenza, tuberculosis and other diseases brought by Europeans.

Regarding the future likelihood of these catastrophic causes of population decline, it is not easy to be optimistic, because our own 20th century provides numerous examples of such catastrophe. There have been large-scale famines leading to extensive depopulation. Probably the worst was the Chinese famine of 1959–61, caused by natural disasters and the dislocations of the Great Leap Forward. It is thought that in those years, 30 million deaths took place because of starvation (Banister and Kincannon 1984). In the 1980s certain regions in Ethiopia and the African Sahel have been depopulated for similar reasons. As for disease, some 20th-century epidemics have reached vast proportions, in particular the influenza epidemic of 1918–19, which took 20 million lives worldwide. War and armed conflict have had even more serious depopulating effects in this century than earlier, as warring states and factions have increasingly resorted to the mass extermination of civilians. Large areas of Russia and Poland suffered population declines for this reason between 1941 and 1945, and similar declines are alleged to have occurred elsewhere during the century (Cambodia, Armenia, Uganda, Punjab).

The third cause of population decline, net emigration, is frequently encountered, but unlike mortality increases, it can often be regarded as benign. Emigration occurs in response to 'push' factors or 'pull' factors. In any individual case it is often difficult to tell whether 'push' or 'pull' is stronger, but it is certainly safe to say that in many instances, the decision to emigrate should be seen as a hopeful determination to explore new opportunities rather than as an escape from distress. Indeed, in a dynamic, expanding economy, it is to be expected that changes in demand and technology will shift the comparative advantages and disadvantages of particular regions, and that some regions will lose population to others as labour markets respond to these shifts.

At the regional or sub-national level, net emigration is often substantial enough to produce an actual decline in population. For example, in the United States between 1980 and 1983, four of the 50 states lost population, even though in all states the number of deaths during that period was less than the number of births. At the national level, net emigration is less commonly a cause of depopulation, largely because of the legal and other obstacles to international migration.

We turn now to the consequences of population decline, which, as noted above, will be found to vary according to the cause of the decline. The consequences of decline have been investigated with particular thoroughness in France, where the subject has been a matter of active political and academic discussion since the defeat of France by a more populous Germany in the war of 1870–71. In general, the tendency of the French population to stagnate has been deplored. A typical statement is found in the preamble to the Family Code of 1939, a set of pro-natalist measures adopted by the Daladier government on the eve of World War II (cited by Tomlinson et al. 1985):

> Our military and economic forces are in danger of wasting away; the country is ruining itself little by little; by contrast, the individual tax burden is increasing the whole time; each citizen is having to pay more to support the social welfare system; industry is gradually deprived of its market; land

> remains untilled; overseas expansion loses its momentum; and beyond our frontiers, our intellectual and artistic prestige is extinguished.

There are three themes in this bleak picture which have remained important in demographic analysis and which deserve further comment here: the increased burden of dependency said to result from a declining population, the weakening of military forces and the fall in aggregate demand.

The burden-of-dependency argument contends that in a declining population there is an increase in the ratio of dependants to workers. This causes heavier burdens on workers, both because of the increased taxes they must pay to finance public services provided to the dependent part of the population, and because of the increased levels of private consumption they must support. A fall in the rate of saving is the probable result. But there are some qualifications which should be made to this argument. First, if the population decline is due to the emigation of young adult males – a not untypical situation – there may well be an increase in the ratio of dependants to non-emigrant workers, but no corresponding additional burden on nonemigrants, since the dependants of emigrants will be supported in part by remittances.

Second, if the population decline is caused by a reduction in fertility, the rising fraction of elderly in the population will be at least partly offset by a diminishing fraction of children, with little change occurring in the ratio between all dependants and all workers (except in the very long run). The American case is illustrative. Between 1960, near the start of the current fertility decline in the United States, and 1983, the fraction of the population aged 65 or over rose from 9 per cent to 12 per cent, but the fraction aged under 18 fell from 36 per cent to 27 per cent, so that the fraction aged 18–64 actually rose from 55 per cent to 61 per cent. These numbers may even understate the real reduction in dependency burdens occurring during this period, since the fertility decline facilitated an increase in labour-force participation rates among females, reducing still further the number of dependants per worker.

While fertility declines like those occurring in the United States may not lead to much change in the ratio between all dependants and all workers, they certainly produce changes in the structure of dependency. Whether these structural changes lead to an additional fiscal burden on workers depends on the relative costs of public services for the elderly (pensions, health care) and those for children (education).

A third qualification which should be made to the burden-of-dependency argument is as follows: to the extent that the elderly finance their own consumption out of earlier saving, undertaken through a funded pension scheme or otherwise, their presence does not constitute an economic burden. For this reason and others, there is much complexity in the 'economics of aging populations', which has become an area of active enquiry in Europe and elsewhere as anxieties have developed on such issues as the future financing of social security.

The military implications of population decline do not seem very clear, despite what French strategists have argued. A country can gain the upper hand over a more populous adversary by conscripting a larger fraction of its population, by possessing more advanced weaponry, by receiving assistance from allies, or by any of several other methods. In the 20th century there is no shortage of examples of smaller countries defeating larger (Japan against Russia in 1904, Germany against Russia in 1917, Japan against China in 1937, Israel against Egypt in 1967, Vietnam against the United States in 1975).

The aggregate-demand argument is Keynesian in nature, and suggests that in a declining population, there will be large reductions in demand for certain kinds of investment goods and consumption goods (e.g. housing and children's clothing). Weak demand in these markets could lead to a deficiency of aggregate demand and to an equilibrium with considerable unemployment. However, if there is a Keynesian problem of this nature, a Keynesian solution could also exist. Expansionary fiscal and monetary measures could in principle restore aggregate demand to its full-employment level.

There are other elements in the case against a declining population, a case developed in recent

years with particular vigour by Alfred Sauvy and Julian Simon (see, for example, Dumont and Sauvy 1984; Simon 1981). Many of these elements are difficult to evaluate, since they concern the allegedly deleterious effects of depopulation on certain intangible characteristics of a society that are not easily measured – such as the dynamism of its artists, or its spirit of adventure, or its readiness to innovate. Also difficult to evaluate is the 'Beethoven–Einstein' argument, which says that a smaller population has a smaller probability of producing a great genius. (If that is true, perhaps such a population is also less likely to produce an evil genius on the scale of Hitler.)

Generally absent from the alarmist views on population decline is the admission that decline does have some beneficial tendencies. These may indeed be swamped by the undoubted negative tendencies, but not necessarily so. Perhaps the most powerful benefit of population decline is its immediately favourable effect on the ratio between physical resources and the labour force. In the short run, the stock of natural resources and capital is fixed, and so any reduction in labour inputs will raise the ratio of natural resources to labour, the ratio of capital to labour, the marginal product of labour, and most probably the wage rate. In the longer run, what happens to the capital–labour ratio when the labour force is diminishing is more difficult to say: the outcome depends among other things on what is happening to dependency burdens and the rate of saving. But even in the longer run, the stock of many types of utilized natural resources will be practically independent of the size of the labour force, and to that extent a smaller labour force is likely to mean a higher income per capita. To make this point, it suffices to look at the economies of Kuwait and Nigeria, which in recent years have produced roughly the same substantial volume of crude oil. But Kuwait's population is only two per cent of Nigeria's, and largely in consequence, its per capita income is about 20 times higher.

The reasoning here is the same as that employed in standard neoclassical models of migration. It is assumed that higher wages in one area will attract migrants; this movement will lower the marginal product of labour in the area of destination and raise it in the area of origin, thus narrowing wage differentials and leading to an equilibrium rate of migration. The point of interest in the present context is that declines in the labour force tend to raise output per worker, certainly in the short run and perhaps in the long run as well.

Closely related to these economic benefits from depopulation are some environmental benefits. The increase in natural resources per capita which tends to raise income per capita also tends to alleviate problems like air and water pollution, the rapid depletion of mineral resources, urban congestion and excessive use of recreational space. The environmental advantages of smaller populations have been one of the main themes of contemporary anti-natalist movements like Zero Population Growth.

In sum, it is not difficult to think of benefits as well as costs of population decline. In many of the countries now facing population decline as a result of their recent fertility history, the benefits and costs are regarded as fairly evenly balanced, or at least, 'the sense of urgency over population decline is still far from acute' (McIntosh 1981). According to the World Bank (1984), there were 22 countries which in 1982 had a total fertility rate less than 2.1. Seventeen of these were high-income OECD countries, three were East European (East Germany, Hungary and Yugoslavia), and the others were Cuba and Singapore. In some of these countries, like France and Hungary, there is considerable anxiety about depopulation. But in others, many people seem to feel that 'smaller is better'.

See Also

- ▶ Ageing Populations
- ▶ Demographic Transition
- ▶ Social Security
- ▶ Stagnation

References

Banister, J., and L. Kincannon. 1984. Perspectives on China's 1982 census. Paper presented at the International Seminar on China's 1982 Population Census, Beijing.

Dumont, G.F., and A. Sauvy. 1984. *La montée des déséquilibres démographiques: quel avenir pour une France vieillie dans un monde jeune?* Paris: Economica.
McIntosh, C.A. 1981. Low fertility and liberal democracy in Western Europe. *Population and Development Review* 7(2): 181–207.
Simon, J. 1981. *The ultimate resource*. Princeton: Princeton University Press.
Tomlinson, R., M.M. Huss, and P.E. Ogden. 1985. France in peril: The French fear of dénatalité. *History Today* 35: 24–31.
World Bank. 1984. *World development report 1984*. New York: Oxford University Press.

Default and Enforcement Constraints

Fabrizio Perri

Abstract

This article illustrates when limited enforcement of contracts induces enforcement constraints (limits to intertemporal exchange) or default (the breaking of intertemporal promises with the associated punishment), and sheds light on how enforcement policies should be related to the observed frequency of default. When limited enforcement is the only friction equilibrium default is never observed, yet tightening enforcement of contracts is socially beneficial. When limited enforcement coexists with other frictions, default occurs in equilibrium but tightening enforcement might be socially undesirable. The reason is that equilibrium default, although detrimental to intertemporal exchange, might lead to improved allocation of resources across states.

Keywords

Arrow–Debreu promises; Default; Enforcement constraints; Intertemporal exchange; Lagrange multipliers; Limited enforcement of contracts with default; Limited enforcement of contracts without default; Risk sharing

JEL Classification

D4; D10

Intertemporal exchange, that is the exchange of resources today for a promise of resources at a later date in a given state, is key for promoting economic efficiency. For example, to finance an investment, a government borrows capital abroad in exchange for a promise of repayment once the investment has paid off. Or, to finance consumption, an individual who loses her job borrows resources in exchange for the promise of repayment once she gets a new job. If the enforcement of promises is limited, the extent of intertemporal exchange can be reduced by so-called enforcement constraints and, under some conditions, *default*, that is, the breaking of promises, can arise. This article presents a simple general equilibrium set-up to analyse these issues and provide some direction for the design of enforcement policies. Key references for the theory of limited enforcement without default are Kehoe and Levine (1993), Kocherlakota (1996) and Alvarez and Jermann (2000), while for limited enforcement with default see Zame (1993) and Dubey et al. (2005).

The Set-Up

The goal of this set-up is to capture the need for intertemporal exchange, as described in the examples above. There are two agents which live for two periods and consume a single good. Agent 1, the borrower, owns a technology such that, if k units of the good are invested in period 1, AK^{α}, $0 < \alpha < 1$, units are produced in period 2, where A is a random variable realized in period 2, with positive support and distribution $F(A)$ known to both agents. Agent 2, the lender, is endowed with e units of the consumption good in period 1. Consumption allocations of agent i are consumption at date 1, c_{i1} and the function $c_{i2}(\mathrm{A})$ which assigns period 2 consumption for each possible realization of A. Borrower's utility is given by $u(c_{11}) + \int u(c_{12}(A))dF(A)$ where u is a concave utility function satisfying Inada conditions. The lender has linear utility given by $c_{21} + \int c_{22}(A)dF(A)$. Linear utility implies that lender's equilibrium utility is constant across different market structures so that borrower's utility is the only statistic needed to Pareto-rank

equilibria. In all the economies described below the following resource constraints hold

$$c_{11} + c_{21} + k = ec_{12}(A) = Ak^{\alpha} \text{ for every } A$$

A Frictionless Benchmark

Assume agents can trade a complete set of Arrow–Debreu promises which are fully and costlessly enforceable. The budget constraints of the borrower are

$$c_{11} + k = \int p(A)dF(A) \tag{1}$$

$$c_{12}(A) = Ak^{\alpha} - p(A) \text{ for every } A \tag{2}$$

where $p(A)$ denotes the amount that the borrower promises to repay in state A. Equilibrium allocations display complete risk sharing, that is, the ratio of marginal value of consumption of the two agents is constant across dates and states of the world. We denote with c^{AD} the constant, across dates and states, level of consumption of the borrower in this economy.

Limited Enforcement

This section describes an economy denoted as ADLE (Arrow–Debreu Limited Enforcement) and shows that limited enforcement prevents full risk sharing, reduces investment and welfare. Assume that in period 2 the borrower can walk away from any promise made to the lender by suffering a default deadweight cost proportional to its output and equal to δAk^{α} where $\delta > 0$ is a parameter that measures the strength of enforcement. This implies that any Arrow–Debreu promise $p(A) > \delta Ak^{\alpha}$ will not be honoured by the borrower and thus will not be purchased by the lender. Also, promises satisfying $p(A) \leq \delta Ak^{\alpha}$ will be fully honoured and priced as in the frictionless economy. So limited enforcement limits the use of state-contingent promises but does not induce default. A convenient way of capturing this, following Alvarez and Jermann (2000), is to assume that the borrower faces constraints on the sales of each promise so as to guarantee no default. These *enforcement* constraints have the form

$$p(A) \leq \delta Ak^{\alpha} \text{ for every } A \tag{3}$$

as the borrower can sell each promise only up to the point where the cost of keeping it is equal to the cost of defaulting on it. Equilibrium allocations can be characterized by substituting budget constraints (1) and (2) into the borrower's utility and taking first-order conditions with respect to k and p(A) subject to constraints (3). This yields

$$u'(c_{11}) = \int \left[A\alpha k^{\alpha-1}u'(c_{12}(A)) + A\alpha k^{\alpha-1}\delta\mu(A)\right] dF(A) \tag{4}$$

where

$$\mu(A) = u'(c_{11}) - u'(c_{12}(A))$$

are the Lagrange multipliers on the enforcement constraints. If the cost of default δ is sufficiently small and the distribution of A is sufficiently spread out, $c_{11} = c(A) = c^{AD}$ is not a solution of (4) as enforcement constraints on the high A promises would be violated. The solution is then characterized by a level of productivity A^* such that for all $A > A^*$ enforcement constraints are binding and $c(A) = (1-\delta)Ak^{\alpha} > c_{11}$. For $A \leq A*$ enforcement constraints are not binding and $c(A) = c_{11} < c^{AD}$. Complete risk sharing involves the borrower selling promises to repay in states with high A, in order to finance consumption today (when she has no output) and consumption tomorrow in states with low A. But if the distribution of A is spread out, complete risk sharing calls for promises of a large transfer of resources from the borrower to the lender in the states with high A. When enforcement is limited (δ is low) the lender, in period 1, correctly anticipates that these transfers will not be made and buys a smaller amount of the promises. So, relative to complete risk sharing, the borrower has fewer resources in period 1 and in the period 2 states with low A, but consumes more in period 2 states high A. This allocation of consumption

D

increases the marginal value of resources in period 1 relative to the expected marginal value of resources in period 2 and thus reduces k relative to the full enforcement case. Finally, equilibria in economies with strong enforcement (high δ) Pareto-dominate equilibria with weak enforcement (low δ). To see this, note that, for the borrower, the equilibrium allocation in the weak enforcement economy is budget-feasible in the strong enforcement economy, so, if it is not chosen, it must yield her lower utility.

ADLE economies have been used extensively in a variety of applications such as asset pricing (Alvarez and Jermann 2000), international business cycles (Kehoe and Perri 2002) and consumption inequality (Krueger and Perri 2006). All these studies show that limited enforcement prevents complete risk sharing, and for this reason it provides a much better fit with the data than standard Arrow–Debreu economies. This environment, though, cannot be used to understand equilibrium default (that is, the actual break of a promise and the suffering of the associated cost) as the trade in contingent promises makes incurring the default cost unnecessary. In order to understand when default arises and what its consequences are, the next section considers an economy in which contingent promises cannot be traded, either because markets are exogenously missing or because the borrower has private information about realizations of A.

Limited Enforcement and Non-contingent Promises

The borrower finances consumption and investment only by selling a non-contingent promise p which can be defaulted on in state A by suffering the default cost δAk^{α}. Since the cost of repaying the promise does not vary with the state while the default cost is increasing with A, if there is equilibrium default it will happen in the low A states. In particular, if the borrower invests k and sells a promise p, she will default in all the states such that $A \leq \frac{p}{\delta k^{\alpha}}$.

As a consequence, the equilibrium price of the promise is given by

$$q(p,k) = 1 - F\left(\frac{p}{\delta k^{\alpha}}\right). \tag{5}$$

The problem of the borrower is then

$$\max_{p,k} u(q(p,k)p - k) + \int_0^{\frac{p}{\delta k^{\alpha}}} u((1-\delta)Ak^{\alpha})dF(A) + \int_{\frac{p}{\delta k^{\alpha}}}^{\infty} u(Ak^{\alpha} - p)dF(A). \tag{6}$$

The equilibrium is characterized by a couple p, k which solve (5) and (6). It can be immediately shown that equilibria in this economy are, generically, Pareto-inferior to equilibria in the corresponding ADLE economy. Also, for many parameter values equilibria in this set-up differ from those in the ADLE economy along two important dimensions: (*a*) there is a positive measure of states for which default occurs and (b) there is a positive measure of values for δ for which welfare is *decreasing* in the strength of enforcement. As a simple example, consider the case in which A can take only two values: a high value A_h with probability π and a low value A_l with probability $1 = \pi$, with $\pi > A_l/A_h$. In this case there is a range of values for δ for which the equilibrium promise and capital satisfy

$$\delta A_l k^{\alpha} < p < \delta A_h k^{\alpha}, \tag{7}$$

so that default happens only when state A_l is realized and consequently $q(p,k) = \pi$. Now consider the effect of a marginal reduction in δ. Equation (7) shows that, if the borrower kept k and p unchanged in response to the change in δ, default patterns, and hence $q(p, k)$, would not change; however reducing δ increases the returns of borrower in the default state so its utility would increase relative to the initial equilibrium. Here weakening enforcement allows the borrower to implicitly transfer, through default, more resources to the low A state and thus to achieve a better allocation of risk across states. In the ADLE economy this transfer was achieved through the

Arrow–Debreu promises so default was not necessary. When promises cannot be made state-contingent, increasing payoffs in the default states is the only way of obtaining this transfer.

In this simple example weakening enforcement does not affect default frequency, but in more general set-ups it does and as a consequence increases equilibrium interest rates and hampers intertemporal exchange. This effect is detrimental for welfare. But the example above suggests that the detrimental effect can be offset by the positive effect of the better risk allocation across states. Note that this result does not rely on the two-state assumption, and it can be shown to hold, for example, also when A is log-normally distributed.

Summary

Limiting contract enforcement in otherwise frictionless environments constrains intertemporal exchange and hampers risk sharing, investment and welfare, but does not induce default. When additional frictions, such as incomplete markets or private information, limit the span of tradable promises, then limited enforcement can play a positive role by inducing equilibrium default, which can be used as a (costly) way of providing better allocation of risk across states. The analysis sheds light on how enforcement policies should be related to the observed frequency of default.

When limited enforcement is the only friction, default is never observed, yet tightening enforcement is socially beneficial. When limited enforcement coexists with other frictions, default happens in equilibrium but this does not necessarily mean that enforcement should be tightened. Indeed, tightening enforcement without ameliorating the additional friction might reduce default but also risk sharing and welfare.

See Also

► Risk Sharing
► Sovereign Debt

Bibliography

Alvarez, F., and U. Jermann. 2000. Efficiency, equilibrium, and asset pricing with risk of default. *Econometrica* 68: 775–797.

Dubey, P., J. Geanakoplos, and M. Shubik. 2005. Default and punishment in general equilibrium. *Econometrica* 73: 1–37.

Kehoe, T., and D. Levine. 1993. Debt-constrained asset markets. *Review of Economic Studies* 60: 865–888.

Kehoe, P., and F. Perri. 2002. International business cycles with endogenous incomplete markets. *Econometrica* 70: 907–928.

Kocherlakota, N. 1996. Implications of efficient risk sharing without commitment. *Review of Economic Studies* 63: 595–609.

Krueger, D., and F. Perri. 2006. Does income inequality lead to consumption inequality? Evidence and theory. *Review of Economic Studies* 73: 163–193.

Zame, W. 1993. Efficiency and the role of default when securities markets are incomplete. *American Economic Review* 83: 1142–1164.

Defence Economics

Keith Hartley and Martin C. McGuire

Abstract

Defence economics is a new field of economics. Its development and research agenda have reflected current events. Examples include the superpower arms race of the cold war, disarmament following the end of the cold war, international terrorism, peacekeeping and conflict. A brief history is presented; the field is defined and the facts of world military spending are outlined; the defence economics problem, namely, the need for difficult choices, is considered; and conflict and terrorism are used to illustrate some of the new developments in the field.

Keywords

Arms races; Arms trade; Cost-plus contracts; Crowding out; Defence economics; Disarmament costs; Economic theories of military alliances; Ethnicity; European Security and

Defence Policy (EU); Fixed-price contracts; Free rider problem; Game theory; Military employment contract; Military outsourcing; Military wage differential; Military–industrial complex; Nationalism; One-shot games; Principal and agent; Private finance initiatives; Procurement; Public–private partnerships; Purchasing power parity; Religion; Repeated games; Research and development; Strategic behaviour; Substitution effect; Substitution principle; Technology; Terrorism, economics of; Tit for tat; Two world wars, economics of the; Voting paradoxes; War and economics

JEL Classifications
Z0

Defence economics is a relatively new part of the discipline of economics. One of the first specialist contributions in the field was by C. Hitch and R. McKean, *The Economics of Defense in the Nuclear Age* (Hitch and McKean 1960). This book applied basic economic principles of scarcity and choice to national security. It focused on the quantity of resources available for defence and the efficiency with which such resources were used by the military. For example, defence consumes scarce resources that are therefore not available for social welfare spending (for example, missiles versus education and health trade-offs). Once resources are allocated to defence, military commanders have to use them efficiently, combining their limited quantities of arms, personnel and bases to 'produce' security and protection. Within such a military production function, there are opportunities for substitution. For example, capital (weapons) can replace (and have replaced) military personnel; imported arms can replace nationally produced weapons; and nuclear forces have replaced large standing armies. Defence economics is about the application of economic theory to defence-related issues.

The development of defence economics and its research agenda reflected current events. For example, during the cold war there was a focus on the superpower arms races, alliances (NATO and the Warsaw Pact), nuclear weapons and 'mutually assured destruction'. The end of the cold war resulted in research into disarmament, the challenges of conversion and the availability of a peace dividend. Since the end of the cold war, the world remains a dangerous place with regional and ethnic conflicts (for example, Bosnia, Kosovo, Iraq), threats from international terrorism (for example, terrorist attacks on USA on 11 September 2001), rogue states and weapons of mass destruction (that is, biological, chemical and nuclear weapons). NATO has accepted new members (for example, former Warsaw Pact states) and has developed new missions, and the European Union has developed a European Security and Defence Policy. Changing threats and new technology require the armed forces and defence industries to adjust to change and new challenges. Peacekeeping has become a major mission for armed forces and is an example of the trend towards globalization.

The modern era of globalization involves more international transactions in goods, services, technology and factors of production, which brings new security challenges for both nation states and the international community. Defence firms have become international companies with international supply networks. Globalization also highlights the importance of international collective action to respond to new threats such as international terrorism and to maintain world peace (for example, through international peacekeeping missions under UN, NATO or EU control). But international collective action experiences the standard problems of burden-sharing and free riding.

This article outlines the development of defence economics; it defines the field and describes the 'stylized facts' of world military expenditure; the defence economics problem is considered; and a case study of conflict and terrorism illustrates some of the new developments in the field.

A Brief History

Defence issues have existed throughout history as nations have been involved in armed conflict of various forms and durations (for example, the

Hundred Years War). Great powers have used military force to dominate regions and parts of the world (for example, Alexander the Great; Roman legions; Genghis Khan; Ottoman Turks; Nazi Germany), with such powers rising and falling (Kennedy 1988). Conflict has also been characterized by major technical changes ranging from bows and arrows to cannons and machine guns, from sailing ships to iron and steel warships and nuclear- powered vessels, from horse cavalry to tanks, from flag communications to radios and satellite communications and from balloons to aircraft, missiles, nuclear weapons and space systems. Historically, the economic base for conflict was first an agricultural society, then an industrial society followed by a knowledge economy.

Some of the classical economists studied war and conflict (for example, Smith, Ricardo, Malthus, J. S. Mill: see Goodwin 1991, chapter 2). For these economists, war departed from much of their conventional thinking: it involved chaos and disorder rather than market equilibrium, and it required government action rather than private market behaviour. Yet it remains surprising that, with a long history of wars, including two world wars and the superpower arms race of the Cold War, relatively few economists have been attracted to the field. A review of the economics literature on conflict concludes that 'We were surprised at the relative absence of applied economics studies of actual conflicts' (Sandler and Hartley 2003, p. xl). There are various possible explanations for the relative absence of economists studying war and conflict. These include data and security problems, the difficulty of applying conventional market analysis to the chaos and disequilibrium of conflict, a traditional reluctance to analyse the public sector (with defence assumed to be exogenous), and the feeling that defence and security issues are not as important as other social welfare issues, with war viewed as an immoral and unethical subject. Furthermore, security issues have not been as an attractive career path for economists (compared with issues such as inflation, unemployment, growth and developing countries), and conflicts are usually of short duration so that they offer only limited research prospects before peace returns to remove war-related problems (Goodwin 1991, pp. 1–2).

Definitions

Defence economics studies all aspects of war and peace and embraces defence, disarmament and conversion. This definition includes studies of both conventional and non-conventional conflict such as civil wars, revolutions and terrorism. It involves studies of the armed forces and defence industries and the efficiency with which these sectors use scarce resources in providing defence output in the form of peace, protection and security. Reductions in defence spending (such as those following the end of the cold war) result in disarmament, which involves reallocating resources from the defence to the civilian sector. This raises questions about the impact of disarmament on the employment and unemployment of both military personnel and defence industry workers; the possibilities for converting military bases and arms industries to civil uses (the Biblical swords to ploughshares); and the role of public policies in assisting the transition and reallocation of resources.

The coverage of the subject is extensive and involves economic theory, empirical testing and policy-related issues, including applications of public choice analysis. Both defence and peace have distinctive economic characteristics in that they are public goods which are non-rival and non-excludable. There are large literatures dealing with the determinants of military expenditure, including economic theories of military alliances and arms races (that is, threats) and the impact of defence spending on economic growth and development. Armed forces are major buyers of both equipment (arms/weapons) and military personnel, and such procurement choices affect defence industries and both local and national labour markets. For example, government procurement of weapons involves choices between competition and preferential purchasing and between various types of contracts (for example, fixed-price, cost-plus), each with different implications for contractor efficiency and profitability. There is a related literature on industrial and alliance policies

comparing the economics of supporting a national defence industrial base with alternative industrial policies such as international collaboration, licensed production or importing foreign equipment. Imports also involve the international arms trade, its economic impacts on both buyers and suppliers, and policy initiatives to regulate such trade. More generally, there is an extensive literature on arms control and disarmament, the adjustment costs of disarmament, the economics of conversion and the contribution of public policy to minimizing such adjustment costs. Finally, there have been some new developments involving the application of economics to the study of conflict, civil wars, revolutions and terrorism (Brauer 2003; Hegre and Sandler 2002; Sandler and Hartley 1995, 2007).

Defence economics became established in the 1960s with the publication of a number of pioneering contributions, mostly by US economists. These contributions applied economics to some novel areas and included economic models of alliances (Olson and Zeckhauser 1966), the economics of arms races (Richardson 1960; Schelling 1966), the procurement of weapons and military personnel (Peck and Scherer 1962; Oi 1967), and the impact of military spending on economic development (Benoit 1973). A further development confirming the emergence of defence economics as an accepted part of the discipline of economics was the launch in 1990 of a field journal, *Defence Economics*, later renamed *Defence and Peace Economics* (initially it was published four times per year, but in 2000 it was expanded to six issues per year).

Inevitably, defence economics generates controversy reflected in myths and emotion. Critics point to the 'wastes' of defence spending and its 'crowding-out' of 'valuable' civil expenditure. Classic examples include the sacrifice of schools and hospitals associated with major weapons projects such as modern combat aircraft and aircraft carriers (for example, the US F-22 aircraft and the European Typhoon). Peace economists are similarly critical of defence economics and military spending: they focus on peace topics such as disarmament and the maintenance of peace, arms control and international security, conflict analysis and management, and crises and war studies. Defence economists are not, however, 'warmongers': they are instead interested in understanding the economics of the military–industrial–political complex and all aspects of defence whereby a proper understanding of these issues will contribute to a more peaceful world. A starting point in showing how economists analyse defence is to review the 'stylized facts' of world military spending.

The Stylized Facts of World Military Spending

What is known about military spending, and where are the gaps in the data? Good quality data exist on world military spending, the world's armed forces and the arms trade. Cross-section and time-series data are available at the country level; some examples are shown in Table 1. The data on world military expenditure show aggregate spending by the USA accounting for 45% of total world military spending and NATO accounting for some 70%. Similarly, in 2004 the USA dominated defence R&D spending, accounting for some 75% of the world total and 31% of world arms exports.

Table 1 shows examples of defence shares of GDP to illustrate the burdens of defence spending, especially for developing nations such as Eritrea, India and Pakistan (an arms race situation) and for the Middle East (a conflict region). Burundi and Sudan have defence burdens similar to or greater than those of the UK and Germany. Table 1 also shows other measures of the economic burdens of defence for the world's poorer nations (that is, nations which cannot feed, house or educate their populations and which have poor health records). Developing nations accounted for 70% of the world total of 21.3 million military personnel, and such totals further show the importance of military manpower economics. Similarly, developing nations are major importers of arms, while the developed nations are the major arms exporters. Such data provide an introduction to some of the major themes of defence economics, namely, the determinants of military expenditure,

Defence Economics, Table 1 World military spending and armed forces, various years

World military expenditure	*US$ billion, 2004*
NATO	722
USA	467
France	52
Germany	38
UK	54
China	37
Russia	23
World total	**1,035**
Defence share of GDP	%[a]
USA	3.9
France	2.6
Germany	1.4
UK	2.3
Eritrea	19.4
Burundi	5.9
Sudan	2.4
India	2.1
Pakistan	4.4
Israel	9.1
Jordan	8.9
Oman	12.2
Defence research and development[b]	*US$ billion, 2004* (2001 prices and PPP rates)
USA	67.5
Russia	6.1
UK	4.7
USA and EU total	80.9
Estimated world total of defence R&D	**90.0+**
World armed forces	*Number of military personnel, 1999 ('000s)*
Developed nations	6,550
Developing nations	14,700
NATO	4,580
USA	1,490
UK	218
Eritrea	215
China	2,400
World	**21,300**
World arms trade	*US$ million, 2000–2004 (1990 prices)*
Major importers	
China	11,677
India	8,526
Greece	5,263
UK	3,395
Turkey	3,298
World total 84,490	
Major exporters	
Russia	26,925
USA	25,930
France	6,358
Germany	4,878
UK	4,450
World total	**84,490**

Sources: US DoS (2002), NATO (2005), OECD (2004), SIPRI (2005)
PPP purchasing power parity
[a]Defence share data for USA, France, Germany and UK are for 2004; all other data are for 2003
[b]Defence R&D data are for government-funded defence R&D

arms races, alliances, the relationship between defence spending and economic development, the arms trade and the economics of military personnel.

Micro-level data are more limited but there are some useful sources especially on defence contractors and defence industries. Table 2 provides examples of such micro-level data based on the 100 largest arms-producing companies (SIPRI 2005) and employment in national defence industries (BICC 2005). Again, these data are available on a cross-section and time-series basis, and the company data include total sales, total profits and aggregate employment. From Table 2 it can be seen that the USA has six of the world's top ten arms companies and that the American firms have a substantial scale advantage over their European rivals: the average size of a US firm from the top ten is almost twice the corresponding average of the European companies. These data are the basis for research questions about the determinants of firm size, the impact of economies of scale, scope and learning, and the determinants of performance in terms of labour productivity and profitability.

Table 2 also shows data on defence industry employment. The industrialized nations accounted for 63% of total employment in the world's defence industries, with the developing countries accounting for the remaining 37%. The USA, China and Russia have the largest defence industries by employment, accounting for 75% of

Defence Economics, Table 2 Defence companies and industries

Major defence companies	Arms sales, 2003 (US$ million)
Lockheed Martin (USA)	24,910
Boeing (USA)	24,370
Northrop Grumman (USA)	22,720
BAE Systems (UK)	15,760
Raytheon (USA)	15,450
General Dynamics (USA)	13,100
Thales (France)	8,350
EADS (Europe)	8,010
United Technologies (USA)	6,210
Finmeccanica (Italy)	5,290
Major defence industries	**Employment numbers, 2003 ('000 s)**
Industrialized countries	4,710
Developing countries	2,769
NATO	3,452
EU	645
USA	2,700
China	2,100
Russia	780
France	240
UK	200
World total	**7,479**

Source: BICC (2005) and SIPRI (2005)

the world total. Overall, the world military–industrial complex employed almost 29 million personnel in the armed forces and defence industries, reinforcing its role as a major employer of labour, including some highly qualified R&D staff and other highly skilled workers. Such scarce labour has alternative uses in the civilian sector, raising questions as to whether defence spending 'crowds out' valuable civil investment and diverts scientific manpower from civil research projects.

Despite the available data, there remain significant gaps in our knowledge of the world's military sector. Typically, new defence projects are surrounded by secrecy; there are problems in identifying some defence goods (for example, dual use goods, such as civil airliners which can be used as military transport aircraft); there is a lack of good-quality data on defence R&D, including employment in defence R&D; and little is known about China, especially its defence R&D programmes (Hartley 2006a). International comparisons of military expenditure data are also sensitive to the choice of exchange rate adjustments, with country rankings sensitive to the use of market exchange rates or purchasing power parity rates (SIPRI 2005). At the firm and industry levels, analysis of the military business in terms of defence output, employment and profitability is complicated because the typical output comprises a mix of military and civil components, making it difficult to compare the performance of defence contractors and civil firms. Further gaps exist in our knowledge of the world regional distribution of military bases and defence plants, so that it is difficult to assess the economic dependence of various regions on defence spending. Little is known about defence industry supply chains both within countries and within the global economy. Finally, there is a need for more reliable data on the international trade (including illegal transactions) in small arms (these are often the main weapons used in many regional conflicts, such as in Bosnia).

The Defence Economics Problem

This is the standard choice problem of economics, but applied to defence. Typically, following the end of the cold war defence budgets have been either constant or falling in real terms; and these limited budgets are faced with rising input costs of both capital and labour. Equipment costs have been rising by some 10% per annum in real terms, which means a long-run reduction in the numbers of weapons acquired for the armed forces (for example, the US Air Force's original requirement for F-22 combat aircraft for 750 units was later reduced to some 180 aircraft). Similarly, with an all-volunteer force, the costs of military personnel have to rise faster than wage increases in the civil sector. This wage differential is required to attract and retain military personnel by compensating them for the net disadvantages of military life. Here, the military employment contract is unique in that armed forces personnel

are subject to military discipline; they are required to deploy to any part of the world at short notice; they could remain overseas indefinitely; and some might never return (that is, death and injury are a feature of this contract). This combination of constant or falling defence budgets and rising input costs means that governments and defence policymakers cannot avoid the need for difficult choices in a world of uncertainty (that is, where the future is unknown and unknowable, and no one can accurately predict the future).

Faced with this defence choice problem, governments have adopted various solutions. They can adopt a policy of 'equal misery' whereby each of the services is subject to budget cuts (for example, reduced training, cancelling some new equipment projects and delaying others); or they can undertake a major revision of a nation's defence commitments (for example, a defence review such as the UK's 1998 Strategic Defence Review); or they can seek to improve efficiency in the armed forces and defence industries (for example, via a competitive equipment procurement policy and military outsourcing). Other policy options include joining a military alliance (such as NATO; EU) or avoiding the defence choice problem by increasing the defence budget (as in the USA since 11 September 2001); but then choices are needed between defence and social welfare spending.

Economics offers three broad policy principles for formulating an efficient defence policy, namely, final outputs, substitution and competition. Take first the principle of *final outputs*. Measuring defence output is notoriously difficult, but it can be expressed in such general terms as peace, security and threat reduction. The UK has solved the problem by committing (and funding) its armed forces to having the capacity to fight simultaneously three small to medium conflicts (for example, Bosnia, Kosovo, Sierra Leone) or one large-scale conflict as part of an international coalition (for example, the Gulf War, Iraq). This approach is a departure from the traditional focus on measuring inputs in terms of the numbers of infantry regiments, warships, tanks and combat aircraft. Such a focus fails to address the key issue of the contribution of these inputs to final defence output in the form of peace and protection. A focus on inputs also fails to address the marginal contribution of each of the armed forces: what would be the implications for defence output if, say, the air force were expanded by 5–10%, or the navy was reduced by 5–10%?

The second economic principle is that of *substitution.* There are alternative methods of achieving protection, each with different cost implications. Possible examples of partial substitutes include reserves replacing regular personnel, civilians replacing regulars (for example, police in Northern Ireland replacing army personnel), attack helicopters replacing tanks, ballistic and cruise missiles and unmanned combat air vehicles replacing manned strike and bomber aircraft, air power replacing land forces, and imported equipment replacing nationally produced equipment. Some of these substitutions might alter the traditional monopoly property rights of each of the armed forces. For example, surface-to-air missiles operated by the army might replace manned fighter aircraft operated by the air force, and maritime anti-submarine aircraft operated by the air force might replace frigates supplied by the navy.

The third economic principle is that of *competition* as a means of achieving efficiency. Standard economic theory predicts that, compared with monopoly, competition results in lower prices, higher efficiency, and competitively determined profits and innovation in both products and industrial structure. For equipment procurement, competition means allowing foreign firms to bid for national defence contracts and awarding fixed-price contracts rather than cost-plus contracts; it also means ending any 'cosy' relationship between the defence ministry and its national champions and any preferential purchasing and guaranteed home markets.

Competition can be extended to activities undertaken by the armed forces. Here, there is a public sector monopoly problem whereby the armed forces have traditionally undertaken a range of activities 'in house' without being subject to any rivalry. Military outsourcing allows private contractors to bid for and undertake such activities. Examples include accommodation, catering, maintenance, repair, training, transport and management

tasks (for example, managing stores or depots and firing ranges). In some cases, outsourcing involves private finance initiatives whereby the private sector finances the activity (for example, new buildings or an aircrew simulator training facility) and then enters into a long-term contract with the defence ministry to provide services to the armed forces in return for rental payments. Another variant is a public–private partnership whereby the private sector finances an activity or asset in return for rental payments from the defence ministry, but the contractor is allowed to sell any peacetime spare capacity to other users (for example, tanker aircraft capacity which when not needed in peacetime can be rented to other users).

Application of the policy guidelines to an efficient defence policy requires that individuals and groups in the military–industrial–political complex are provided with sufficient incentives to behave efficiently. There are the inevitable principal–agent problems where agents have considerable opportunities to pursue their own interests which may conflict with those of their principals (for example, leading a quiet life rather than bearing the costs of change). Individuals and groups in the armed forces and defence ministries will be reluctant to apply the substitution principle if there are no personal or group incentives and rewards for achieving efficient substitution (that is, interest groups can be barriers to change). Compare the private sector, where there are market and institutional arrangements promoting efficiency in the form of rivalry between suppliers, the profit motive and the capital market as a 'policing and monitoring' mechanism through the threats of takeover and bankruptcy. Such market arrangements are absent in the armed forces (and elsewhere in the public sector).

There is also the challenge of achieving 'top level' efficiency in defence provision. Economic theory solves this challenge as a standard optimization problem involving the maximization of a social welfare function subject to resource or budget constraints (where welfare is dependent on civil goods and security, with security provided by defence). Operationalizing this apparently simple optimization rule is much more difficult. Individual preferences for defence are subject to its public good characteristics and free riding problems and the continued difficulty of defining defence output. In democracies, society's preferences are usually expressed through voting at elections. However, elections are limited as a means of obtaining an accurate indication of society's preferences for defence and its willingness to pay. Elections occur infrequently; they are usually for a range of policies of which defence is only one element in the package (which includes policies on, for example, education, health, transport, the environment, foreign policy and taxation); and the 'voting paradox' shows the difficulty of deriving a society's preferences using the voting system. Nor do voters have reliable information on the output of defence spending.

Defence economics explains military spending using a demand model of the form:

$$\mathrm{ME} = M(P, Y, T, A, Pol, S, Z)$$

where ME = real military spending; P = relative prices of military and civil goods and services; Y = real national income; T = threats in the form of the military expenditure of a rival nation (arms race models); A = membership of a military alliance and the real military expenditure of the allies (such as NATO); Pol = variable for the political composition of the government (for example, left- or right-wing, with the latter favouring 'strong defences'); S = a variable representing the security and strategic environment (such as the end of the cold war; conflicts such as Korea, Vietnam, the Gulf War and Iraq); and Z = other relevant influences (for example, land mass to be protected). Estimation of the demand model usually proceeds without a price variable, mainly because most nations do not provide relative price data. This omission can be justified if the price of military goods and services has inflated at the same rate as civil goods and services; but such an assumption is not always realistic. A survey of empirical results is presented in Sandler and Hartley (1995, 2007).

Conflict and Terrorism

The demand model for military expenditure recognized the relevance of threats such as terrorism

and conflict as determinants of defence spending. Traditionally, conflict and terrorism have been the preserve of disciplines other than economics. For example, debates and decisions about war involve political, military, moral and legal judgements. But conflict has an economic dimension, namely, its costs. Wars are not costless: they can involve massive costs (for example, the Second World War). Economics has also made further contributions in analysing the causes of conflict and in identifying potential targets during conflict (for example, the Second World War selection of aircraft factories, dams, submarine yards and oil fields as targets for Allied bombing raids on Germany).

Economic models start by analysing conflict as the use of military force to achieve a reallocation of resources within and between nations (that is, civil wars and international conflict). Nations invade to capture or steal another nation's property rights over its resources (such as land, minerals, oil, population, water). Conflict has a distinctive feature: it destroys goods and factors of production, and it is easier to destroy than to create. In peacetime, civilian economies aim to create more goods and services through growth and expanding a nation's production possibility frontier. Conflict uses military force and destructive power to enable a nation to acquire resources from another state, so expanding its production boundary through military force (Vahabi 2004).

Conflict and terrorism provide opportunities for applying game theory. They involve strategic behaviour, interactions and interdependence between adversaries ranging from small groups of terrorists, rebels and guerrillas to nation states. Strategic interaction means that conflict can be analysed as games of bluff, chicken and 'tit-for-tat' with first-mover advantage and possibilities of one-shot or repeated games. For example, first-mover advantage might indicate a pre-emptive strike (for example, Pearl Harbour in 1941; Kuwait in 1990). However, there are other, noneconomic explanations of conflict. These include religion, ethnicity and grievance (for example, Germany after the First World War); the desire for a nation state (such as Palestine); the absence of democracy; and mistakes and misjudgement.

The costs of war are a relatively neglected dimension of conflict. War involves both one-off and continuing costs. One-off costs are those of the actual conflict, while continuing costs are any post-conflict costs including those of occupation and peacekeeping. A further distinction is necessary between military and civilian costs. In principle, the military costs of conflict are the marginal resource costs arising from the conflict (that is, those costs which would not otherwise have been incurred). Examples include the costs of preparation and deployment prior to a conflict; the costs of the conflict, including the costs of basing forces overseas and the use of ammunition, missiles and equipment, including human capital and equipment losses in combat; the post-conflict occupation and peacekeeping missions and the costs of returning armed forces to their home nation.

There are further costs of conflict in the form of impacts on the civilian economies of the nations involved in the war. For example, the US and UK involvement in the Iraq war that began in 2003 had possible short- and long-term impacts for both economies. There were possible impacts on oil prices, share prices, the airline business, tourism, defence industries, private contractors, aggregate demand and future public spending plans. Further substantial costs were imposed on the Iraq economy in the form of deaths and injuries of military and civilian personnel, together with the damage and destruction of physical assets. Table 3 shows some examples of the costs of various conflicts for the UK and USA. The general point remains that wars are costly and require scarce resources which have alternative uses (that is, wars involve the sacrifice of hospitals, schools and social welfare programmes). Questions also arise as to whether the benefits of conflict exceed its costs.

Defence economists have also contributed to the analysis of terrorism using both choice-theoretic and game-theoretic models. Terrorism shows that non-conventional conflict is also costly. The attacks of 11 September 2001 on the USA resulted in almost 3,000 deaths and economic losses of $80–90 billion (Barros et al. 2005). Other terrorist-related costs include nations spending on homeland security measures, on terrorist-related intelligence, on security measures in airports, the

Defence Economics, Table 3 Costs of conflict

UK: Conflict	Military costs to UK (US$ billion, 2005 prices)
World War I	357
World War II	1,175
Gulf War	6.0
Bosnia	0.7
Kosovo	1.7
Iraq	6.0 +
USA: Conflict	Military costs to USA (US$ billions, 2005 prices)
World War I	208
World War II	3,148
Korea	365
Vietnam	537
Gulf War	83
Iraq	440
Estimated civilian costs:	Civilian costs (US$ billion, 2005 prices)
Iraq war	
Costs to US economy[a] *from Iraq war*	557
Costs to world economy[b] *from Iraq war*	1,183
Iraq war: costs to Iraq	US$ billion (2005 prices)
Reconstruction costs	20–60

Source: Hartley (2006b)
[a]US civilian costs are of lost GDP for the period 2003–2010
[b]Cost to world economy is lost GDP for the period 2003–2010

lost time waiting at airports to clear security, the losses of liberty and freedoms and the war on terror (for example, in Afghanistan and Iraq).

Choice-theoretic models of terrorism apply standard consumer choice theory with terrorists maximizing a utility function subject to budget constraints. The utility function can be specific, such as a choice between attack modes, say, skyjackings and bombings, or more generally involve a choice between terrorist and peaceful activities. The approach offers some valuable insights into terrorist behaviour and possible policy solutions. The model shows that terrorist behaviour and activities can be influenced by governments acting to reduce terrorist funds (that is, an income effect), by changing relative prices (that is, promoting a substitution effect), and by efforts to change terrorist preferences towards more peaceful activities (for example, Northern Ireland). The substitution effect is an especially powerful insight showing that policies which increase the relative price of one attack mode, such as skyjackings, will encourage terrorists to substitute an alternative and lower-cost method of attack (for example, assassinations, bombings, or kidnappings: Frey and Luechinger 2003; Anderton and Carter 2005).

Conclusion

Defence economics is now established as a reputable sub-discipline of economics. It shows how economic theory and methods can be applied to the defence sector embracing the armed forces, defence industries and the political–institutional arrangements for making defence choices. But this is only the beginning. Massive opportunities remain for further research in the field. Changes in threats, new technology and continued budget constraints will require further adjustments in the armed forces and defence industries, and will generate a new set of research problems. Examples include space warfare, the economics of nuclear weapons policy, assessing the efficiency of armed forces, improving the efficiency of military alliances and developing more efficient approaches to international governance and international collective action.

See Also

- Arms Races
- Arms Trade
- Terrorism, Economics Of
- War and Economics
- World Wars, Economics Of

Bibliography

Anderton, C., and J. Carter. 2005. On rational choice theory and the study of terrorism. *Defence and Peace Economics* 16: 275–282.

Barros, C., C. Kollisa, and T. Sandler. 2005. Security challenges and threats in a post-9/11 world. *Defence and Peace Economics* 16: 327–329.

Benoit, E. 1973. *Defense and economic growth in developing countries*. Boston: DC Heath.
BICC (Bonn International Centre for Conversion). 2005. *Conversion survey 2005*. Baden-Baden: Nomos Verlagsgellschaft.
Brauer, J. 2003. Economics of conflict, war and peace in historical perspective. *Special Issue, Defence and Peace Economics* 14: 151–236.
Frey, B., and S. Luechinger. 2003. How to fight terrorism: Alternatives to deterrence. *Defence and Peace Economics* 14: 237–249.
Goodwin, C., ed. 1991. *The economics of national security*. Durham: Duke University Press.
Hartley, K. 2006a. Defence R&D: Data issues. *Defence and Peace Economics* 17 (3): 1–10.
Hartley, K. 2006b. The economics of conflict. In *The Elgar companion to public economics: Empirical public economics*, ed. A. Ott and R. Cebula. Cheltenham: Edward Elgar.
Hegre, H., and T. Sandler. 2002. Economic analysis of civil wars. *Special Issue of Defence and Peace Economics* 13: 429–496.
Hitch, C., and R. McKean. 1960. *The economics of defense in the nuclear age*. Cambridge, MA: Harvard University Press.
Kennedy, P. 1988. *The rise and fall of the great powers*. London: Fontana Press.
NATO. 2005. *NATO – Russia compendium of financial and economic data relating to defence*. Brussels: NATO.
OECD. 2004. *Main science and technology indicators*. Paris: OECD.
Oi, W. 1967. The economic cost of the draft. *American Economic Review* 57 (2): 39–62.
Olson, M., and R. Zeckhauser. 1966. An economic theory of alliances. *The Review of Economics and Statistics* 48: 266–279.
Peck, M., and F. Scherer. 1962. *The weapons acquisition process*. Boston: Harvard University Press.
Richardson, L. 1960. *Arms and insecurity: A mathematical study of the causes and origins of war*. Pittsburgh: Homewood.
Sandler, T., and K. Hartley. 1995. *The economics of defense*, Cambridge Surveys of Economic Literature. Cambridge: Cambridge University Press.
Sandler, T., and K. Hartley, eds. 2003. *The economics of conflict, 3 vols*. Cheltenham: Edward Elgar.
Sandler, T., and K. Hartley, eds. 2007. *Handbook of defence economics*. Vol. 2. Amsterdam: North-Holland.
Schelling, T. 1966. *Arms and influence*. New Haven: Yale University Press.
SIPRI (Stockholm International Peace Research Institute). 2005. *SIPRI yearbook 2005*. Oxford: Oxford University Press.
US DoS (US Department of State). 2002. *World military expenditures and arms transfers, 1999–2000*. Washington, DC: Bureau of Verification and Compliance, US Department of State.
Vahabi, M. 2004. *The political economy of destructive power*. Cheltenham: Edward Elgar.

Deficit Financing

George L. Perry

Government budget deficits directly affect both the level of aggregate demand and its composition. Less directly, by influencing the amount of national saving and investment, they also influence the growth rate of real income in the longer run. The expected size and predictability of each of these effects is the subject of continuing empirical investigation. Because revenues and some transfer payments automatically rise and fall with cyclical movements in the economy, it is important at the outset to distinguish between actual deficits and structural deficits. The latter are calculated as the deficits that would prevail at some trend level of GNP, while actual deficits grow as the economy falls below this trend and shrink as the economy rises above it. In the rest of this discussion, deficits will mean structural deficits defined in this way, so that changes in the deficit refer to shifts in the deficit that would exist at a given utilization rate of economic resources.

The effects of deficits on the level of aggregate demand, commonly referred to as fiscal policy, became an important focus of governments' budget planning after Keynesian stabilization analysis became absorbed into policy-making. We first consider the basic relationship developed in Keynesian analysis before considering complications that may diffuse it. In the basic case, effects of larger or smaller deficits are symmetric and come about through changes in either government expenditures or tax revenues at given levels of income. Higher levels of government purchases raise demand directly while higher transfers or lower taxes raise incomes, which lead to higher levels of private demand. Whether an expansion of demand results entirely in higher real output or shows up partly in the price level depends on other considerations, such as how much slack exists in the economy and need not concern us at this level

of exposition. For now we assume at least part of any change in GNP is a change in real GNP.

Because higher aggregate demand leads to higher levels of employment and incomes, any initial effects of deficits on demand are amplified through subsequent induced increases in spending out of the induced higher levels of income. So long as these increments to spending are a fraction less than 1.0 of the increments to gross national product, this process converges to a higher equilibrium level of aggregate income and output. The ratio of the eventual higher level of GNP to the initial fiscal stimulus is known as the multiplier. Thus if the multiplier on government purchases is 2.0, an initial $1 billion increase in the deficit resulting from $1 billion more in government purchases leads to a level of GNP $2 billion higher than the initial level. An equivalent way of expressing these effects is to note that an initial expansion of the deficit is a reduction in government, and therefore national, saving. In response, GNP expands to the point where national saving again equals investment.

We may now consider the main qualifications to this basic fiscal policy model. They are all possible reasons why offsets may occur to the apparent increments to demand coming from a fiscal action.

The first issue has to do with monetary policy and is partly definitional. Pure fiscal policy effects, which we are discussing here, should mean the effects that occur when the budget deficit shifts but monetary policy is unchanged. Depending on the definition of unchanged monetary policy that is used, a portion of the fiscal effects on demand may be offset by higher interest rates. The most common notion of unchanged monetary policy is an unchanged money supply. If the GNP were determined simply as a proportion of the money supply, then on this definition of unchanged monetary policy, whatever added demands came from the budget deficit would necessarily be offset by reduced demands elsewhere. This 'crowding out' would occur as a result of a rise in interest rates that directly reduced domestic interest-sensitive demands such as housing or business investment or that reduced the foreign trade balance by appreciating the exchange rate.

However, both theory and empirical evidence reject this model of a fixed relation between money and GNP. The interest rate increase that would reduce some private demands will also lead to economizing on money balances, thus breaking the fixed link between GNP and money demand. Nonetheless, to the extent that a fixed money supply forces interest rates to change in response to a fiscal change, a fixed money supply will reduce the effect on GNP that we attribute to a pure fiscal impact. Some private demands will change in response to the change in interest rates, offsetting part of effect on total GNP of the fiscal change.

Under alternative definitions of an unchanged monetary policy, we arrive at different assessments of what is here called the pure fiscal impact. Other candidates for defining an unchanged monetary policy include unchanged levels of bank reserves or borrowed reserves. Because the supply of money is itself elastic with respect to interest rate changes, the rise in interest rates that accompanies a change in fiscal policy is somewhat smaller under this definition than if the money supply is assumed fixed. As a consequence, a greater impact on demand is attributed to pure fiscal policy. Finally, if we define unchanged monetary policy as an unchanged real interest rate, fiscal policy would have the full impact on GNP described in the basic model above. Although such a policy would be unsustainable with overfull employment, it is not logically inferior to a constant money supply definition. Furthermore, targeting interest rates corresponds to the way monetary policy has often been conducted.

The next set of qualifications to the basic fiscal policy model concerns the behavioural response of private sector agents. One issue concerns the possible difference in consumers' response to temporary and permanent fiscal changes. The permanent income hypothesis relates current consumption to consumers' expected permanent income. A fiscal change that is known to be temporary will therefore have a much smaller effect on current consumption than would the same size fiscal change if it were taken to be permanent. However, if many consumers are constrained in their spending by a lack of liquidity, because they

cannot freely borrow at near market interest rates against their future incomes, consumption will not be governed by permanent income. In this case, by relieving the binding liquidity constraint, temporary fiscal changes could have nearly the same effect on current spending as permanent ones.

Although the issue is unsettled because if is difficult to model consumers' expectations of future income, the balance of the evidence suggests that personal tax reductions that are known to be temporary, such as one-time tax rebates, have a smaller effect on spending than do other types of fiscal changes. But this is not a general result for all types of temporary tax change, some of which may have exceptionally large effects. An enlargement of investment tax credits for a limited period of time may have such an exceptionally large effect by shifting investment projects forward in time to take advantage of the temporary tax incentive. The reverse effect could occur from a temporary suspension of a tax credit. Such temporary changes have been used for stabilization by various governments in the past. However such special inducements that shift demand through time only alter demand now at the expense of demand later.

A more extreme argument against fiscal changes affecting GNP is the so-called Ricardian equivalence hypothesis, which asserts that deficits directly bring forth an offsetting change in private saving. This idea, which has been associated in modern times with Robert Barro, presumes that consumption decisions are based on an optimizing strategy over an infinitely long time horizon so that people today adjust their own consumption and saving in response to the after-tax incomes they expect to flow to themselves and their heirs over the indefinite future. Because in this model added deficits today will require higher taxes in the future, consumers fully offset increased government deficits with increased personal saving, thereby eliminating any effect of deficits on GNP.

Although it has renewed interest in modelling fiscal effects more carefully, there is little empirical support for this extreme proposition. However whether deficits, which directly change total national saving, induce some partial offest in saving in other sectors remains an unsettled empirical question. Such direct offsets appear to be more likely in response to some sources of change in deficits than others. Quite apart from the Barro-like effects just discussed, conventional consumption functions predict that a minor fraction of changes in disposable income will be saved, so that a shift in the deficit coming from personal tax reductions would induce a small rise in personal saving. A shift in the deficit coming from reduced business taxation could produce a shift in net business saving depending on how much the tax change affects business investment. In part, how important such effects are will depend on the time horizon in question. For example, some tax changes may have significant effects on investment in the first instance as business adjusts to a different desired capital stock. But once the new desired stock is achieved, investment demand will be changed much less. Further time lags may be involved as firms adjusted dividend payouts and individuals adjust their consumption. But leaving aside such transitory complications, in a steady state the directly induced effects of deficits on private saving appear to be small. Therefore shifts in deficits do shift total saving, total demand and GNP.

We turn next to the effects of a shift in the deficit with the level of GNP held constant. This case is relevant for analysis of the medium run, when departures of real GNP from its trend are averaged out. It is also relevant whenever monetary policy or real limits on expansion are assumed to constrain total real GNP. As before, the shift in the deficit represents a shift in government saving; but since GNP cannot change, this shift must be offset by a corresponding shift in saving net of investment of one or more other sectors.

Much of the adjustment to deficits appears to take place through induced changes in interest rates rather than directly. Higher interest rates reduce business investment, residential construction and spending on consumer durables. They may also affect personal saving, and therefore consumption more generally, although there is little evidence that such effects are large enough to be important. In some circumstances, a major part of a shift in the deficit may be offset by a

decline in net foreign investment or, equivalently, a decline in the current account balance. Such an effect was an important part of the adjustment to the historically large US budget deficits of the 1980s.

The offsets to a deficit will not generally remain unchanged through time. At first, a modest rise in interest rates may induce an appreciation of the currency and a decline in the current account balance. This minimizes the effect of the deficit on domestic investment. But as foreigners' holdings of the deficit country's assets continue to increase, it may take ever-higher interest rates to maintain the currency at its appreciated level; and this, in turn, will reduce domestic investment further, thus shifting more of the adjustment to the deficit onto domestic sectors.

Because, in general, larger deficits lead to higher interest rates at any level of GNP, they reduce the share of GNP devoted to investment and increase the share devoted to consumption and government spending. To the extent that investment is crowded out by higher interest rates, the future capital stock will be smaller, thus reducing future real incomes and consumption. Even if domestic investment is sustained by increased net investment by foreigners, the earnings on this investment will accrue to foreigners, so again real domestic incomes will be reduced. If the deficit is increased as a consequence of higher government investment or other growth-inducing expenditure such as research and education, growth might not suffer absolutely. It would still be reduced relative to a budget that financed such outlays with higher taxes that suppressed present consumption.

These outcomes do not imply that a zero deficit, or any particular level of surplus or deficit, is optimal at all times or even, on average, in the long run. Sustained deficits can be too large in the sense that they lead to an explosive growth in the ratio of debt to GNP. But apart from such a limiting case, the appropriate deficit to GNP ratio will depend on the prevailing ratio of private saving to GNP and on the desired ratio of foreign investment or disinvestment to GNP. These ratios have varied substantially across countries for reasons that have to do with the generosity of public retirement programmes, established lending practices and policies for homeownership and other factors that determine private saving propensities and foreign investment schedules. In part they reflect different states of maturity in economies that make the return to saving and investment higher in some than in others. But whatever these fundamental characteristics of economies may be, within a range, varying deficits can be used to alter the ratio of national saving and investment to GNP.

See Also

- ▶ Burden of the Debt
- ▶ Crowding Out
- ▶ Demand Management
- ▶ Fine Tuning
- ▶ Public Debt
- ▶ Ricardian Equivalence Theorem

Deficit Spending

M. J. Artis

Interest in the economics of deficit finance began to all intents and purposes with the absorption of the economics of the *General Theory.* Before that, though with a few exceptions, the economic discussion of the public finances was based on the assumption of a fully employed economy and the notion of using deficit finance to stimulate output was accordingly not at issue.

Despite the fact that the economics of deficit finance begin with the Keynesian Revolution, it has been conclusively established by Kregel (1985) that Keynes himself 'did *not* ever directly recommend government deficits as a tool of stabilization policy' (Kregel, p. 32). Keynes played a conservative political hand and viewed budget deficits with a 'clearly enunciated lack of enthusiasm'. Although Kregel's discovery is both true and startling, the founder of the Keynesian theory

of public finance, Abba Lerner, described what he termed the concept of functional finance as 'first put forward in complete form by J.M. Keynes in England' (Lerner 1943). This seems to be, therefore, another example of Keynes himself being unaware of the full implications of his own theory or, alternatively, of Keynes himself being aware of political reasons why it would be inappropriate to declare publicly the full implications of his theory for the public finances. (It remains unclear which of these propositions has the greater part of the truth.)

The doctrine of functional finance says that the balance of spending and taxation in the budget should be manipulated so as to produce the desired result for output and employment and not in the interests of realizing a balance or surplus (or deficit) *per se.* This is entirely in tune with the income–expenditure analysis of the determination of income which became the central interpretation of the *General Theory;* since output is driven by demand, output can be altered by government action to raise or lower its own expenditures and to raise or lower, via taxation, the spending of the private sector. It is in this (simple and straightforward) sense that deficit finance and Keynesian economics are closely and correctly linked together. Strictly speaking (and this was fully recognized by Lerner), the objective is not output *per se* but 'internal balance'; this is important, because in conditions of full employment, potential excess demand and inflation the doctrine may indicate that a budget surplus is more appropriate than a deficit. The suggestion that Keynesian economics leads to excessive budget deficits does not therefore seem at all correct, although it is one to be encountered in the writings of some critics.

The deficit in the budget *per se* is of course an endogenous item, in the sense that tax revenues and some components of expenditures depend directly upon the level of output and economic activity. In order to obtain measures of deficit finance which are free of this endogeneity, it has become customary to estimate the 'structural' deficit, or the deficit at a normalized level of activity. Measures of the structural budget deficit are standard fare as summary measures of the stance of fiscal policy.

In recent years, the dominance of the principles of functional finance has declined and arguments have been erected (or resurrected) to show that deficit finance may not have the properties ascribed to it in the principles of functional finance; in particular, it has been argued that deficit finance is no different from deferred taxation and deferred taxation no different from current taxation. Hence the case for deficit finance has to be made on some different ground. This argument, perversely referred to as the 'Ricardian equivalence' doctrine (perverse because Ricardo, having entertained it, rejected it) takes its point of departure in a perfect foresight, full information (and fully employed) economy. In such an economy, if individuals are infinitely lived (or care about the welfare of their heirs), a current deficit financed by the issue of bonds creates the expectation of corresponding tax liabilities in the future. The wealth embodied in the bonds (equal to the present discounted value of the flow of coupons and repayment of principal) is precisely offset by the present discounted value of the stream of extra taxes required to service the coupons and repayment of principal. The two are equivalent and cancel out. The bond issue might as well be cancelled in favour of an increase in taxes since private sector savings must rise to meet the obligation to pay future taxes in any case. This argument against deficit finance, put forward most forcefully by Barro (1974) must be regarded as unacceptably extreme. A number of objections may be made to it, as a doctrine of real world relevance; a break in the chain of inheritance, lack of information, imperfect capital markets, less than full employment states are all objections. It is right to qualify these objections by pointing out that it is not in every particular case that the validity of the objection restores the assumptions of the functional finance income–expenditure model as the alternative correct model. Although the Ricardian equivalence theorem appears to be unacceptably extreme it is of interest to note that if it is accepted, a case for loan finance may still exist if only taxes are not lump sum. For if they are not and as usually assumed, welfare losses rise proportionately with the tax rate, then there is a presumption in favour of smoothing tax rates;

hence if expenditures or the tax base are erratic, a case for deficit finance reappears.

In practice, however, it is other considerations which have reduced the dominance of functional finance priciples and resurrected arguments for being concerned about deficit finance. Two, in particular, may be mentioned: first, the connection, real or presumed, between fiscal deficits and monetary growth in periods when monetary targeting has become a central policy objective; second, the structural problem raised by the deceleration of economic growth. Overlying these considerations is the point that with better information flows and increased financial integration, asset markets dominated by forward looking expectations have considerable power to check a fiscal policy that seems adventurous. In particular, if deficit finance is conducted on so large a scale as to raise doubts about its substainability, the market may conclude that rather than change the policy, the result will be explosive growth in the money supply. As a result, bond prices fall currently, and the exchange rate plummets. Scenarios of this type are responsible for an increasing emphasis being placed on targets for the ratio of public sector debit to GNP, an integral control version of deficit/GNP ratios. In contrast to the destabilizing character of the latter, however, targets for the ratio of debt to potential GDP allow the stabilizers to be 'turned on' as output deviated from potential and provide a compromise between the flexibility and complete discretion and the potentially destabilizing rigidity of deficit targeting. Whether the compromise is the best that can be achieved remains to be discovered. All that is really required is that the market should trust the government, in following the principles of functional finance, not abuse them. To suppose that this trust can be inspired by adopting a target which implies a large degree of sacrifice of these principles may be wrong.

See Also

- ▶ Deficit Financing
- ▶ Demand Management
- ▶ Finance

Bibliography

Barro, R. 1974. Are government bonds net wealth? *Journal of Political Economy* 82(6): 1095–1117.

Kregel, J. 1985. Budget deficits, stabilization policy and liquidity preference: Keynes' post-war proposals. In *Keynes's relevance today*, ed. F. Vicarelli. London: Macmillan.

Lerner, A.P. 1943. Functional finance and the federal debt. *Social Research* 10: 38–51.

Defoe, Daniel (1660–1731)

K. Tribe

Born Daniel Foe in 1660, son of a London tradesman and Nonconformist, Defoe's early life was that of a merchant with a diversity of interests and ambitions. After his support for the Monmouth rebellion of 1685 he welcomed the accession of the Prince of Orange in 1688, later being given employment by the government. A financially advantageous marriage followed, and then his fortunes reversed with the collapse of his ventures in 1692 owing £17,000. His efforts at clearing his debts first turned him towards journalism, and this was to be his major occupation for the remainder of his life. In the early years of the 17th century he met with some literary success, but in 1702 he was imprisoned for libel. His release in 1704 was conditional on his undertaking to establish an intelligence network for the Government, and in the succeeding years he travelled widely, gathering information and assessing popular opinion. In 1713 he was imprisoned once more, this time for anti-Jacobite writings; pardoned in 1715, he returned to literary work and in the period until his death in 1731 produced the majority of the works for which he is known today.

Defoe published a number of tracts upon directly economic issues, chief among them his *Plan of the English Commerce* which argued that the employment of labour on the working-up of domestic produce (in particular, wool) was the true path to prosperity. He is perhaps best known

for his novel *Robinson Crusoe* and his two accounts of English society, *Journal of the Plague Year* and *Tour through the whole Island of Great Britain.* The first of these was published anonymously in 1719 and was, until Defoe admitted authorship, thought to be a true account of the life of a castaway. In his defence, Defoe suggested that he had included much of his own experience, and it is shown today that the work is based on the experience of Alexander Selkirk. This blending of 'fiction' and 'fact' is typical of the other two works; for while they are based upon Defoe's observations, the form in which they are cast is fictional. The *Journal* records events that occurred when Defoe was five and the *Tour*, published in 1724–6, is in fact a compilation drawing in part on the travels of Defoe some 20 years previously. For all this, they are no less valuable as accounts of contemporary society, and were regarded as models by later observers.

Bibliography

Moore, J.R. 1960. *A checklist of the writings of Daniel Defoe*. Bloomington: Indiana University Press.

Degree of Monopoly

Kurt W. Rothschild

If the term 'monopoly' is taken in its literal meaning, then there is no room for such a thing as a 'degree of monopoly'. For 'monopoly' means – taking into account the Greek origins of the term – a single seller; and there cannot be any 'degrees' of singleness. In fact, all through the 19th and well into the 20th century, economic thinking tended to look at monopoly in this way. Monopoly referred to the market form with a single seller as opposed to Competition, where several firms appear on the market. When the two market forms and their consequences were analysed it was soon realized that the two types were not quite sufficient to cover all decisive elements, and some in-between forms were taken into account as, for instance, in Cournot's duopoly analysis or in Marshall's insights into imperfect competition. But all the time monopoly remained more or less unscathed as a clearly defined juxtaposition to competitive market forms.

This situation became undermined from two different sides: rather gradually from a practical–political angle, when at the turn of the century a growing concern with big-business practice led to demands for anti-monopolistic legislation, and – later on – more dramatically in the theoretical sphere when the almost simultaneous appearance of Joan Robinson's (1933) and Edward H. Chamberlin's (1933) treatises on monopolistic competition provided a new perspective for market form analysis. In practical affairs it had soon become obvious that *exclusive* control of supplies of a certain commodity by a single firm was rather an exception, but that all the suspected evils of monopoly – high prices, displacement of actual or potential competitors, curtailment of production etc. – could also be detected when big firms or cartels dominate a market, even if there are numerous smaller competitors. Monopolistic power thus became connected with the question of concentration, and varying degrees of concentration could be seen as expressions of varying degrees of monopoly. This led to various proposals of a *descriptive–statistical nature* to measure degrees of monopolistic domination.

On the theoretical plane the development originated from a growing sophistication in the analysis of 'pure' market forms. On the one hand it became clear that there cannot be such a thing as a completely isolated monopolist free from competitive pressures, because there always exist substitutes which limit his room for manoeuvring; and on the other hand the heterogeneity of goods, location, and availabilities so departmentalizes competitive markets that the individual firm can have a certain amount of monopoly-like freedom for price-setting which could not exist on perfectly competitive markets. The classical juxtaposition of monopoly and competition had lost its simplicity; the 'pure' cases turned out to be

limiting concepts in a world characterized by an intermixture of monopolistic and competitive elements. Thus the 'monopolistic competition revolution' gave rise to a series of attempts to find a suitable *theoretical* tool for measuring the 'degree of monopoly' with the main stress being put on the conceptual and analytical basis while the question of quantitative expression was largely neglected or remained unsolved.

Before giving a short presentation of the more important indices used for measuring the 'degree of monopoly' it is necessary to enumerate some of the formidable difficulties that beset *any* attempt to find a suitable (single) expression which could provide a unique and comprehensive index. First of all there is the firm–industry problem. As long as a monopoly is conceived in its narrowest sense – a single supplier of a certain commodity – the problem does not arise: the monopolistic firm coincides with the entire industrial branch. But once we allow for *degrees* of monopoly in multi-firm industries and for heterogeneous goods, all indices which try to measure monopolistic power within an industry come up against the problem of where to draw the lines for a meaningful group. If we want to estimate the monopolistic position of firms in the motor-car industry, are we to take as the decisive industrial group all motor-cars, or motor-cars of a certain size, motor-cars of a certain type, or what? Obviously one wants to draw the line where products cease to be serious substitutes for the commodity in question. But this involves necessarily a somewhat arbitrary decision and the results will be affected by it. Some writers, following Triffin (1940), have argued that in a world of heterogeneous goods and inter-industrial competition the concept of industry should be dropped altogether and the degree of monopoly of a firm should be measured exclusively vis-à-vis *single* other firms with the aid of cross-elasticities of demand. These would be zero in the case of complete monopolistic independence. But this approach would lead to an enormous number of cross-elasticities for every firm, and it neglects the fact that we do deal with industrial groups and problems in practice.

A further problem arises from the fact that indices of relative size within an industrial group do not tell us sufficiently how far other factors – like regional dispersion, marketing activities etc. – influence the monopolistic status of big and small firms. Measures which rely on realized prices and profits can only tell us something about *actual*, but not about *potential* monopoly power. Finally, there is always – in view of business secrecy and incomplete statistics – a serious data and estimation problem when it comes to quantitative judgements.

But the most important reason for coexistence of various degree of monopoly indices is that the monopoly problem has different aspects which require different measuring rods. Thus, the problem of monopoly power may be seen as a problem of relative market power within an industry, that is the problem of big firms versus small firms. This aspect plays an important role in anti-trust and fair-competition legislation. From the point of view of traditional economic theory, the question of monopolistic price formation with its effects on optimal allocation and economic welfare is the dominant one. Others – as for instance Marx or Kalecki – have stressed the distributional aspects of monopoly power, particularly with regard to the wage–profit relation. Finally, a political–economic viewpoint looks at the problem of the influence of monopolies on the state in the age of 'Monopoly Capitalism'.

The search for suitable indices originated in connection with political concerns over the growing concentration in certain industries. This gave rise to a demand for descriptive–statistical measures to be used as diagnostic instruments. A widely used index of long standing is the so-called 'concentration ratio' which measures the weight of the biggest enterprises in an industry on the basis of the percentage share of the biggest firms in total output, or sales, or employment, or capital assets.

The advantages of this index are obvious: the required data are usually available and its meaning is easily appreciated. It can certainly act as a rough indicator of levels and changes in monopolistic positions in individual industries. If shares of the biggest firms in *total* manufacturing output (employment etc.) are measured, we obtain hints with regard to the Monopoly–State problem. The

main disadvantage of the concentration ratio is that it completely disregards information about the number and size of firms beyond the few leading firms. But their structure can influence the monopolistic context. As an alternative (or complement) to the concentration ratio we, therefore, find suggestions to measure the degree of monopoly with the aid of distributional indices like the Lorenz curve or the Herfindahl index which measure overall inequalities of distribution – greater inequalities (in output etc.) to be taken as higher degrees of monopoly. This is, however, hardly satisfactory because it does not give sufficient weight to concentration at the upper end (biggest firms) which is decisive for the monopoly problem.

When we turn to the theoretically oriented indices of degree of monopoly we meet a greater variety. Most of the proposals were born in the two decades after the publication of Chamberlin's and Robinson's books; since then the interest in further developments has died down. The pioneering study appeared in 1934 when Lerner (1934) proposed an index based on the theoretical idea that pure competition should be taken as the benchmark. Taking into account that in pure competition equilibrium prices will equal marginal costs he took as his indicator of degree of monopoly the excess of price over marginal cost relative to price (i.e. (p-c)/p, with p = price and c = marginal cost). This index equals zero in case of pure competition and can rise to a maximum of one when marginal costs are zero. What it measures is the deviation of a firm's price from the competitive ideal, with consequences for allocation and welfare. But since the index is (in equilibrium, with marginal revenue equal to marginal costs) equivalent to the reciprocal value of the price elasticity of demand, it does not take into account cost and supply considerations. The main advantage of the Lerner index is that it does not require the definition of an industry group. The same is true for Weintraub's (1949) suggestion of an index which measures the ratio between the actual output of a firm and the output it would produce under pure competition. This index equals one in the fully competitive case and becomes smaller with growing monopolistic deviations. Difficulties arise with regard to suitable data.

Also based on the individual firm and avoiding the ambiguous industry concept is an index by Bain (1941) which starts from the commonly acknowledged idea that monopoly power is acquired in order to obtain higher profits. Bain, therefore, proposes to use the ratio between a firm's profit rate and a 'normal' competitive profit rate as a degree of monopoly indicator. Since actuarial profit data do not meet theoretical requirements this approach also runs into estimation problems. To lay stress on profits is an advantage, but the index cannot distinguish monopoly-caused profits from other types (demand shifts, windfalls, etc.).

Some other indices take into account the (intra-industrial) interdependence of firms. Rothschild (1942), referring back to Chamberlin's 'two demand curves' facing a firm – a special one, when it alone varies the price, and a general one, when all firms together change their price – takes as his index the ratio of the slopes of the special and the general demand curve. In full competition the slope of the special curve and the index are equal to zero, and the index rises to one in pure monopoly when both curves coincide. Estimation problems are as formidable as before. Morgan (1946) and Papandreou (1949) both build on Triffin's idea of making cross-elasticities of demand a decisive criterion, but expand his ideas. Morgan makes monopoly power vis-à-vis other firms a function of the firm's share in their combined output and of the heterogeneity of goods (measured by cross-elasticities). Both enter positively into the index. Papandreou, in a combination of two complex indices, takes into account not only the crosselasticities of demand (which determine the degree of 'insulation' when other prices are changed) but also output capacities (which determine the power of 'penetration' into other markets when demand is increased).

A special word must be said about Kalecki's (1938) 'degree of monopoly' which has become the most important specimen in theoretical literature. *Formally* it is similar to Lerner's but the theoretical underpinnings and uses are quite different. In contrast to Lerner, Kalecki starts from a

(manufacturing) world in which oligopoly and imperfect competition are the rule. Pure competition cannot be the standard. All firms work below capacity; their marginal production costs tend to be constant and equal to average production costs. Prices are 'administered' by a mark-up on average production costs. The higher the mark-up the greater the difference between price and average cost (=marginal cost) and the higher, therefore, the 'degree of monopoly' in the definition of Lerner. Since the mark-up determines gross profits one obtains a theoretical framework where the 'degree of monopoly' is a decisive factor in determining the income distribution between production workers' wages and gross profits. But it is important to realize that in this setting (general under-utilization of capacity) the 'degree of monopoly' has a very wide meaning: in addition to monopoly power in the narrower sense and the higher profits that go with it, it also covers other non-wage items such as salaries and depreciation.

In more recent years Cowling (1978) and others have taken up Lerner's and Kalecki's indices in modified form to study postwar developments in welfare losses and distributional effects of growing monopolization and oligopolization. The fact of growing management influence which can transform monopoly profits into 'costs' and managerial advantages is taken into account and the index of the degree of monopoly is supplemented by an index of the degree of managerial discretion. The greater the latter, the smaller will be the apparent degree of monopoly as measured by reported profits.

The various conceptual and statistical attempts to find a suitable index for the 'degree of monopoly' have contributed to a better understanding of the issues involved. While most of them can shed *some* light on the monopoly problem, this is far too complex and many-sided to be compressed into one single index or to be fully describable in purely quantitative terms.

See Also

- ▶ Elasticity
- ▶ Kalecki, Michal (1899–1970)
- ▶ Lerner, Abba Ptachya (1905–1982)
- ▶ Monopolistic Competition
- ▶ Monopoly

Bibliography

Bain, J.S. 1941. The profit rate as a measure of monopoly power. *Quarterly Journal of Economics* 55: 271–293.

Chamberlin, E.H. 1933. *The theory of monopolistic competition*. Cambridge, MA: Harvard University Press.

Cowling, K. 1978. Monopoly, welfare and distribution. In *Contemporary economic analysis*, ed. M.J. Artis and A.R. Nobay. London: Croom Helm.

Kalecki, M. 1938. The determinants of the distribution of the national income. *Econometrica* 6: 97–112.

Lerner, A.P. 1934. The concept of monopoly and the measurement of monopoly power. *Review of Economic Studies* 1: 157–175.

Morgan, T. 1946. A measure of monopoly in selling. *Quarterly Journal of Economics* 60(3): 461–463.

Papandreou, A.G. 1949. Market structure and monopoly power. *American Economic Review* 39: 883–897.

Robinson, J. 1933. *The economics of imperfect competition*. London: Macmillan.

Rothschild, K.W. 1942. The degree of monopoly. *Economica* 9: 24–39.

Triffin, R. 1940. *Monopolistic competition and general equilibrium theory*. Cambridge, MA: Harvard University Press.

Weintraub, S. 1949. *Price theory*. New York: Pitman Publishing Corporation.

Degree of Utility

P. H. Wicksteed

This phrase was first made current by Jevons in his *Theory of Political Economy,* 1871. Its precise significance will be best elucidated by an analogy. 'Degree of utility' stands in the same relation to 'total utility' as 'velocity' to 'space traversed'. Suppose we have a body projected vertically upwards from rest, at a given speed. We may inquire *first* at what height the body will be found at any moment after its projection, and *second* at what rate it will be moving at any point of its course, and clearly the rate of its movement is the rate at which its height is increasing (whether

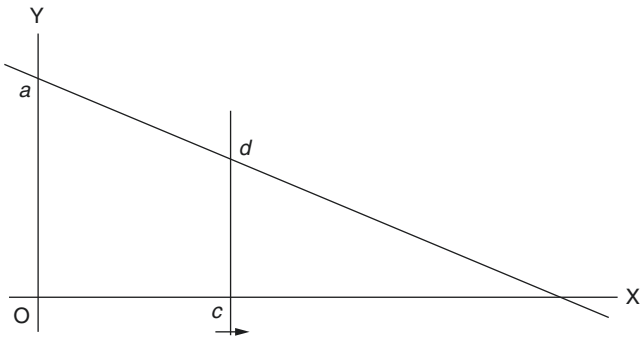
positively as it rises, or negatively as it falls). This rate may be measured in feet per second, or in miles per hour, or in any other suitable unit, but in any case it varies from point to point and does not continue the same during any period, however short.

We must now extend the idea of measurement to such economic conceptions as 'satisfaction' and 'utility'. Measurement consists essentially in determining the ratio of the magnitude investigated to some other magnitude adopted as a standard; and a 'satisfaction' would accordingly be measured if we could determine its ratio to some standard satisfaction, or, which amounts to the same thing, some standard dissatisfaction. Thus if I wish to measure the satisfaction derived by a hungry man from the consumption of a certain quantity of bread, I may inquire how much labour he would perform, under stated conditions, rather than go without it; or what he would pay for it sooner than go without if an unscrupulous monopolist exacted from him the extreme famine price. Thus if we take any standard we choose we can, ideally at least, conceive of any concrete 'utility' or 'satisfaction' being measured in it. But we must remember that such measurements are based on the relative magnitudes of different satisfactions, etc., to one and the same person, and do not profess to give us means of comparing a satisfaction experienced by one mind with a satisfaction experienced by another; for no one can say that the standard unit of satisfaction selected means the same thing to two different men. Nor shall we find that any such absolute measurement is needed for the purpose in hand.

Having premised so much, we may now work out the economic analogue of the projected body. Suppose we take such a commodity as bread supplied to a hungry man. *Firstly*, we may inquire what amount of satisfaction the man has derived from the consumption of any given quantity of bread; in which case we shall be investigating the 'total utility' or 'value in use' of that quantity of bread, to that man, under those conditions. *Secondly*, we may inquire at what *rate* (per ounce, per pound, etc.) the consumption of the bread is conferring satisfaction upon the man at any point in the course of his meal; and in that case we shall be investigating the 'degree of utility' of the bread. This 'degree of utility' will of course vary from point to point. When the man was at his hungriest he would be deriving relatively great satisfaction per ounce of bread consumed, and towards the end of his meal, when nearly satisfied, his satisfaction per ounce would be relatively small; and, theoretically, it will not remain constant during any period, however short. Now this 'degree of utility' is obviously the rate at which the 'total utility' is increasing; just as the velocity of a rising or falling body is the rate at which 'space traversed' or 'height' is increasing (Fig. 1).

The precise relation of velocity to space traversed, and of degree of utility to total utility, is expressed mathematically by saying that the former are the 'differential coefficients', 'first-derived functions', or 'fluxions' of the latter; and, graphically, if the latter are expressed by areas the former will be expressed by lines. In the figure, if we imagine the line *cd* moving from O in the direction of the arrow-

Degree of Utility, Fig. 1



head, at a uniform rate, to represent the lapse of time, and if we imagine the area *a*O*cd*, to represent the space traversed by the projected body in the O*c*, then the intercept *cd* will be the differential coefficient of *a*O*cd* and will represent the velocity of the body, or the rate at which it is rising, at the point of time represented by *c.* Perhaps this will be sufficiently obvious to the non-mathematical reader if he reflects that velocity represents the rate at which height is increasing, as time lapses, and observes that the length of the intercept *cd* likewise determines the rate at which the area *a*O*cd* increases as the vertical line moves in the direction of the arrow-head.

Now let the movement of the vertical from O represent the consumption of the bread, so that O*c* represents the amount consumed up to any given point of the meal; and let *a*O*cd* represent the total satisfaction derived from the consumption up to the point reached, then *cd* will still be the differential coefficient of *a*O*cd*, and will represent the rate per unit (ounce, etc.) at which the consumption of the bread is now increasing the total satisfaction reaped by the consumer. That is to say *cd* represents the degree of utility of bread at the point *c,* the amount represented by O*c* having already been consumed.

It should be observed, however, that when we are dealing with economic quantities, the line *ad* will probably never be a straight line, but always a curve of more or less complexity; and it will seldom or never be possible to determine its actual form with any precision.

The main interest naturally attaches to the degree of utility of that increment of a commodity which the consumer expects to obtain next, or which he may have to relinquish, that is to say the last increment he has secured or the next he hopes to secure. This is called by Jevons the 'final degree of utility'. The analogy of the moving body insisted on above was developed by Professor Léon Walras of Lausanne, and was first suggested by his father, A.A. Walras.

See Also

► Final Degree of Utility

Del Mar, Alexander (1836–1926)

Joseph Aschheim and George S. Tavlas

Born in New York City and educated as a mining engineer at New York University, Del Mar formulated views on monetary economics on the basis of numerous empirical investigations which he undertook both on his own during the Civil War, and while serving as the first director of the US Bureau of Statistics.

Del Mar anticipated modern monetary analysis along a broad front. While an exponent of the long-run neutrality of money, he argued that in the short-run money is non-neutral. Specifically, he placed emphasis on evaluating the impact of monetary changes in the context of dynamic analysis and developed a broad monetary transmission mechanism (termed the 'precession of prices') which depended upon the marketability of assets (1864). Since he perceived that labour was the least marketable of assets, its price was the last to respond to a monetary change. The perceived tendency of wages to lag behind prices, the observed procyclical nature of velocity, due to the effect of price expectations, and Del Mar's link-up between anticipations of price changes and variations in nominal interest rates, allowed him to formulate such concepts as self-generating expectations and money illusion (Tavlas and Aschheim 1985).

From the 1860s to the 1880s, Del Mar undertook perhaps the first attempts in the US literature to estimate both the value of velocity and the annual rate of increase of national wealth. Based on his estimates, in *The Science of Money* (1885) he advocated as a policy-guide the first numerical monetary growth-rate rule in the professional literature. His work, however, was largely overlooked by the profession, except for Irving Fisher. Fisher cited Del Mar over the course of nearly 40 years.

Selected Works

1862. *Gold money and paper money.* New York: A.D.F. Randolf.
1864. The dynamics of finance. *The New Nation* 1: 1–4.
1865. The growth of national wealth. *New York Social Science Review: Devoted to Political Economy and Statistics* 1: 193–218.
1880. *A history of the precious metals.* London: George Bell.
1885. *The science of money.* London: George Bell.
1886. *A history of money in ancient countries.* London: George Bell.

References

Tavlas, G.S., and J. Aschheim. 1985. Alexander Del Mar, Irving Fisher, and monetary economics. *The Canadian Journal of Economics* 18(2): 294–313.

Demand for Money: Empirical Studies

Stephen M. Goldfeld

The relation between the demand for money balances and its determinants is a fundamental building block in most theories of macroeconomic behaviour and is a critical component in the formulation of monetary policy. Indeed, a stable demand function for money has long been perceived as a prerequisite for the use of monetary aggregates in the conduct of policy. Not surprisingly, then, the demand for money has been subjected to extensive empirical scrutiny.

Several broad factors have shaped the evolution of this research. First, there is the evolving nature of theories of the demand for money. The simple versions of the so-called quantity theory were followed by the Keynesian theory of liquidity preference and then by more modern variants. As theory evolved, so did empirical research. A second factor is the growing arsenal of econometric techniques that has permitted more sophisticated examinations of dynamics, functional forms, and expectations. These techniques have also provided researchers with a wide variety of diagnostic tests to evaluate the adequacy of particular specifications.

Finally, and perhaps most importantly, research has been spurred by the apparent breakdown of existing empirical models in the face of newly emerging data. These difficulties have been particularly evident since the mid-1970s. In many countries this period has been marked by unusual economic conditions including severe bouts of inflation, record-high interest rates, and deep recessions. This period also coincided with the widespread adoption of floating exchange rates and, in a number of major industrial countries, with substantial institutional changes brought about by financial innovation and financial deregulation. The period since 1974 thus provided a very severe test of empirical money demand relationships. As we shall see, this period succeeded in exposing a number of shortcomings in existing specifications of money demand functions. Where institutional change was particularly marked, it also led to a change in what we think of as 'money'.

It is perhaps ironic that the emergence of these shortcomings roughly coincided with the adoption by a number of central banks of policies aimed at targeting monetary aggregates. Some have argued that this association is more than mere coincidence. In any event, given the vested interest of policy-makers in the existence of a reliably stable money demand function, it is hardly surprising that employees of central banks were among the most active contributors to the most recent literature on money demand. The Federal Reserve System of the United States, with its dominant market share of monetary economists, was particularly active in this regard.

As noted, appreciation of empirical research on money demand requires a bit of background on monetary theory and it is with this that we begin our discussion. We next consider some

measurement issues and then turn to the early empirical results. After briefly documenting the emerging difficulties with these results, we finally consider recent reformulations of the demand for money.

Theoretical Overview

One of the earliest approaches to the demand for money, the *quantity theory of money* starts with the *equation of exchange.* One version of the equation can be written.

$$MV \equiv PT \tag{1}$$

where M is the quantity of money, V is the velocity of circulation, P is the price level, and T is the volume of transactions. While M, P and T are directly measurable, V is implicitly defined by (1) so (1) is merely an identity. However, if we add the key assumption that velocity, V, is determined by technological and/or institutional factors and is therefore relatively constant, one can recast (1) as a demand function for money in which the demand for real balances, M/P, is proportional to T.

This simple demand for money function was modified by Keynes's (1936) analysis which introduced the speculative motive for holding money along with the transactions motive embodied in (1). The speculative motive views money and bonds as alternative assets with bond holding, in turn, viewed as depending on the rate of return on bonds. This introduction of the interest rate into the demand for money, where it joined the transactions variable suggested by the quantity theory is the main empirical legacy of Keynes. Once the interest rate is introduced, there is no presumption that velocity will be constant from period to period.

Post-Keynesian developments moved in several different directions. One is represented by Friedman (1956), whose restatement of the quantity theory dispensed with the individual motives posited by Keynes and treated money like any other asset yielding a flow of services. This view emphasized the level of wealth as one of the major determinants of money demand. Friedman also suggested that a quite broad range of opportunity cost variables including the expected rate of inflation have theoretical relevance in a money demand function. (Given this emphasis, it is ironic that Friedman's early empirical results (Friedman 1959) seemed to suggest that interest rates were unimportant in explaining velocity movements.)

While Friedman's approach sidestepped the explicit role of money in the transactions process, other influential post-Keynesian developments reconsidered and expanded on the transactions motive. William Baumol (1952) and James Tobin (1956) both applied inventory-theoretic considerations to the transactions demand for money. This led to the so-called *square-root law* with average money holdings given by

$$M = (2bT/r)^{1/2} \tag{2}$$

where r is the interest rate on bonds and b is the brokerage charge or transactions cost for converting bonds into cash. Dividing both sides of equation (2) by the price level, makes the real transactions demand for money depend on 'the' interest rate, real brokerage charges and the level of real transactions. Miller and Orr (1966) extended this analysis to allow for uncertainty in cash flows, providing the insight that a firm's average money holdings depends on the variance of its cash flow viewed as a measure of the uncertainty of the flow of receipts and expenditures.

Keynes's speculative motive has also been reformulated – largely in terms of portfolio theory (Tobin 1958). However, given the menu of assets available in most countries, this approach actually undermines the speculative demand for money. The reason is that if there is a riskless asset (e.g. a savings deposit) paying a higher rate of return than money (presumed to be zero in most models), then money is a dominated asset and will not be held. One can resurrect an asset demand for money by combining the portfolio approach with transaction costs but this has yet to be done in a fully general way. One partial attempt in this direction (Ando and Shell 1975) demonstrates that in a world with a riskless and a risky asset

the demand for money will not depend on the rate of return on the risky asset. This approach suggests using only a small number of interest rates, pertaining to riskless assets, in empirical work.

Some Measurement Issues

Empirical estimation of a money demand function requires choosing explicit variables measuring both money and its determinants. Even if guided by a particular theory, such choices are often less than clear-cut. Given the diversity of theories, the range of possible variables is wider yet. This is immediately evident when one considers how to measure 'money'; the sharp distinction between money and other assets turns out to be a figment of the textbook. Moreover, what passes for money can be readily altered by changing financial institutions.

In general, theories based on the transactions motive provide the most guidance and lead to a so-called *narrow* definition of money that includes currency and deposits transferable by cheque (also called checkable deposits). In some institutional settings a plausible measure of checkable deposits is readily apparent. In the United States, for example, for many years only demand deposits at commerical banks were checkable. In other settings, there may well be a spectrum of checkable assets without any clear-cut dividing line. For example, a deposit account may limit the number of cheques per month or may have a minimum cheque size. Other accounts may permit third-party transfers only if regular periodic payments are involved or may permit cheque writing only with substantial service charges. When such deposit accounts should be included in a transaction-based definition of money is not obvious.

Furthermore, even in a world in which the definition of checkable deposits is relatively unambiguous, it is not clear that currency and checkable deposits should be regarded as perfect substitutes, a view that is implicit in simply adding them together to produce a measure of money. Currency and checkable deposits may differ in transactions costs, risk of loss, and ease of concealment of illegal or tax-evading activities. It may thus be preferable to estimate separate demand functions for currency and checkable deposits.

Once one moves away from a transactions view of the world, the appropriate empirical definition of money is even less clear. A theory that simply posits that money yields some unspecified flow of services must confront the fact that many assets may yield these services in varying degrees. Such theories have typically relied on a relatively broad definition of money but the definitions utilized are inevitably somewhat arbitrary. (This issue is taken up again in section "Recent Reformulations".)

As with the definition of money, alternative theories have different implications for the relevant set of explanatory variables. As we have seen, the most prominent variables suggested by theory include the level of transactions, wealth, the opportunity cost of holding money, and transaction costs. Each of these involves measurement problems, even in a world of certainty. When uncertainty is allowed for, and expectational issues therefore arise, matters are even worse.

The level of transactions (T in equation (2)) is typically measured by the level of income or gross national product (GNP). While the term 'gross' in GNP makes it sound comprehensive, GNP is much less inclusive than a general measure of transactions. In particular, it excludes all sales of intermediate goods, purchases of existing goods, and financial transactions, all of which may contribute to a demand for money. The empirical use of GNP as a proxy for T therefore presumes that GNP and T move in a proportionate way. Unfortunately, this key assumption is extremely difficult to test because reliable data on T are nonexistent. (Moreover, it is not the case that all transactions are equally 'money intensive'. To cope with this empirically might require separately introducing the various components of T or, as an approximation, of GNP.)

As an alternative to GNP, some researchers have used permanent income, typically measured as an exponentially weighted average of current and past-values of GNP. This is generally done in the spirit of the modern quantity theory where permanent income is a proxy for wealth. As an empirical matter, given the high correlation of

GNP and permanent income, a permanent income variable could easily 'work' even if money demand is dominated by transactions considerations. One can, of course, use a measure of wealth directly (only non-human wealth is readily available). This is certainly consistent with the quantity theory view and, given that financial transactions may generate a demand for money, can fit into a transactions view.

Before leaving measures of transactions, we should note one further problem that arises because of issues of aggregation. Most theories of the demand for money apply to an individual behavioural unit but are generally estimated with aggregate data without much attention to the details of aggregation. This failure may lead to the omission of potentially important variables. For example, in the context of a transactions variable, aggregation may suggest that the distribution of income, as well as the level of income, matters. However, with a few exceptions discussed below, we shall not focus on problems of aggregation.

Another set of measurement issues is presented by the opportunity cost of holding money. We consider in turn the two parts to this story: the rate of return on assets alternative to money; and the own rate of return on money. Under the transactions view, the relevant alternative is a 'bond' that is used as a temporary repository of funds soon to be disbursed. As a practical matter this has led to the use of one or more of the following rates: the yield on short-term government securities; the yield on short-term commercial paper; and the yield on time or savings deposits. As we have seen, the relevant set of alternatives under the modern quantity theory is much broader and empirical research in this spirit has also used long-term bond rates, either government or corporate. Indeed, a few studies have attempted to use proxies for the entire term structure of interest rates. In addition, some investigators use the rate of return on corporate equities and/or the expected rate of inflation.

The own rate of return on money obviously depends on the concept of money chosen for analysis. The seemingly simplest case occurs with a narrow definition of money that bears an explicit zero rate of return. In such cases, most investigators have treated the own rate of return as zero. This, however, is not precisely correct since holders of deposits may earn an implicit rate of return, either because they receive services or because service charges may be foregone as the level of deposits rises. Measuring this implicit return is no easy matter. Matters are considerably more complicated when broader definitions of money are used and some components of money bear explicit interest, especially when there are several components each carrying a different rate of return. The aggregate own rate of return would then be a complex function of the interest rates, shares, and elasticities of each of the components. For the most part, researchers have not faced this issue squarely. However, the advent of interest-bearing checkable deposits that exist alongside zero-return demand deposits means that even those using narrow definitions of money must address this issue.

A final variable that appears prominently in equation (2) is the transactions cost, *b*. This is sometimes interpreted as the brokerage charge for selling 'bonds' or as the 'shoe-leather' cost of going to the bank. Whatever the interpretation, however, such variables have generally been conspicuous by their absence from empirical work. Researchers have thus implicitly assumed that real transactions costs are constant. The validity of this assumption has grown increasingly questionable as innovation and technical change have spread through the financial sector. Unfortunately, there are only highly imperfect proxies available to measure *b*. The consequences of this are examined below.

Empirical Findings: The Early Results

Before considering empirical results, a word needs to be said about the types of data that have been used. While there have been some cross-section studies using data at a variety of levels of aggregation, the vast majority of available studies employ highly aggregated time series data. Initially these were confined to annual observations, but increasingly the focus has been on shorter

periods such as quarterly, monthly, or even weekly data. In part this shift stems from the availability of short-period data but, more importantly, from the related perception that the quarterly or monthly time frame is more useful for guiding monetary policy.

The earliest empirical work in monetary economics primarily involved producing estimates of velocity, characterizing its behaviour over time and identifying the institutional factors responsible for longer-run movements in velocity. (For a discussion of this literature, see Selden 1956.) Modern empirical studies of money demand first appeared a few years after the publication of Keynes's *General Theory* in 1936. Not surprisingly, these studies focused on testing the prediction of the hypothesis of liquidity preference that there was an inverse relationship between the demand for money and the interest rate. One approach to this problem was to establish a positive correlation between interest rates and velocity.

A second approach involved distinguishing between 'active' and 'idle' balances and then relating idle balances to the interest rate. Conceptually this amounted to positing a demand function for money of the form

$$M/P = ky + f(r) \tag{3}$$

where γ is income or GNP. With k assumed known, idle balances, given by $(M/P - ky)$, can then be related to r. Tobin (1947), using data from 1922 to 1945, calculated k by assuming idle balances were zero in 1929 and found a relatively close relationship between idle balances and r of a roughly hyberbolic shape. Of course, as was recognized at the time, there is an element of arbitrariness in the definition of idle balances, and it is a short step to estimate equation (3) directly, obviating the necessity of distinguishing between active and idle balances. Indeed, this approach had already been suggested in 1939 by A. J. Brown who estimated a variant of (3). (Brown's paper, which is surprisingly modern, both conceptually and statistically, is also noteworthy for the inclusion of the rate of inflation in the demand for money.)

Initially at least, typical estimates of the demand-for-money function were based on annual data and used a log-linear specification, which has constant elasticities. Thus, a typical equation used in empirical work was of the form

$$\ln(M_t/P_t) - \ln(M_{t-1}/P_{t-1}) = \gamma\left[\left(M_t^* P_t\right) - \ln(M_{t-1}/P_{t-1})\right] \tag{4}$$

As before, γ is a scale variable such as income or wealth and r represents the interest rate. Sometimes several scale variables or interest rates were used; additional variables were also included on occasion. From the late 1950s on many studies estimated equations like (4) for a number of countries. These studies differed in terms of the sample period (sometimes going back as far as the late 1800s) and the specific choice of dependent and independent variables. While these studies hardly produced identical conclusions, at least through the early 1970s a number of common findings did emerge. For the United States (see Laidler 1977): (1) Various interest rates – sometimes several at once – proved to be of statistical significance in (4) with elasticities of short-term and long-term rates generally ranging from – 0.1 to – 0.2 and – 0.2 to – 0.8, respectively. (2) Income, either measured or permanent, and non-human wealth all achieved statistical significance, although typically only when these variables were included one at a time. Some studies viewed the matter as a contest between these several variables, the winner often depending on the sample period, the definition of M, and econometric details. Estimated scale elasticities ranged from about $\frac{1}{2}$ to nearly 2, but most estimates were in the lower end of the range. (3) As judged by a variety of procedures, both formal and informal, the demand function for money exhibited a reasonable amount of stability over time.

While many of the early studies using annual data tended to ignore dynamic aspects of the specification, a number did address this issue, most frequently by the simple device of including a lagged dependent variable in the money demand equation. One rationale for this is the partial adjustment model, which posits the existence of

a 'desired' level of real money balances M^*/P, and further assumes that the actual level of money balances adjusts in each period only part of the way toward its desired level. This idea is captured in the logarithmic adjustment equation

$$\ln(M_t/P_t) - \ln(M_{t-1}/P_{t-1}) = \gamma\left[\ln\left(M_t^* P_t\right) - \ln(M_{t-1}/P_{t-1})\right] \quad (5)$$

where M_t/P_t denotes the actual value of real money balances. The parameter γ governs the speed of adjustment; $\gamma = 1$ corresponds to complete adjustment in one period (i.e. $M_t = M_t^*$). Implementation of (5) is achieved by expressing M_t^*/P_t as a function of y_t and r_t as in (4) and substituting into (5). The resulting equation gives M_t/P_t as a function of y_t, r_t, and M_{t-1}/P_{t-1}. As we shall see below, the partial adjustment model is not without its shortcomings.

Not surprisingly, allowance for dynamics proved of particular importance once investigators began using quarterly data. Dynamics aside, results obtained with quarterly data generally confirmed the findings with annual data. Quarterly data did suggest it was preferable to work with narrow definitions of the money stock. Indeed, some studies suggested there was a further payoff to disaggregating the narrow money stock, either into its components (i.e. currency and checkable deposits) or by type of holder (e.g. household vs. business). On the whole, however, these refinements were not necessary to yield a serviceable quarterly money demand function. A simple specification in which real narrow money balances depended on GNP, a short-term market interest rate, a savings deposit rate, and lagged money balances appeared to be adequate for most purposes (Goldfeld 1973).

As the 1970s unfolded, however, this happy state of affairs unravelled. Difficulties were particularly pronounced with United States data, but instabilities appeared with equations for other countries as well (Boughton 1981; Goldfeld 1976). In the United States these difficulties first surfaced around 1974. Had past behaviour held up, the behaviour of real GNP and interest rates from the end of 1973 to the end of 1975 should have produced a mild decline in money demand in 1974 followed by a recovery in 1975. Instead, real money balances steadily declined, falling by about 7 per cent during this period. The economy seemed to be making do with less money. Or put another way, conventional money demand functions made sizeable and unprecedented overprediction errors. From 1974 to 1976 the cumulative drift was about 9 per cent. Another indication of the difficulty emerged when the post-1973 data were added to the estimation sample. Inclusion of the recent data tended to change the parameter estimates in the conventional money demand function, generally yielding quite unsatisfactory estimates. For example, the parameter γ tended to hover close to zero, implying implausibly long adjustment lags. These same difficulties were picked up by formal econometric tests that rejected the hypothesis that the structure of the money demand function had remained constant. Prior to 1974 these tests had given no indication of instability.

Stimulated by these difficulties, the last decade has witnessed a veritable outpouring of research on money demand. The primary emphasis has been on 'fixing' matters by improving the specification and/or using more appropriate econometric techniques. While progress has been made, even improved specifications have not proved immune from episodes of apparent instability.

Recent Reformulations

A substantial part of recent research has focused on the United States, but the issues are of general relevance for other countries. It should be noted that open-economy considerations, which have received only limited attention in the literature on the United States, would be more relevant for many other countries. On the other hand, the emphasis on financial innovation and deregulation in the case of the United States is probably of lesser importance for many countries.

The idea that financial innovation contributed to the instability of money demand in the United States stemmed from two observations: (1) the errant behaviour of money demand in the

mid-1970s appeared to be concentrated in business holdings of checkable deposits; and (2) marked improvements were evident in business cash management techniques. These improvements, including such arcane-sounding devices as cash concentration accounts, lockboxes and zero balance accounts, altered the nature of the transactions process and permitted firms to economize on the need for transactions balances. These improvements stemmed both from exogeneous technological innovations (e.g. in telecommunications) and from endogenous decisions whereby firms, stimulated by the high opportunity cost of holding cash, invested in new transactions technologies. In the context of the Baumol–Tobin inventory-theoretic model of money demand, those changes can be modelled as a reduction in transactions costs, *b,* while in the Miller–Orr variant one can view these innovations as reducing the uncertainty of receipts and expenditures. While early innovations in the United States appeared concentrated in the business sector, more recent innovations – such as money market mutual funds – and financial deregulation have affected households as well. (As an aside, it should be noted that the constraints of regulation stimulated financial innovation that in turn forced deregulation. To the extent that innovation and deregulation contributed to instability in money demand, regulation, which was in part aimed at improving the workings of monetary policy, sowed the seeds of later difficulties for policy.)

Explicit consideration of financial innovation in an econometric specification has, however, proved extremely difficult. The basic problem is that there are no reliable direct data on transactions costs. What indirect evidence there is stems from the use of time trends to capture exogenous technical change or of some function of previous peak interest rates as a proxy for endogenous reductions in transactions costs. The idea behind the latter variable is that high interest rates create an incentive to incur the fixed costs necessary to introduce a new technology but that once interest rates decline the technology remains in place. The use of a previous peak variable is meant to capture this irreversibility and researchers using such a variable have found that it improves the fit of money demand functions. Unfortunately, however, the resulting estimates do not appear very robust, either to small changes in specification or to the use of additional data. Some economists have played down the potential importance of financial innovations, pointing to the fact that high interest rates did not appear to stimulate the same degree of innovation in other countries. Nevertheless, most empirical researchers remain quite uneasy with their inability to capture adequately relevant changes in transactions costs since it raises the possibility of a continuing source of specification error.

Of course, financial innovation is not the only conceivable source of specification error, and when money demand functions began misbehaving, other elements of the conventional specification were re-examined. In particular, researchers again considered the use of alternative measures of transactions, wealth, and interest rates. They also relaxed the assumption of a constant elasticity implicit in equation (4) and re-examined the benefits of disaggregating money holdings by type of holder (e.g. business vs. households). In contrast with earlier work, these efforts suggested a greater role for wealth and some evidence on the importance of allowing for a nonconstant interest elasticity and for introducing a measure of the own rate of return on money. They also reconfirmed that there are gains to disaggregation by types of holder. Nevertheless, these improvements still left unexplained much of the aberrant behaviour of money demand.

Another approach was to reconsider the definition of money. Since a substantial volume of monetary data is available, economists who are unhappy with the official definitions are free to construct their own. Research along these lines has been in two diametrically opposed directions. The first has regarded the official definitions of even ‘narrow’ money as too broad, at least from a purely transactions point of view. This concern has led some to suggest using a disaggregated approach in which separate empirical demand functions are estimated for each monetary asset. This sidesteps the definitional issue and at the same time permits the use of econometric

techniques that take account of the interrelated nature of the demand functions. In practice, however, the application of this approach has been complicated by the appearance of new financial instruments brought about by deregulation, and such efforts have not been fully successful.

The second approach, noting that the line between transactions and other motives has become empirically murky, has considered whether relatively broad definitions of money could yield a stable demand function. However, conventional broad monetary aggregates obtained by simply adding together quantities of different assets are subject to the criticism that they combine components that offer differing degrees of monetary services. Consequently, most recent research along these lines has involved the weighting of the various components of a broad measure of money by the degree of 'moneyness' or 'liquidity' of each component. Although, the way in which this is done is inevitably somewhat arbitrary, in recent years some progress has been made in applying index-number theory to this issue (Barnett 1982). Indeed, the Federal Reserve now regularly publishes a number of such weighted money measures, sometimes called Divisia indexes. Thus far, this research seems to suggest that only the broadest of such monetary measures appear to yield a stable demand function. Even this result, however, is not without its difficulties. For one, a complete understanding of this result requires an economic explanation of the behaviour of the weights used to construct the measures. (Especially where the weights are based on relative velocity or turnover data, there appears to be some circularity in the construction of the measures that will give the appearance of stability.) Second, it is important for the results to be useful in formulating policy that these weights be forecastable. On the whole, while promising, the verdict on the Divisia approach is still out, either as an explanation of instability or for use in the policy process.

Yet another feature of money demand that has received recent attention is the dynamics of the adjustment process. As noted above, the so-called real partial adjustment model of equation (5) formed the basis of much early work. However, this model has come in for a wide variety of criticism. One aspect of this can be seen by rewriting (5) as follows:

$$\begin{aligned} &\ln M_t - \ln M_{t-1} \\ &= \gamma\left[\ln\left(M_t^*/P_t\right) - \ln(M_{t-1}/P_{t-1})\right] \\ &\quad + \Delta \ln P_t. \end{aligned} \tag{6}$$

As (6) shows, since the coefficient of $\Delta \ln P_t$ is unity, the specification presumes as immediate adjustment to changes in the price level. As this assumption seems unwarranted, more recent research has used the so-called nominal adjustment model given by

$$\ln M_t - \ln M_{t-1} = \gamma\left(\ln M_t^* - \ln(M_{t-1})\right). \tag{7}$$

Estimation of (7) is quite similar to (5) except that the variable M_{t-1}/P_t replaces the variable M_{t-1}/P_{t-1}. A variety of empirical tests suggest that the nominal model is to be preferred, but also indicate clearly that this change does not repair the money demand function.

Other re-examinations of dynamics have suggested that the simple partial adjustment model, either nominal or real, is more fundamentally flawed. Some writers point to the fact that the Miller–Orr transactions model predicts that money holders, facing a fixed cost of adjusting, will either make no adjustment or a complete adjustment. Partial adjustment would not be observed for an individual money holder. However, the applicability of this feature of the Miller–Orr model to aggregate data is not fully clear. Other attempts to derive an adjustment model from an optimizing framework have suggested models with a variable speed of adjustment with the speed parameter γ depending on income or interest rates. However, there has been only limited empirical work with such models.

Considerably more empirical work has been done with models where the speed of response of money holdings to some shock depends on which variable is producing the change in desired money holdings. This would accommodate the

suggestion that changes in real income, especially when such changes are paid in the form of money, should yield quicker adjustments of money holdings than changes in interest rates. To allow for these effects, one must relax the rigid geometrically distributed lag implicit in (5) or (7) and use instead a more general distributed lag specification. Data for the United States do seem to provide some support for this more general adjustment model but, as with other suggested improvements, this change is not sufficient to yield a single acceptable function that fits the post-World War II data.

A final attack on the partial adjustment model involves a more general reconsideration of the adjustment process. The point can be seen most clearly if we assume that the monetary authorities exogenously fix the nominal money supply. In such a world the desired nominal stock of money must adjust to the given stock, presumably by adjustments to variables influencing desired holdings. A particularly simple version of this idea would dispense with the partial adjustment model of (7) and replace it with an adjustment equation for prices as in

$$\ln P_t - \ln P_{t-1} = \lambda\left(\ln M_t - \ln M_t^*\right) \qquad (8)$$

While this obviates the need for a short-run money demand function, long-run money demand appears in (8) via M_t^*.

A variant of this approach would estimate the money demand function by imposing the assumption of rationality on price expectations. For example, one could begin with (4) or even (7) and use it to solve for the price level. Then, via the Fisher equation expressing the nominal rate of interest as the sum of the real rate and the expected rate of inflation, one can use the hypothesis of rational expectations to express the actual price level (or the rate of inflation) as a function of income, the money stock, and the real rate of interest. If we further posit the stochastic process for income, for the money stock (e.g. via a money supply rule) and for the real rate (e.g. the real rate is constant), we can use the resulting equation to estimate the parameters of the money demand function.

The estimation of money demand via (8) or its rational expectations variant is, however, not without its difficulties. One problem is that this approach implies that the inflation rate reacts quickly to changes in output or the money supply. Put another way, it assumes that the rate of inflation moves like an asset price determined in financial markets. This approach conflicts with the evidence of the stickiness of prices in response to shocks of various sorts. One way around this difficulty is to posit that the adjustments to 'disequilibrium' in the money market are effected in interest rates and/or output. (See Laidler and Bentley (1983), for a small model with these features.)

A second difficulty is the assumption that the money supply is exogenously set. For the United States, at least, the assumption seems most relevant for the period October 1979 to October 1982, the three years during which the Federal Reserve officially adopted monetary targeting. However, stated official policy notwithstanding, some have argued that the Federal Reserve never really pursued a policy of monetary targeting while others have suggested that such a policy began well before October 1979. This suggests that it is not always easy to identify changing monetary regimes. Nevertheless, it is clear that changes in the rules governing monetary policy can have implications for the proper specification and estimation of a money demand function. That is, conventional specifications may work in some circumstances but not others. Indeed, it has been suggested that failure to allow for this accounts for at least part of the apparent instability of conventional money demand functions (Gordon 1984).

While it is undoubtedly important to view the money demand function as part of a more complete system, to date this has not been empirically done in a satisfactory way. Part of the problem stems from the need to specify the money supply process in some detail; a task made difficult by changing policy strategies and deregulation. Moreover, there is yet another complication, the

question of the time unit of the analysis. Practitioners of monetary policy tend to have a relatively short decision-making horizon so that capturing the money supply process may require weekly or monthly data. In contrast, most money demand estimation has used quarterly or annual data. Put another way, proper attention to the dynamics of the monetary sector may require more care in the choice of the time unit of analysis. It may also require some sophisticated econometric techniques to perform estimation in the face of changing monetary regimes.

Conclusion

The current state of affairs finds the empirical money demand function to be in a bit of disarray, especially if one judges success by our ability to specify a single function that appears stable over the postwar period. To be sure, there are ample potential explanations – perhaps embarrassingly many – for the observed difficulties with conventional models. However, data inadequacies or econometric problems mean that it is not always easy to incorporate these explanations into an empirical demand function for money. Some have concluded from this that greater instability in money demand is a fact, not to be repaired in any simple way. It is the challenge of future research to overcome these difficulties. Given progress to date, it seems likely that further research will yield a more satisfactory statistical explanation of money demand. However, the flimsy nature of past apparent successes and the theoretical and empirical difficulties alluded to above alert us to the need for substantial scrutiny in evaluating new models. Ultimately, of course, such models need to stand the forward-looking test of time; that is, they need to continue to hold outside the period of estimation.

See Also

► Quantity Theory of Money
► Rational Expectations

Bibliography

Ando, A., and K. Shell. 1975. Demand for money in a general portfolio model. In *The Brookings model: Perspectives and recent developments*. Amsterdam: North-Holland.

Barnett, W. 1982. The optimum level of monetary aggregation. *Journal of Money, Credit, and Banking* 14(4): 687–710, Part II.

Baumol, W.J. 1952. The transactions demand for cash: An inventory theoretic approach. *Quarterly Journal of Economics* 66: 545–556.

Boughton, J.M. 1981. Recent instability of the demand for money: An international perspective. *Southern Economic Journal* 47(3): 579–597.

Brown, A.J. 1939. Interest, prices, and the demand schedule for idle money. *Oxford Economic Papers* 2: 46–69. Reprinted, in *Oxford studies in the price mechanism*, ed. T. Wilson and P. Andrews. Oxford: Clarendon Press, 1951.

Friedman, M. 1956. The quantity theory of money – A restatement. In *Studies in the quantity theory of money*, ed. M. Friedman. Chicago: University of Chicago Press.

Friedman, M. 1959. The demand for money: Some theoretical and empirical results. *Journal of Political Economy* 67: 327–351.

Goldfeld, S.M. 1973. The demand for money revisited. *Brookings Papers on Economic Activity* (3): 577–638.

Goldfeld, S.M. 1976. The case of the missing money. *Brookings Papers on Economic Activity* (3): 683–730.

Gordon, R.J. 1984. The short-run demand for money: A reconsideration. *Journal of Money, Credit, and Banking* 16(4): 403–434, Part I.

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

Laidler, D.E.W. 1977. *The demand for money: Theories and evidence*. New York: Dun-Donnelley.

Laidler, D.E.W., and B. Bentley. 1983. A small macromodel of the post-war United States. *Manchester School of Economics and Social Studies* 51(4): 317–340.

Miller, M.H., and D. Orr. 1966. A model of the demand for money by firms. *Quarterly Journal of Economics* 80: 413–435.

Selden, R. 1956. Monetary velocity in the United States. In *Studies in the quantity theory of money*, ed. M. Friedman. Chicago: University of Chicago Press.

Tobin, J. 1947. Liquidity preference and monetary policy. *Review of Economics and Statistics* 29: 124–131.

Tobin, J. 1956. The interest-elasticity of transactions demand for cash. *Review of Economics and Statistics* 38: 241–247.

Tobin, J. 1958. Liquidity preference as behavior towards risk. *Review of Economic Studies* 25: 65–86.

Demand for Money: Theoretical Studies

Bennett T. McCallum and Marvin S. Goodfriend

In any discussion of the demand for money it is important to be clear about the concept of money that is being utilized; otherwise, misunderstandings can arise because of the various possible meanings that readers could have in mind. Here the term will be taken to refer to an economy's *medium of exchange:* that is, to a tangible asset that is generally accepted in payment for any commodity. Money thus conceived will also serve as a store of value, of course, but may be of minor importance to the economy in that capacity. The monetary asset will usually also serve as the economy's medium of account – that is, prices will be quoted in terms of money – since additional accounting costs would be incurred if the unit of account were a quantity of some asset other than money. The medium-of-account role is, however, not logically tied to the medium of exchange (Wicksell 1906; Niehans 1978).

Throughout much of Western history, most economies have adopted as their principal medium of exchange a commodity that would be valuable even if it were not used as money. Recently, however, fiat money – intrinsically worthless tokens made of paper or some other cheap material – has come to predominate. Under a commodity money arrangement, the exchange value of money will depend upon the demand for the monetary commodity in its non-monetary as well as its monetary uses. But in a discussion of money demand, as distinct from a discussion of the price level, any possible non-monetary demand for the medium of exchange – which will be absent anyhow in fiat money system – can legitimately be ignored.

The quantity of money demanded in any economy – indeed, the set of assets that have monetary status – will be dependent upon prevailing institutions, regulations and technology. Technical progress in the payments industry will, for instance, tend to alter the quantity of money demanded for given values of determinants such as income. This dependence does not, however, imply that the demand for money is a nebulous or unusable concept, any more than the existence of technical progress and regulatory change in the transportation industry does so for the demand for automobiles. In practice, some lack of clarity pertains to the operational measurement of the money stock, as it does to the stock of automobiles or other commodities. But in an economy with a well-established national currency, the principle is relatively clear: assets are part of the money stock if and only if they constitute *claims* to currency, unrestricted legal claims that can be promptly and cheaply exercised (at par). This principle rationalizes the common practice of including demand deposits in the money stock of the United States, while excluding time deposits and various other assets.

The rapid development during the 1960s and 1970s of computer and telecommunications technologies has led some writers (e.g. Fama 1980) to contemplate economies – anticipated by Wicksell (1906) – in which virtually all purchases are effected not by the transfer of a tangible medium of exchange, but by means of signals to an accounting network, signals that result in appropriate debits and credits to the wealth accounts of buyers and sellers. If there were literally *no* medium of exchange, the wealth accounts being claims to some specified bundle of commodities, the economy in question would be properly regarded and analysed as a non-monetary economy, albeit one that avoids the inefficiencies of crude barter. If, by contrast, the accounting network's credits were claims to quantities of a fiat or commodity medium of exchange, then individuals' credit balances would appropriately be included as part of the money stock (McCallum 1985).

Basic Principles

An overview of the basic principles of money demand theory can be obtained by considering a

hypothetical household that seeks at time t to maximize

$$u(c_t, l_t) + \beta u(c_{t+1}, l_{t+1}) + \beta^2 u(c_{t+2}, l_{t+2}) + \ldots \quad (1)$$

where c_t and l_t are the household's consumption and leisure during t and where, $\beta = 1/(1 + \delta)$, with $\delta > 0$ the rate of time preference. The within-period utility function, $u(\cdot,\cdot)$ is taken to be well behaved so that unique positive values will be chosen for c_t and l_t. The household has access to a productive technology described by a production function that is homogeneous of degree one in capital and labour inputs. But for simplicity we assume that labour is supplied inelastically, so this function can be written as $y_t = f(k_{t-1})$, where y_t is production during t and k_{t-1} is the stock of capital held at the end of period $t-1$. The function $f(\cdot)$ is well behaved, so a unique positive value of k_t will be chosen for the upcoming period. Capital is unconsumed output, so its price is the same as that of the consumption good and its rate of return between t and t + 1 is $f'(k_t)$.

Although this set-up explicitly recognizes the existence of only one good, it is intended to serve a simplified representation – one formally justified by the analysis of Lucas (1980) – of an economy in which the household sells its specialized output and makes purchases (at constant relative prices) of a large number of distinct consumption goods. Carrying out these purchases requires *shopping time,* s_t, which subtracts from leisure: $l_t = 1 - s_t$, where units are chosen so that there is 1 unit of time per period available for shopping and leisure together. (If labour were elastically supplied, then labour time would have to be included in the expression.) In a monetary economy, however, the amount of shopping time required for a given amount of consumption will depend negatively upon the quantity of real money balances held by the household (up to some satiation level). For concreteness, we assume that

$$s_t = \psi(c_t, m_t) \quad (2)$$

where $\psi(\cdot,\cdot)$ has partial derivatives $\psi_1 > 0$ and $\psi_2 \leq 0$. In (2), $m_t = M_t/P_t$, where M_t is the nominal stock of money held at the end of t and P_t is the money price of a consumption bundle. (A variant with M_t denoting the start-of-period money stock will be mentioned below.) The transaction variable is here specified as c_t rather than $c_t + \Delta k_t$ to reflect the idea that only a few distinct capital goods will be utilized, so that the transaction cost to expenditure ratio will be much lower than for consumption goods.

Besides capital and money, there is a third asset available to the household. This asset is a nominal bond; i.e., a one-period security that may be purchased at the price $1/(1 + R_t)$ in period t and redeemed for one unit of money in $t + 1$. The symbol B_t will be used to denote the number (possibly negative) of these securities purchased by the household in period t, while $b_t = B_t/P_t$.

In the setting described, the household's budget constraint for period t may be written as follows:

$$\begin{aligned} f(k_{t-1}) + v_t \geq c_t + k_t - k_{t-1} + m_t \\ - (1 + \pi_t)^{-1} m_{t-1} \\ + (1 + R_t)^{-1} b_t \\ - (1 + \pi_t)^{-1} b_{t-1} \end{aligned} \quad (3)$$

Here v_t is the real value of lump-sum transfers (net of taxes) from the government, while π_t is the inflation rate, $\pi_t = (P_t - P_{t-1})/P_{t-1}$. Given the objective of maximizing (1), first-order conditions necessary for optimality of the household's choices include the following, in which φ_t and λ_t are Lagrangian multipliers associated with the constraints (2) and (3), respectively:

$$u_1(c_1, 1 - s_t) - \phi_t \psi_1(c_t, m_t) - \lambda_t = 0 \quad (4)$$

$$-u_2(c_t, 1 - s_t) + \phi_t = 0 \quad (5)$$

$$-\phi_t \psi_2(c_t, m_t) - \lambda_1 + \beta \lambda_{t+1}(1 + \pi_{t+1})^{-1} = 0 \quad (6)$$

$$-\lambda_t + \beta \lambda_{t+1}[f'(k_t) + 1] = 0 \quad (7)$$

$$-\lambda_t (1 + R_t)^{-1} + \beta \lambda_{t+1}(1 + \pi_{t+1})^{-1} = 0 \quad (8)$$

These conditions, together with the constraints (2) and (3), determine current and planned values of c_t, s_t, m_t, k_t, b_t, φ_t, and λ_t for given time paths of v_t, R_t, and π_t (which are exogenous to the household) and the predetermined values of k_{t-1}, m_{t-1}, and b_{t-1}. (There is also a relevant transversality condition, but it can be ignored for the issues at hand.) Also l_t values can be obtained from $l_t = 1 - s_t$ and, with P_{t-1} given, P_t, M_t, and B_t values are implied by the π_t, m_t, and b_t sequences.

The household's optimizing choice of m_t can be described in terms of two distinct concepts of a money-demand function. The first of these is a proper demand function; that is, a relationship giving the chosen quantity as a function of variables that are either predetermined or exogenous to the economic unit in question. In the present context, the money-demand function of that type will be of the form:

$$m_t = \mu(k_{t-1}, m_{t-1}, b_{t-1}, v_t, v_{t+1}, \ldots, R_t, R_{t+1}, \ldots, \pi_t, \pi_{t+1}, \ldots) \tag{9}$$

where the variables dated $t + 1, t + 2, \ldots$ must be understood as anticipated values. Now, it will be obvious that this relationship does not closely resemble those normally described in the literature as 'money demand functions'. There is a second type of relationship implied by the model, however, that does have such a resemblance. To obtain this second expression, one can eliminate ${}^{\beta\lambda}{}_{t+1}({}^{1+}\pi_{t+1})^{-1}$ between equations (6) and (8), then eliminate λ_t and finally φ_t from the resultant by using (4) and (5). These steps yield the following:

$$-u(c_t, 1 - s_t)\psi_2(c_t, m_t) = [u_1(c_t, 1 - s_t) - u_2(c_t, 1 - s_t)\psi_1(c_t, m_t)] \times \left[1 - (1 + R_t)^{-1}\right]. \tag{10}$$

Then $\psi(c_t, m_t)$ can be used in place of s_t, and the result is a relationship that involves *only* m_t, c_t, and R_t. Consequently, (10) can be expressed in the form:

$$f(m_t, c_t, R_t) = 0 \tag{11}$$

and if the latter is solvable for m_t one can obtain:

$$M_t / P_t = L(c_t, R_t). \tag{12}$$

Thus the model at hand yields a *portfolio-balance* relationship between real money-balances demanded, a variable measuring the volume of transactions conducted, and the nominal interest rate (which reflects the cost of holding money rather than bonds). It can be shown, moreover, that for reasonable specifications of the utility and shopping-time functions, will be increasing in its first argument and decreasing in the second. $L(\cdot,\cdot)$

There are, of course, two problems in moving from a demand function (of either type) for an individual household to one that pertains to the economy as a whole. The first of these involves the usual problem of aggregating over households that may have different tastes and/or levels of wealth. It is well known that the conditions permitting such aggregation are extremely stringent in the context of any sort of behavioural relation; but for many theoretical purposes it is sensible to pretend that they are satisfied. The second problem concerns the existence of economic units other than households – 'firms' being the most obvious example. To construct a model analogous to that above for a firm, one would presumably posit maximization of the present value of real net receipts rather than (1), and the constraints would be different. In particular, the shopping-time function (2) would need to be replaced with a more general relationship depicting resources used in conducting transactions as a function of their volume and the real quantity of money held. The transaction measure would not be c_t for firms or, therefore, for the economy as a whole. But the general aspects of the analysis would be similar, so we shall proceed under the presumption that the crucial issues are adequately represented in a setting that recognizes only economic units like the 'households' described above.

The distinction between the proper money-demand function (9) and the more standard portfolio-balance relation (12) is important in the context of certain issues. As an example, consider

the issue of whether wealth or income should appear as a 'scale variable' (Meltzer 1963). From the foregoing, it is clear that wealth is an important determinant of money demand in the sense that k_{t-1}, m_{t-1}, and b_{t-1} are arguments of the demand function (9). Nevertheless, formulation (12) indicates that there is no separate role for wealth in a portfolio-balance relation if appropriate transaction and opportunity-cost variables are included.

An issue that naturally arises concerns the foregoing discussion's neglect of randomness. How would the analysis be affected if it were recognized that future values of variables cannot possibly be known with certainty? In answer, let us suppose that the household knows current values of all relevant variables including P_t, R_t, and v_t when making decisions on m_t and c_t, but that its views concerning variables dated $t+1$, $t+2,\ldots$ are held in the form of non-degenerate probability distributions. Suppose also that there is uncertainty in production, so that the marginal product of capital in $t+1$, $f'(kt)$, is viewed as random. Then the household's problem becomes one of maximizing the expectation of (1), with $u(\cdot,\cdot)$ a von Neumann-Morgenstern utility function, given information available in period t. Consequently, the first-order conditions (4), (5), (6), (7), and (8) must be replaced with ones that involve conditional expectations. For example, equation (7) would be replaced with:

$$-\lambda_t + \beta E_t\{\lambda_{t+1}[f'(k_t)+1]\} = 0 \qquad (7')$$

where $E_t(\cdot)$ denotes the expectation of the indicated variable conditional upon known values of P_t, R_t, v_t, and so on. With this modification, the nature of the proper demand function becomes much more complex – indeed, for most specifications no closed form solution analogous to (9) will exist. Nevertheless, the portfolio–balance relation (12) will continue to hold exactly as before, for the steps described in its derivation above remain the same except that it is $E_t[\beta\lambda_{t+1}(1+\pi_{t+1})^{-1}]$ that is eliminated between equations corresponding to (6) and (8). From this result it follows that, according to our model, the relationship of M_t/P_t to the transaction and opportunity–cost variables is invariant to changes in the probability distribution of future variables.

Another specification variant that should be mentioned reflects the assumption that it is money held at the start of a period, not its end, that facilitates transactions conducted during the period. If that change in specification were made and the foregoing analysis repeated, it would be found that the household's concern in period t would be to have the appropriate level of real money balances at the start of period $t+1$. The portfolio–balance relation analogous to (12) that would be obtained in the deterministic case would relate m_{t+1} to c_{t+1} and R_t, where $m_{t+1} = M_{t+1}/P_{t+1}$ with M_{t+1} reflecting money holdings at the end of period t. Consequently, M_{t+1}/P_t would be related to R_t, planned c_{t+1}, and P_t/P_{t+1}. Thus the theory does not work out as cleanly as in the case considered above even in the absence of randomness, and is complicated further by the recognition of the latter. The fundamental nature of the relationships are, however, the same as above.

Another point deserving of mention is that if labour is supplied elastically, the portfolio-balance relation analogous to (12) will include the real wage-rate as an additional argument. This has been noted by Karni (1973) and Dutton and Gramm (1973). More generally, the existence of other relevant margins of substitution can bring in other variables. If stocks of commodities held by households affect shopping-time requirements, for example, the inflation rate will appear separately in the counterpart of (12) (see Feige and Parkin 1971).

Finally, it must be recognised that the simplicity of the portfolio-balance relation (12) would be lost if the intertemporal utility function (1) were not time-separable. If, for example, the function $u(c_t, l_t)$ in (1) were replaced with $u(c_t, l_t, l_{t-1})$ or $u(c_t, c_{t-1}, l_t)$, as has been suggested in the business cycle literature, then the dynamic aspect of the household's choices would be more complex and a relation like (12) – i.e. one that includes only contemporaneous variables – could not be derived.

Historical Development

The approach to money-demand analysis outlined above, which features intertemporal optimization choices by individual economic agents whose transactions are facilitated by their holdings of money, has evolved gradually over time. In this section we briefly review that evolution.

While the earlier literature on the quantity theory of money contained many important insights, its emphasis was on the comparison of market equilibria rather than individual choice; that is, on 'market experiments' rather than 'individual experiments', in the language of Patinkin (1956). Consequently, there was little explicit consideration of money-demand behaviour in pre-1900 writings in the quantity theory tradition. Indeed, there was little emphasis on money demand *per se* even in the classic contributions of Mill (1848), Wicksell (1906) and Fisher (1911), despite the clear recognition by those analysts that some particular quantity of real money holdings would be desired by the inhabitants of an economy under any specified set of circumstances. Notable exceptions, discussed by Patinkin (1956, pp. 386–417), were provided by Walras and Schlesinger.

In the English language literature, the notion of money demand came forth more strongly in the 'cash balance' approach of Cambridge economists, an approach that featured analysis organized around the concepts of money demand and supply. This organizing principle was present in the early (*c* 1871) but unpublished writings of Marshall (see Whitaker 1975, p. 165–8) and was laid out with great explicitness by Pigou (1917). The Cambridge approach presumed that the quantity of money demanded would depend primarily on the volume of transactions to be undertaken, but emphasized volition on the part of money-holders and recognized (sporadically) that the ratio of real balances to transaction volume would be affected by foregone 'investment income' (i.e., interest earnings). In this regard Cannan (1921), a non-Cambridge economist who was influenced by Marshall, noted that the quantity of money demanded should be negatively related to anticipated inflation – an insight previously expressed by Marshall in his testimony of 1886 for the Royal Commission on the Depression of Trade and Industry (Marshall 1926). In addition, Cannan developed very clearly the point that the relevant concept is the demand for a *stock* of money.

Although the aforementioned theorists developed several important constituents of a satisfactory money-demand theory, none of them unambiguously cast his explanation in terms of marginal analysis. Thus a significant advance was provided by Lavington (1921, p. 30), in a chapter entitled 'The Demand for Money', who attempted a statement of the marginal conditions that must be satisfied for optimality by an individual who consumes, holds money, and holds interest-bearing securities. But despite the merits of his attempt, Lavington confused – as Patinkin (1956, p. 418) points out – the subjective sacrifice of permanently adding a dollar to cash balances with that of adding it for only one period. Thus it was left for Fisher (1930, p. 216) to provide a related but correct statement. The discussions of both Lavington and Fisher are notable for identifying the interest rate as a key determinant of the marginal opportunity cost of holding money.

In a justly famous article, Hicks (1935) argued persuasively that progress in the theory of money would require the treatment of money demand as a problem of individual choice at the margin. Building upon some insightful but unclear suggestions in Keynes's *Treatise on Money* (1930), Hicks investigated an agent's decision concerning the relative amounts of money and securities to be held at a point in time. He emphasized the need to explain why individuals willingly hold money when its return is exceeded by those available from other assets and – following Lavington and Fisher – concluded that money provides a service yield not offered by other assts. Hicks also noted that the positive transaction cost of investing in securities makes it unprofitable to undertake such investments for very short periods. Besides identifying the key aspects of marginal analysis of money demand. Hicks (1935) pointed out that an individual's total wealth will influence his demand for money. All of these points were

developed further in chapters 13 and 19 of Hicks's *Value and Capital* (1939). The analysis in the latter is, some misleading statements about the nature of interest notwithstanding, substantively very close to that outlined in the previous section of this article. Hicks did not, however, provide formal conditions relating to money demand in his mathematical appendix.

The period between 1935 and 1939 witnessed, of course, the publication of Keynes's *General Theory* (1936). That work emphasized the importance for macroeconomic analysis of the interest-sensitivity of money demand – 'liquidity preference', in Keynes's terminology– and was in that respect, as in many others, enormously influential. Its treatment of money demand *per se* was not highly original, however, in terms of fundamentals. (This statement ignores some peculiarities resulting from a presumably inadvertent attribution of money illusion; on this topic, again see Patinkin 1956, pp. 173–4.)

The importance of several items mentioned above – payments practices, foregone interest and transaction costs – was explicitly depicted in the formal optimization models developed several years later by Baumol (1952) and Tobin (1956). These models, which were suggested by mathematical inventory theory, assume the presence of two assets (money and an interest-bearing security), a fixed cost of making transfers between money and the security, and a lack of synchronization between (exogenously given) receipt and expenditure streams. In addition, they assume that all payments are made with money. Economic units are depicted as choosing the optimal frequency for money-security transfers so as to maximize interest earnings net of transaction costs.

In Baumol's treatment, which ignores integer constraints on the number of transactions per period, the income and interest-rate elasticities of real money demand are found to be ½ and – ½, respectively. Thus the model implies 'economies of scale' in making transactions. Tobin's (1956) analysis takes account of integer constraints, by contrast, and thus implies that individuals respond in a discontinuous fashion to alternative values of the interest rate. In his model it appears entirely possible for individual economic units to choose corner solutions in which none of the interest-bearing security is held. A number of extensions of the Baumol–Tobin approach have been made by various authors; for an insightful survey the reader is referred to Barro and Fischer (1976).

Miller and Orr (1966) pioneered the inventory approach to money demand theory in a stochastic context. Specifically, in their analysis a firm's net cash inflow is generated as a random walk, and the firm chooses a policy to minimize the sum of transaction and foregone-interest costs. The optimal decision rule is of the (*S, s*) type: when money balances reach zero or a ceiling, *S,* the firm makes transactions to return the balance to the level *s.* In this setting there are again predicted economies of scale, while the interest-rate elasticity is – 1/3. For extensions the reader is again referred to Barro and Fisher (1976).

The various inventory models of money demand possess the desirable feature of providing an explicit depiction of the *source* of money's service yield to an individual holder. It has been noted (e.g. by Friedman and Schwartz 1970) that the type of transaction demand described by these models is unable to account for more than a fraction of the transaction balances held in actual economies. Furthermore, their treatment of expenditure and receipt streams as exogenous is unfortunate and they do not generalize easily to fully dynamic settings. These points imply, however, only that the inventory models should not be interpreted too literally. In terms of fundamentals they are closely related to the basic model outlined in the previous section.

A quite different approach was put forth by Tobin (1958), in a paper that views the demand for money as arising from a portfolio allocation decision made under conditions of uncertainty. In the more influential of the paper's models, the individual wealth-holder must allocate his portfolio between a riskless asset, identified as money, and an asset with an uncertain return whose expected value exceeds that of money. Tobin shows how the optimal portfolio mix depends, under the assumption of expected utility maximization, on the individual's degree of risk aversion, his wealth, and the mean-variance

characteristics of the risky asset's return distribution. The analysis implies a negative interest sensitivity of money demand, thereby satisfying Tobin's desire to provide an additional rationalization of Keynes's (1936) liquidity-preference hypothesis. The approach has, however, two shortcomings. First, in actuality money does not have a yield that is riskless in real terms, which is the relevant concept for rational individuals. Second, and more seriously, in many actual economies there exist assets 'that have precisely the same risk characteristics as money and yield higher returns' (Barro and Fischer 1976, p. 139). Under such conditions, the model implies that no money will be held.

Another influential item from this period was provided by Friedman's well-known 'restatement' of the quantity theory (1956). In that paper, as in Tobin's, the principal role of money is as a form of wealth. Friedman's analysis emphasized margins of substitution between money and assets other than bonds (e.g. durable consumption goods and equities). The main contribution of the paper was to help rekindle interest in monetary analysis from a macroeconomic perspective, however, rather than to advance the formal theory of money demand.

A model that may be viewed as a formalization of Hicks's (1935, 1939) approach was outlined by Sidrauski (1967). The main purpose of Sidrauski's paper was to study the interaction of inflation and capital accumulation in a dynamic context, but his analysis gives rise to optimality conditions much like those of equations (4), (5), (6), (7), and (8) of the present article and thus implies money-demand functions like (9) and (12). The main difference between Sidrauski's model and ours is merely due to our use of the 'shopping time' specification, which was suggested by Saving (1971). That feature makes real balances an argument of each individual's utility function only indirectly, rather than directly, and indicates the type of phenomenon that advocates of the direct approach presumably have in mind. Thus Sidrauski's implied money-demand model is the basis for the one presented above, while a stochastic version of the latter, being fundamentally similar to inventory or direct utility-yield specifications, is broadly representative of current mainstream views.

Ongoing Controversies

Having outlined the current mainstream approach to money-demand analysis and its evolution, we now turn to matters that continue to be controversial. The first of these concerns the role of uncertainty. In that regard, one point has already been developed; i.e., that rate-of-return uncertainty on other assets cannot be used to explain why individuals hold money in economies – such as that of the US – in which there exist very short-term assets that yield positive interest and are essentially riskless in nominal terms. But this does not imply that uncertainty is unimportant for money demand in a more general sense, for there are various ways in which it can affect the analysis. In the basic model outlined above, uncertainty appears explicitly only by way of the assumption that households view asset returns as random. In that case, if money demand and consumption decisions for a period are made simultaneously then the portfolio-balance relation (12) will be – as shown above – invariant to changes in the return distributions. But the same is not true for the proper demand function (9). And the arguments c_t and R_t of (12) will themselves be affected by the extent of uncertainty, for it will affect households' saving, as well as portfolio, decisions. The former, of course, impact not only on c_t but also on the economy's capital stock and thus, via the equilibrium real return on capital, on R_t. In addition, because R_t is set in nominal terms, its level will include a risk differential for inflation uncertainty (Fama and Farber 1979).

Furthermore, the invariance of (12) to uncertainty breaks down if money must be held at the start of a period to yield its transaction services during that period. In this case, the money demand decision temporally precedes the related consumption decision so the marginal service yield of money is random, with moments that depend on the covariance matrix of forecast errors for consumption and the price level. Thus the extent of uncertainty, as reflected in this covariance

matrix, influences the quantity of real balances demanded in relation to R_t and plans for c_{t+1}.

There is, moreover, another type of uncertainty that is even more fundamental than rate-of-return randomness. In particular, the existence of uncertainty regarding exchange opportunities available at an extremely fine level of temporal and spatial disaggregation – uncertainties regarding the 'double coincidence of wants' in meetings with potential exchange partners – provides the basic *raison d'être* for a medium of exchange. In addition, the ready verifiability of money enhances the efficiency of the exchange process by permitting individuals to economize on the production of information when there is uncertainty about the reputation of potential trading partners. Thus uncertainty is crucial in explaining why it is that money holdings help to facilitate transactions – to save 'shopping time' in our formalization. In this way randomness is critically involved, even when it does not appear explicitly in the analysis. (Alternative treatments of uncertainty in the exchange process have been provided by Patinkin 1956; Brunner and Meltzer 1971; King and Plosser 1986).

An important concern of macroeconomists in recent years has been to specify models in terms of genuinely structural relationships; that is, ones that are invariant to policy changes. This desire has led to increased emphasis on explicit analysis of individuals' dynamic optimization problems, with these expressed in terms of basic taste and technology parameters. Analysis of that type is especially problematical in the area of money demand, however, because of the difficulty of specifying rigorously the precise way – at a 'deeper' level than (2), for example – in which money facilitates the exchange process. One prominent attempt to surmount this difficulty has featured the application of a class of overlapping-generations models – i.e. dynamic equilibrium models that emphasize the differing perspectives on saving of young and old individuals – to a variety of problems in monetary economics. The particular class of overlapping-generations models in question is one in which, while there is an analytical entity termed 'fiat money', the specification deliberately excludes any shopping-time or related feature that would represent the transaction-facilitating aspect of money. Thus this approach, promoted most prominently in the work of Wallace (1980), tries to surmount the difficulty of modelling the medium-of-exchange function of money by simply ignoring it, emphasizing instead the asset's function as a store of value.

Models developed under this overlapping-generations approach typically possess highly distinctive implications, of which the particularly striking examples will be mentioned. First, if the monetary authority causes the stock of money to grow at a rate in excess of the economy's rate of output growth, no money will be demanded and the price level will be infinite. Second, steady-state equilibria in which money is valued will be Pareto optimal if and only if the growth rate of the money stock is non-positive. Third, open-market changes in the money stock will have no effect on the price level. It has been shown, however, that these implications result from the models' neglect of the medium-of-exchange function of money. Specifically. McCallum (1983) demonstrates that all three implications vanish if this neglect is remedied by recognition of shopping-time considerations as above. That conclusion suggests that the class of overlapping-generations models under discussion provides a seriously misleading framework for the analysis of monetary issues. This weakness, it should be added, results not from the generational structure of these models, but from the overly restrictive application of the principle that assets are valued solely on the basis of the returns that they yield; in particular, the models fail to reflect the non-pecuniary return provided by holdings of the medium of exchange. On these points see also Tobin (1980).

Recognizing this problem but desiring to avoid specifications like (2), some researchers have been attracted to the use of models incorporating a *cash-in-advance* constraint (e.g. Lucas 1980; Svensson 1985). In these models, it is assumed that an individual's purchases in any period cannot exceed the quantity of money brought into that period. Clearly, imposition of this type of constraint gives a medium-of-exchange role to the model's monetary asset and thereby avoids the problems of the Wallace-style overlapping-

generations models. Whether it does so in a satisfactory manner is, however, more doubtful. In particular, the cash-in-advance formulation implies that start-of-period money holdings place a *strict* upper limit on purchase during the period. This is a considerably more stringent notion than that implied by (2), which is that such purchases are possible but increasingly expensive in terms of time and/or other resources. Thus the demand for money will tend to be less sensitive to interest-rate changes with the cash-in-advance specification than with one that ties consumption and money holding together less rigidly. More generally, the cash-in-advance specification can be viewed as an extreme special case of the shopping-time function described in (2), in much the same way as a fixed-coefficient production function is a special case of a more general neoclassical technology. For some issues, use of the special case specification will be convenient and not misleading, but care must be exerted to avoid inappropriate applications. It seems entirely unwarranted, moreover, to opt for the cash-in-advance specification in the hope that it will be more nearly structural and less open to the Lucas critique (1976) than relations such as (2). Both of these specificational devices – and probably any that will be analytically tractable in a macroeconomic context – should be viewed not as literal depictions of technological or social constraints, but as potentially useful metaphors that permit the analyst to recognize in a rough way the benefits of monetary exchange. (On the general topic, see Fischer 1974).

A final controversy that deserves brief mention pertains to an aspect of money demand theory that has not been formally discussed above, but which is of considerable importance in practical applications. Typically, econometric estimates of money-demand functions combine 'long run' specifications such as (12) with a *partial adjustment* process that relates actual money-holdings to the implied 'long run' values. Operationally, this approach often results in a regression equation that includes a lagged value of the money stock as an explanatory variable. (Distributed-lag formulations are analytically similar.) Adoption of the partial adjustments mechanism is justified by appeal to portfolio-adjustment costs. Specifically, some authors argue that money balances serve as a 'buffer stock' that temporarily accommodates unexpected variations in income, while others attribute sluggish adjustments to search costs.

From the theoretical perspective, however, the foregoing interpretation for the role of lagged-money balances (or distributed lags) appears weak. If is difficult to believe that tangible adjustment costs are significant, and in their absence there is no role for lagged money balances, in formulations as such as (12) when appropriate transaction and opportunity-cost variables are included. Furthermore, typical estimates suggest adjustment speeds that are too slow to be plausible.

These points have been stressed by Goodfriend (1985), who offers an alternative explanation for the relevant empirical findings. A model in which there is full contemporaneous adjustment of money-holding to transaction and opportunity-cost variables is shown to imply a positive coefficient on lagged money when these determinants are positively autocorrelated and contaminated with measurement error. Under this interpretation, the lagged variable is devoid of behavioural significance; it enters the regression only because it helps to explain the dependent variable in a mongrel equation that mixes together relations pertaining to money-demand and other aspects of behaviour. (This particular conclusion is shared with the 'buffer stock' approach described by Laidler (1984), which interprets the conventional regression as a confounding of money-demand with sluggish price-adjustment behaviour.) Furthermore, the measurement error hypothesis can account for positive auto-correlation of residuals in the conventional regression and, if measurement errors are serially correlated, the *magnitude* of the lagged-money coefficient typically found in practice.

See Also

- ▶ Liquidity Preference
- ▶ Quantity Theory of Money
- ▶ Velocity of Circulation

Bibliography

Barro, R.J., and S. Fischer. 1976. Recent developments in monetary theory. *Journal of Monetary Economics* 2(2): 133–167.

Baumol, W.J. 1952. The transactions demand for cash: An inventory theoretic approach. *Quarterly Journal of Economics* 66: 545–556.

Brunner, K., and A. Meltzer. 1971. The uses of money: Money in the theory of an exchange economy. *American Economic Review* 61(5): 784–805.

Cannan, E. 1921. The application of the theoretical apparatus of supply and demand to units of currency. *Economic Journal* 31: 453–461.

Dutton, D.S., and W.P. Gramm. 1973. Transactions costs, the wage rate, and the demand for money. *American Economic Review* 63(4): 652–665.

Fama, E.F. 1980. Banking in the theory of finance. *Journal of Monetary Economics* 6(1): 39–57.

Fama, E.F., and A. Farber. 1979. Money, bonds, and foreign exchange. *American Economic Review* 69(4): 639–649.

Feige, E., and M. Parkin. 1971. The optimal quantity of money, bonds, commodity inventories, and capital. *American Economic Review* 61(3): 335–349.

Fischer, S. 1974. Money and the production function. *Economic Inquiry* 12(4): 517–533.

Fisher, I. 1911. *The purchasing power of money.* New York: Macmillan.

Fisher, I. 1930. *The theory of interest*. New York: Macmillan.

Friedman, M. 1956. The quantity theory of money: A restatement. In *Studies in the quantity theory of money*, ed. M. Friedman. Chicago: University of Chicago Press.

Friedman, M., and A.J. Schwartz. 1970. *Monetary statistics of the United States*. New York: Columbia Press for the National Bureau of Economic Research.

Goodfriend, M. 1985. Reinterpreting money demand regressions. *Carnegie-Rochester Conference Series on Public Policy* 22(Spring): 207–242.

Hicks, J.R. 1935. A suggestion for simplifying the theory of money. *Economica* 2: 1–19.

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

Karni, E. 1973. The transactions demand for cash: Incorporation of the value of time into the inventory approach. *Journal of Political Economy* 81(5): 1216–1225.

Keynes, J.M. 1930. *A treatise on money.* 2 vols, London: Macmillan.

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

King, R.G., and C.I. Plosser. 1986. Money as the mechanism of exchange. *Journal of Monetary Economics* 17(1): 93–115.

Lavington, F. 1921. *The English capital market*. London: Methuen.

Lucas Jr., R.E. 1976. Econometric policy evaluation: A critique. *Carnegie-Rochester Conference Series on Public Policy* 5(Autumn): 19–46.

Lucas Jr., R.E. 1980. Equilibrium in a pure currency economy. In *Models of monetary economies*, ed. J.H. Kareken and N. Wallace. Minneapolis: Federal Reserve Bank of Minneapolis.

Marshall, A. 1926. In *Official papers by Alfred Marshall*, ed. J.M. Keynes. London: Macmillan.

McCallum, B.T. 1983. The role of overlapping-generations models in monetary economics. *Carnegie-Rochester Conference Series on Public Policy* 18 (Spring): 9–44.

McCallum, B.T. 1985. Bank deregulation, accounting systems of exchange, and the unit of account: A critical review. *Carnegie-Rochester Conference Series on Public Policy* 23(Autumn): 13–45.

Meltzer, A.H. 1963. The demand for money: The evidence from the time series. *Journal of Political Economy* 71: 219–246.

Mill, J.S. 1848. *Principles of political economy.* 2 vols, London: John W. Parker.

Miller, M.H., and D. Orr. 1966. A model of the demand for money by firms. *Quarterly Journal of Economics* 80: 413–435.

Niehans, J. 1978. *The theory of money.* Baltimore: Johns Hopkins University Press.

Patinkin, D. 1956. *Money, interest, and prices*. New York: Harper and Row.

Pigou, A.C. 1917. The value of money. *Quarterly Journal of Economics* 32: 38–65.

Saving, T.R. 1971. Transactions costs and the demand for money. *American Economic Review* 61(3): 407–420.

Sidrauski, M. 1967. Rational choice and patterns of growth in a monetary economy. *American Economic Association Papers and Proceedings* 57: 534–544.

Svensson, L.E.O. 1985. Money and asset prices in a cash-in-advance economy. *Journal of Political Economy* 93(5): 919–944.

Tobin, J. 1956. The interest-elasticity of transactions demand for cash. *Review of Economics and Statistics* 38: 241–247.

Tobin, J. 1958. Liquidity preference as behavior toward risk. *Review of Economic Studies* 25: 65–86.

Tobin, J. 1980. Discussion. In *Models of monetary economics*, ed. J.H. Kareken and N. Wallace. Minneapolis: Federal Reserve Bank of Minneapolis.

Wallace, N. 1980. The overlapping generations model of fiat money. In *Models of monetary economies*, ed. J.H. Kareken and N. Wallace. Minneapolis: Federal Reserve Bank of Minneapolis.

Whitaker, J.K. (ed.). 1975. *The early economic writings of Alfred Marshall, 1867–1890.* 2 vols. New York: Free Press.

Wicksell, K. 1906. *Lectures on political economy.* Trans. E. Classen, London: Routledge & Kegan Paul, 1935, Vol. II.

Demand Management

G. D. N. Worswick

The expression 'demand management' came into general use after World War II. The idea has its roots in the *General Theory of Employment, Interest and Money* (1936), in which Keynes had argued that in capitalist economies the aggregate demand for goods and services could fall short of the capacity of the economy to produce, with resultant unemployment. A deficiency of demand could be made good by governments increasing their expenditure or lowering taxes, or by the monetary authorities lowering interest rates to stimulate investment. Contrariwise, if there was excess demand, fiscal and monetary policy could be used in a restrictive manner. In the United States, 'demand management' never wholly displaced the term 'stabilization policy'.

Demand management presupposes that the working of the economy is sufficiently well understood for reasonable assessments to be made of the likely evolution of such variables as private and public consumption and investment, exports and imports and the level of prices. If the assessment indicated an 'inflationary gap' between total expenditure and available supplies, that would call for higher taxes or lower public expenditure. In the case of demand deficiency, the task was to estimate the additional demand needed to bring the economy to capacity output at full employment. Demand management was not intended to influence capacity output itself, which would be determined by such longer-term factors as the growth of the labour force, technical progress and the stock of capital equipment: it was aimed at correcting shorter-term deviations of output from its sustainable trend. The feasibility of policy intervention depends on the successive periods of time elapsing between the recognition of the need for action, taking the decision to act, the making of the policy change and the effect on the targeted variable. The effectiveness of demand management depends on the accuracy of forecasts: it also depends on the political constitution, which determines the speed and frequency with which policy interventions, such as tax changes, can be made. In Britain the annual Budget was initially the main pillar of demand management, tax changes rather than alterations in expenditure being the main instruments; later on, adjustments between Budgets became more frequent. When restraint was called for, fiscal policy was supplemented by the control of consumer credit and by regulating the investment programme of the public sector. Monetary policy was directed towards the balance of payments. During the 1950s and 1960s most European countries followed policies of active demand-management, though there were differences in the mix of fiscal and monetary policy; the United States did not embark on fiscal expansion until the 1960s.

In the quarter-century following World War II, the average annual growth of output was historically high, the rate of unemployment very low and the amplitude of fluctuations about the trend of output also historically low. Some would argue that the contribution of demand management to high activity and reduced instability can only be tested with the aid of macroeconometric models, and these, with the appropriate data series, are not available for most of the period in question. But a number of more limited studies have been made. Commenting on these, three members of the OECD Secretariat (Llewellyn et al. 1985) concluded that, in general, policy was stabilizing. They also endorsed the observation of Matthews (1968) that because economic agents and entrepreneurs believed that the authorities would use policy to control activity, the private sector invested on a scale and with a smoothness which contributed to the stability of the whole economy. Throughout the 1950s and 1960s there were in nearly all advanced countries regular annual rises in nominal wages and prices, but at rates which would be considered moderate by subsequent standards, price increases averaging less than 2 per cent a year in some countries and not more than 5 per cent a year in any. These increases were seen not so much as symptoms of excess demand

as evidence of wage–price and wage–wage spirals, and a number of countries operated formal or informal 'incomes policies' designed to contain the rise in the general wage level. Despite occasional misgivings, the wage inflation was reckoned acceptable, and monetary policy was accommodatory.

The postwar 'golden age' came to an end in the early 1970s. After 1973 there was a sharp slowdown in productivity growth in all advanced countries, accompanied in most cases by rising unemployment and rising inflation. The 'stagflation' of the 1970s was followed by world recession in the early 1980s, when output in the OECD countries as a whole almost stopped rising altogether and unemployment reached levels not seen since the 1930s. On the face of it, this was a time when the thrust of demand management might have been expected to be expansionary. In fact, with some notable exceptions, both fiscal and monetary policy became increasingly restrictive. How can this paradox be explained?

Two factors may be singled out as contributing to the doubling of the average rate of inflation in OECD countries in the 1970s: a change into a higher gear of wage inflation in a number of countries at the end of the 1960s, and the fourfold increase in the price of oil of OPEC 1 in 1973–4. The latter was of particular significance since it boosted cost inflation while at the same time acting to reduce demand in oil-importing countries. While the typical inflation of the 1960s might have been acceptable, the higher rate was not, and notwithstanding the recession of output and employment in the mid-1970s, governments began to direct demand management towards reducing inflation. An element of the expansionist mode of demand management can still be seen in the outcome of the Bonn Summit meeting of the major powers in 1978, whereby those countries with low inflation and balance of payments surpluses, notably Germany and Japan, were to engineer a modest domestic expansion and act as 'locomotives' to pull up the rest of the world, but this initiative was swiftly overtaken by OPEC 2, re-igniting inflation and renewing the determination of major countries to pursue restrictive policies. This intensified the recession until the United States broke ranks in 1982.

So long as the response to cost inflation is restrictive fiscal and monetary policy, then clearly demand management is not available to combat unemployment. However, there has also been a theoretical reappraisal of the potential role of demand management. The theme of the 'monetarist' reaction launched by Friedman in 1968 was that there was no lasting trade-off between unemployment and inflation. In his view, if unemployment was pushed below a certain 'natural' rate, determined by the characteristics of the real economy, there would be not merely higher, but accelerating, inflation. In addition, the numerous and uncertain time-lags between the diagnosis of the need for action and the effects on the economy of the appropriate policy change rendered discretionary demand-management hazardous. Accordingly, Friedman recommended the adoption of a simple rule governing the growth of the money supply. Monetarism was influential in leading some countries to adopt monetary targets; in particular, the monetary aggregate 'Sterling M3' was the focal point of the Medium Term Financial Strategy adopted by the British government in 1980. Inflation did come down, though it brought monetarism down with it, for the chosen measure of money supply rose quite out of line with the price level, and the target was officially abandoned five years later. A more radical critique of demand management emerged from the 'rational expectations' analysis being developed at the end of the 1970s. This analysis raises important questions in economic theory and in econometric modelling, but it is the marriage of the rational expectations hypothesis with the 'natural rate' in the 'New Classical Economics' which has the most serious consequences for demand management, for it leads to the denial of any possible influence on the real economy of any systematic policies of this kind. This, however, is a result for a theoretical economy in which all markets clear all the time. It does not apply to a 'disequilibrium' economy in which important markets, such as the labour market, do not clear.

Though by no means conclusive, these critiques have raised important questions about the

scope and limitations of demand management. Economists in the Keynesian tradition argue that its potential to influence real output and employment still remains; in their view what happened in the 1970s and 1980s was a change of the objective of demand management towards reducing inflation. If it is to be restored to its original role, an alternative means of restraining inflation is required, and that means a viable incomes policy. Others advocate switching the target of demand management from real output to the stabilization of the growth of the nominal value of GDP. By itself, this is not unlike what monetarism hoped to achieve, but New Keynesians, such as Meade, go a step further and also invoke incomes policy. But their idea is that wage settlements should be targeted on maximizing employment, whereas conventional incomes policy is conceived as restraining nominal incomes.

The risk of renewed inflation if demand management was given back its old role of reducing unemployment appears to be less in some countries than others, but in a number of countries this restoration is unlikely unless a viable incomes policy can be put in place. An equally powerful inhibition is the fear of exchange-rate depreciation. Those who advocated floating exchange rates had argued that the abandonment of fixed rates would allow countries to follow their own domestic policies of demand management without balance of payments crises cutting them short. But it seems to have been a case of jumping out of the frying pan into the fire. Countries attempting to expand out of recession unilaterally have quickly experienced a sharp fall in their exchange rate: only the dominant United States economy has so far been able to combine internal expansion with stability of the external value of its currency. Joint action among other countries seems to be required to counter the destabilizing effects of excessive currency fluctuations. Some tentative moves in this direction began to appear in the mid-1980s. In retrospect, the Bretton Woods arrangements are seen to have constituted a high peak of international economic cooperation, and some advocate the creation of a 'new Bretton Woods' on a world-wide basis. Others pin their faith on the strengthening of regional groupings, such as the European Community. In any case, the prospects for demand management resuming its role as an instrument for expansion and high employment are bound up with the extent to which major countries can contrive to coordinate their separate national policies.

See Also

► Deficit spending

Bibliography

Friedman, M. 1968. The role of monetary policy. *American Economic Review* 58: 1–17.

Keynes, J.M. 1936/1973. The general theory of employment, interest and money. In *Collected writings of John Maynard Keynes*, vol. VII. London: Macmillan.

Llewellyn, J., S. Potter, and L. Samuelson. 1985. *Economic forecasting and policy*. London: Routledge & Kegan Paul.

Matthews, R.C.O. 1968. Why has Britain had full employment since the war? *Economic Journal* 78: 555–569.

Vines, D., J. Maciejowski, and J.E. Meade. 1983. *Demand management*. London: Allen & Unwin.

Demand Price

John K. Whitaker

JEL Classifications
D1

Earlier economic literature doubtless contains casual usages of the phrase 'demand price', but its appropriation as a technical term appears to date from Alfred Marshall's *Principles of Economics* (Marshall, 1890: see Marshall, 1920, pp. 95–101). Marshall applied the term in the contexts of both individual and market demand. Starting with a commodity (tea) purchasable in integral units of a pound's weight, an individual's demand price for the xth pound is the price he is just willing to pay for it given that he has already acquired $x - 1$ pounds. The basic assumption is

that this demand price is lower the larger is x. A schedule of demand prices for all possible quantities (values of x) defines the consumer's demand schedule. Its graph is naturally drawn with quantity on the horizontal axis. In the case of a perfectly divisible commodity, the demand price of quantity x must be redefined as the price *per unit* which the consumer would be willing to pay for a tiny increment, given that he already possesses amount x. The demand schedule then graphs as a continuous negatively sloped demand curve showing demand price in this sense as a function of x.

If the individual is free to buy any quantity at a fixed price, his 'marginal demand price' is the demand price for that quantity 'which lies at the margin or terminus or end of his purchases' (Marshall, 1920, p. 95). For a perfectly divisible commodity, marginal demand price must equal market price. For a commodity purchasable in integral units only, market price may lie anywhere below marginal demand price, but not so low as to make the next unit marginal.

Marshall's discussion of consumer behaviour is based on two general assumptions, although these are informally relaxed at various points. The first is that the utility obtained from consuming a commodity depends only on the amount of that commodity. The second is that the marginal utility of 'money', or expenditure on all other goods, remains approximately constant with respect to variation in the expenditure on any particular commodity – the presumption being that the latter expenditure is only a small fraction of total expenditure. These assumptions have convenient consequences for the concept of demand price. If $u(x)$ denotes the utility a consumer obtains from consuming quantity x of a given good in a specified period, while λ is the constant marginal utility of money to him, then demand price for quantity x is $(\mathrm{d}u/\mathrm{d}x)/\lambda$ in the case of divisible quantity and $[u(x) - u(x-1)]/\lambda$ in the case when only integral quantities are feasible. In either case, given the value of λ, demand price depends on x alone and is proportional to marginal utility. The hypothesis of diminishing demand price is tantamount to that of diminishing marginal utility. A further advantage is that the demand price for quantity x is independent of the pecuniary terms on which the earlier units were, or are to be, acquired, as these terms will not change the marginal utility of money.

Although demand price is, on the above assumptions, proportional to marginal utility it has the great advantage of being measured in operational money units. This permits a monetary measure of the net benefit or consumer surplus obtained from the option of buying the commodity in question on specified monetary terms, rather than having to divert the expenditure to other goods. The distinction between demand price and market price is an operational version of the classical distinction between value in use and value in exchange.

The concept of demand price features prominently in Marshall's analysis of the market for a single commodity sold at a fixed price which is uniform to all buyers. Demand price is now interpreted as the maximum *uniform* price at which any specified aggregate quantity of the commodity can be sold on the market during a given period. The negatively sloped market demand curve is simply a lateral addition of the individual demand curves and expresses the common demand price as a function of the aggregated quantity. Marshall recognized (1920, p. 457n) that it would be more natural when dealing with market demand to view quantity as a function of price, as Cournot (1838, pp. 44–55) had done, but chose the converse approach to maintain symmetry with his treatment of supply. Believing in the importance of scale economies in production, he deemed it generally impossible to treat quantity supplied per unit of time as a single-valued function of market price. Instead, adopting what he took to be the businessman's perspective, he introduced the concept of 'supply price'; the minimum uniform price at which any given quantity will be supplied to the market.

Market equilibrium occurs at any quantity whose demand price and supply price are equal, so that the market demand curve intersects the market supply curve – the latter the graph of supply price as a function of aggregate quantity

supplied, a lateral sum of individual supply curves. Equilibrium is locally stable if the demand curve cuts the supply curve from above at the equilibrium quantity. This result is justified by the argument that the rate of supply will increase if the current market price (always determined by demand price) exceeds supply price at the current quantity, so that additional production offers excess profit, decreasing in the opposite case (Marshall, 1920, pp. 345–7). The resulting dynamic process is usually referred to as the Marshallian adjustment process.

It is probably due to Marshall's influence that English-speaking economists still graph demand and supply curves with quantity on the horizontal axis even though adopting a more Walrasian perspective which treats quantities demanded and supplied as functions of market price.

Marshall's conception of the demand price of a lone commodity, segregated from other commodities by an assumed constancy of the marginal utility of money, does not feature prominently in modern theoretical work. Instead, a multi-commodity formulation of utility and demand is typically adopted. Consider a consumer maximizing the utility function $u(x_1, x_2, \ldots, x_n)$ subject to the budget $p_i x_i$ constraint $\sum_{p_i x_i} = M$. (Here the x_i are quantities and the p_i prices of the n commodities and M is a preset total expenditure level. The utility function, u, is assumed strictly increasing, strictly quasi-concave, and differentiable.) This maximization implies the consumer's direct demand functions $x_i = d_i(p_1/M, p_2/M, \ldots, p_n/M)$, $i = 1,2, \ldots, n$, sometimes (but with dubious justification) referred to as Marshallian demand functions to distinguish them from Hicksian compensated or constant-utility demand functions.

These demand functions can usually be inverted to yield the indirect or inverse demand functions $p_i/M = g_i(x_1, x_2, \ldots, x_n)$, $i = 1, 2, \ldots, n$. However, these can be obtained more immediately from the budget constraint and the first-order conditions $\partial u/\partial x_i = \lambda p_i$, $i = 1, 2, \ldots, n$ (where λ is the Lagrange multiplier associated with the budget constraint). We have, for $i = 1, 2, \ldots, n$,

$$\frac{p_i}{M} = \frac{\partial u/\partial x_i}{\lambda M} = \frac{\partial u/\partial x_i}{\sum \lambda p_j x_j} = \frac{\partial u/\partial x_i}{\sum x_j (\partial u/\partial x_j)} \equiv g_i(x_1, x_2 \ldots, x_n) \quad (1)$$

(The g_i are clearly unaffected by a monotone increasing transformation of u and reduce to $(\partial u/\partial x_i)/u$ if u is homogeneous of degree one.) The indirect demand functions (1) are the natural generalization of Marshall's demand-price concept at the individual level, defining an n-vector of normalized prices at which a given n-vector of commodities will be demanded.

Indirect demand functions may be useful in the contexts of central planning or rationing, where they can indicate the prices planners should choose to clear markets given the quantities available, or the notional prices at which ration allotments would just be freely purchases (see Pearce, 1964, pp. 57–64). But unfortunately, although indirect demand functions are readily obtained for the individual, they are not as easily aggregated to the market level as are direct demand functions. The asymmetry arises from the fact that individuals face identical prices but do not make identical quantity choices. Thus, market-level indirect demand functions must generally be obtained by first aggregating the individual direct demand functions and then inverting the resulting market functions.

The modern duality approach to consumer behaviour has revealed fundamental symmetries in the roles of prices and quantities. The alternatives of viewing quantity demanded as a function of price or demand price as a function of quantity can now be seen as only one of a variety of dual alternatives which considerably enrich theoretical and econometric analysis. (See Gorman, 1976, for a simple treatment.)

See Also

▶ Marshall, Alfred (1842–1924)

Bibliography

Cournot, A.A. 1838. *Mathematical principles of the theory of wealth*. New York: Macmillan, 1897.
Gorman, W.M. 1976. Tricks with utility functions. In *Essays in economic analysis*, ed. J.J. Artis and A.R. Nobay. London: Cambridge University Press.
Marshall, A. 1890. *Principles of economics*, vol. I. Macmillan: London.
Marshall, A. 1920. *Principles of economics: An introductory volume*. London: Macmillan. Eighth edition of Marshall (1890).
Pearce, I.F. 1964. *A contribution to demand analysis*. Oxford: Clarendon Press.

Demand Theory

Volker Böhm and Hans Haller

Abstract

Demand theory describes and explains individual choice of consumption bundles. Traditional theory considers optimizing behaviour when the consumer's choice is restricted to consumption bundles that satisfy a budget constraint. The budget constraint is determined by price–income pairs. A demand correspondence assigns to each price–income pair a non-empty set of optimal consumption bundles. A demand function assigns to each price–income pair a unique optimal consumption bundle. Optimality of consumption bundles is based on a preference relation. The theory derives existence and properties of demand correspondences (demand functions) from assumptions on preference relations and, if applicable, their utility representations.

Keywords

Budget sets; Cardinal utility; Completeness; Consumption plans; Consumption sets; Contingent commodities; Continuity; Continuous preference orders; Convexity; Demand correspondences; Demand functions; Demand sets; Demand theory; Expenditure functions; Giffen goods; Hicksian (income-compensated) demand function; Inferior goods; Integrability of demand; Inverse demand function; Lebesgue measure approach; Normal goods; Ordinal utility; Preference maximization; Preference orders; Quasi-concavity; Reflexivity; Representability of preferences; Revealed preference theory; Separability; Slutsky matrix; Slutsky, E.; Transitivity; Utility maximization; Walras, L.

JEL Classifications

D11

The main purpose of demand theory is to describe and explain observed consumer choices of commodity bundles. Market parameters, typically prices and income, determine constraints on commodity bundles. Given a combination of market parameters, a commodity bundle or a non-empty set of commodity bundles, which satisfies the corresponding constraints, is called a demand vector or a demand set. The mapping which assigns to every admissible combination of market parameters a unique demand vector (or a non-empty demand set) is called a demand function (or a demand correspondence, respectively). Traditional demand theory considers the demand function (or correspondence) as the outcome of some optimizing behaviour of the consumer. Its primary goal is to determine how alternative assumptions on the constraints, objectives and behavioural rules of the consumer affect his observed demands for commodities. The traditional model of the consumer postulates preferences over alternative commodity bundles to describe the objectives of the consumer. Its behavioural rule consists in maximizing these preferences on the set of feasible commodity bundles which satisfy the budget constraint imposed by the market parameters. If there is a unique preference maximizer under each budget constraint, then preference maximization determines a demand function. If there is at least one preference maximizer under each budget constraint, then preference maximization determines a demand correspondence.

Once the traditional view is adopted, the occurrence of demand correspondences cannot be

avoided. Compatibility of observed demand, which is always unique, with some demand correspondence poses a minor problem in general. However, the correspondence should be obtained through preference maximization. The last requirement leads to the main issues of modern demand theory: Which demand correspondences are compatible with preference maximization? Given any conditions necessary for demand correspondences to be compatible with preference maximization, are they sufficient? Which demand correspondences are compatible with a special class of preferences? What type of preferences yields a particular class of demand correspondences? When addressing these issues, modern demand theory attempts to link two concepts: preferences and demand.

Historically, the important concept was utility rather than preference. Before Fisher (1892) and Pareto (1896), utility was conceived as cardinal: that is, it was assumed to be a measurable scale for the degree of satisfaction of the consumer. Fisher and Pareto were the first to observe that an arbitrary increasing transformation of the utility function has no effect on demand. Edgeworth (1881) had already written utility as a general function of quantities of all commodities and had employed indifference curves. It is now widely accepted in demand theory that only ordinal utility matters. That is, a utility function serves merely as a convenient device to represent a preference relation, and any increasing transformation of the utility function will serve this purpose as well.

Representability by utility functions imposes some restrictions on preferences. The problem of representability of a preference relation by a numerical function was solved by Debreu (1954, 1959, 1964) based on work by Eilenberg (1941), and by Rader (1963) and Bowen (1968). While still assuming cardinal utility, Walras (1874) developed the first 'theory of demand'. His demand was a function of all prices and the endowment bundle, obtained through utility maximization. Slutsky (1915) finally assumed an ordinal utility function with enough restrictions to yield a maximum under any budget constraint and testable properties of the resulting demand functions. In particular, he obtained negativity of diagonal elements and symmetry of the 'Slutsky matrix'.

Antonelli (1886) was the first to go the opposite way: construct indifference curves and a utility function from the so-called inverse demand function. Pareto (1906) took the same route. Katzner (1970) reports on recent results in this direction. The construction of preference relations from demand functions was achieved in two ways:

1. Samuelson (1947) and Houthakker (1950) introduced the concept of revealed preference into demand theory. Considerable progress in relating utility and demand in terms of revealed preference was achieved by Uzawa (1960), further refinements being due to Richter (1966).
2. Hurwicz and Uzawa (1971) contributed to the following so-called integrability problem: construct a twice continuously differentiable utility representation from a continuously differentiable demand function which satisfies certain integrability conditions (including symmetry and negative semi-definiteness of the Slutsky matrix).

Kihlstrom et al. (1976) unified the two approaches by relating the axioms of revealed preference to properties of the Slutsky matrix.

Since there exists a sizable literature on demand theory, many of the concepts and results are well established and well-known. These have become so much part of standard knowledge in economic theory that they are included in any contemporary microeconomic textbook and other surveys. It would substantially reduce the space available for a presentation of the new results of recent decades if an extended introductory account of demand theory were to be included here as well.

Commodities and Prices

Consumers purchase or sell commodities, which can be divided into goods and services. Each commodity is specified by its physical quality, its location, and the date of its availability. In the case of uncertainty, the state of nature in which the

commodity is available may be added to the specification of a commodity. This leads to the notion of a contingent commodity (see Arrow 1953; Debreu 1959). We assume as in traditional theory that there exists a finite number l of such commodities. Quantities of each commodity are measured in real numbers. A *commodity bundle* is an l-dimensional vector $x = (x_1,\ldots,x_l)$. The set of all l-dimensional vectors $x = (x_1,\ldots,x_l)$ is the l-dimensional Euclidean space $\mathbb{R}^l$ which we interpret as the *commodity space*. $|x_h|$ indicates the quantity of commodity $h = 1,\ldots,l$. Commodities are assumed to be perfectly divisible, so that their quantity may be expressed as any (non-negative) real number. The standard sign convention for consumers assigns positive numbers for commodities made available to the consumer (inputs) and negative numbers for commodities made available by the consumer (outputs). Hence, a priori any commodity bundle $x \in \mathbb{R}^l$ is conceivable.

The price p_h of a commodity h, $h = 1,\ldots,l$, is a real number which is the amount in units of account that has to be paid in exchange for one unit of the commodity. For the consumer, p_h is given and has to be paid now for the delivery of commodity h under the circumstances (location, date, state) specified for commodity h. A *price system* or *price vector* is a vector $p = (p_1,\ldots,p_l)$ in $\mathbb{R}^l$ and contains the prices for all commodities. The value of a commodity bundle x given the price vector p is $px = \sum_{h=1}^{l} p_h x_h$. This means that commodity bundles are *priced linearly*.

Consumption Sets and Budget Sets

Typically, some commodity bundles cannot be consumed by a consumer for physical reasons. Those consumption bundles which can be consumed form the consumer's *consumption set*. This is a non-empty subset X of the commodity space $\mathbb{R}^l$. A consumer must choose a bundle x from his consumption set X in order to subsist. Traditionally, inputs in consumption are described by positive quantities and outputs by negative quantities. So in particular, the labour components of a consumption bundle x are all non-positive, unless labour is hired for a service. One usually assumes that the consumption set X is closed, convex, and bounded below. Vectors $x \in X$ are sometimes called *consumption plans*.

Given the sign convention on inputs and outputs and a price vector p, the value px of a consumption plan x defines the net outlay of x, that is the value of all purchases (inputs) minus the value of all sales (outputs) for the bundle x. Trading the bundle x in a market at prices p implies payments and receipts for that bundle. Therefore, the value of the consumption plan should not exceed the initial wealth (or income) of the consumer which is a given real number w. If the consumer owns a vector of initial resources ω and the price vector p is given, then w may be determined by $w = p\omega$. The consumer may have other sources of wealth: savings and pensions, bequests, profit shares, taxes, or other liabilities. Given p and w, the set of possible consumption bundles whose value does not exceed the initial wealth of the consumer is called the *budget set* and is defined formally by

$$\beta(p,w) = \{x \in X \mid px \leq w\}.$$

The ultimate decision of a consumer is to choose a consumption plan from his budget set. Those vectors in $\beta(p,w)$ which the consumer eventually chooses form his *demand set* $\phi(p,w)$.

Preferences and Demand

The choice of the consumer depends on his tastes and desires. These are represented by his *preference relation* $\succsim$ which is a binary relation on X. For any two bundles $x, y \in X$, $x \succsim y$ means that x is at least as good as y. If the consumer always chooses a most preferred bundle in his budget set, then his demand set is defined by

$$\phi(p,w) = \left\{x \in \beta(p,w) \mid x' \in \beta(p,w) \text{ implies } x \succsim x' \text{ or not} x' \succsim x\right\}.$$

Three basic axioms are usually imposed on the preference relation $\succsim$ which are taken as a definition of a rational consumer:

Axiom 1 (reflexivity). If $x \in X$, then $x \succsim x$, that is, any bundle is as good as itself.

Axiom 2 (transitivity). If $x, y, z \in X$ such that $x \succsim y$ and $y \succsim z$, then $x \succsim z$.

Axiom 3 (completeness). If $x, y \in X$, then $x \succsim y$ or $y \succsim x$.

A preference relation $\succsim$ which satisfies these three axioms is a complete preordering or weak order on X and will be called a *preference order*. Already Axioms 2 and 3 define a preference order, since Axiom 3 implies Axiom 1. A preference relation $\succsim$ on X induces two other relations on X, the relation of strict preference, $\succ$, and the relation of indifference, $\sim$.

Definition Let $\succsim$ be a preference relation on the consumption set X. A bundle x is said to be *strictly preferred to* a bundle y, that is $x \succ y$, if and only if $x \succsim y$ and not $y \succsim x$. A bundle x is said to be *indifferent to* a bundle y, that is $x \sim y$, if and only if $x \succsim y$ and $y \succsim x$.

Lemma *Suppose $\succsim$ is reflexive and transitive. Then*

(i) *$\succ$ is irreflexive, that is, not $x \succ x$, and transitive*;

(ii) *$\sim$ is an equivalence relation on X, which means that $\sim$ is reflexive, transitive, and symmetric: that is, $x \sim y$ if and only if $y \sim x$.*

For $Z \subseteq X$, $x \in Z$, x is called *maximal in Z*, if for all $z \in Z$: not $z \succ x$. x is called a *best element of Z* or *most preferred in Z*, if for all $z \in Z : x \succsim z$. Best elements are maximal; maximal elements are not necessarily best elements. If $\succsim$ is complete, then best and maximal elements coincide. Obviously for any price vector p and initial wealth w,

$$\phi(p,w) = \{x \in \beta(p,w) \mid x \text{ is maximal in } \beta(p,w)\}.$$

Axioms 1–3 are not qsted in most of consumer theory. However, transitivity and completeness may be violated by observed behaviour. Recent developments in the theory of consumer demand indicate that some weaker axioms suffice to describe and derive consistent demand behaviour (see, for example, Sonnenschein 1971; Katzner 1971; Shafer 1974; Kihlstrom et al. 1976; Kim and Richter 1986). In an alternative approach, one could start from a strict preference relation as the primitive concept. This may sometimes be convenient. However, the weak relation $\succsim$ seems to be the more natural concept. If the consumer chooses x, although y was a possible choice as well, then his choice can only be interpreted in the sense of $x \succsim y$, but not as $x \succ y$.

For the remainder of this section, let us fix a preference order $\succsim$ on X and a non-empty subset B of $\mathbb{R}^{l+1}$ such that for every $(p,w) \in B$, there is a unique $\succsim$-best element in $\beta(p,w)$: that is, maximization of $\succsim$ defines a *demand function* $f{:}B \to X$ such that $\phi(p,w) = \{f(p,w)\}$ for all $(p,w) \in B$.

Let $x, x' \in X$, $x \neq x'$. We call x *revealed preferred to* x' and write xRx', if there is $(p,w) \in B$ such that $x = f(p,w)$ and $px' \leq px$. xRx' implies that both x and x' belong to the budget set $\beta(p,w)$ and x is chosen. Since f is derived from $\succsim$-maximization, xRx' implies $x \succ x'$. We call x *indirectly revealed preferred to* x' and write xR^*x', if there exists a finite sequence $x_0 = x, x_1, \ldots, x_n = x'$ in X such that $x_0Rx_1, \ldots, x_{n-1}Rx'$. Obviously, R^* is transitive. Since $\succ$ is transitive, xR^*x' implies $x \succ x'$. Consequently, the following must hold (otherwise $x \succ x$!):

$$\begin{gathered}(\text{SARP})\ xR^*x' \Rightarrow \text{not}\,(x'R^*x).\\ (\text{SARP})\ \text{implies}\\ (\text{WARP})\ xRx' \Rightarrow \text{not}\,(x'R^*x).\end{gathered}$$

(SARP) is the *strong axiom of revealed preference*; (WARP) is the *weak axiom*. Hence $\succsim$-maximization implies the strong axiom and a fortiori the weak axiom. For the inverse implication, see Chipman et al. (1971, chs. 1, 2, 3 and 5). For $l \geq 3$, there exist demand functions which satisfy (WARP) but not (SARP), whereas for $l = 2$, (WARP) and (SARP) are equivalent; see Section 3.J of Mas-Colell et al. (1995) and Kihlstrom et al. (1976, p. 977).

Continuous Preference Orders and Utility Functions

Axioms 1–3 have intuitive appeal. This is less so with the topological requirements of the following Axiom 4.

Axiom 4 (continuity). For every $x \in X$, the sets $\{y \in X | y \succsim x\}$ and $\{y \in X | x \succsim y\}$ are closed relative to X.

If $\succsim$ is a preference order, then Axiom 4 is equivalent to: For every $x \in X$, the sets $\{y \in X | y \succ x\}$ and $\{y \in X | x \succ y\}$ are open in X.

Closedness of $\{y \in X | y \succsim x\}$ requires that for any sequence y^n, $n \in \mathbb{N}$, in X such that y^n converges to $y \in X$ and $y^n \succsim x$ for all n, the limit y also satisfies $y \succsim x$. Openness of $\{y \in X | y \succ x\}$ means that if $y \succ x$, then $y' \succ x$ for any y' close enough to y.

The sets $\{y \in X | y \succsim x\}$ are called *upper contour sets* of the relation $\succsim$ and the sets $\{y \in X | x \succsim y\}$ are called *lower contour sets* of $\succsim$. For $x \in X$, the set $I(x) := \{y \in X | y \sim x\}$ is called the *indifference class* of x with respect to $\succsim$ or the $\succsim$-*indifference surface* through x or the $\succsim$-*indifference curve* through x. In the case $\succsim$ is reflexive and transitive, $I(x)$ is the equivalence class of x with respect to the equivalence relation $\sim$.

There is a preference order $\succsim$ on $\mathbb{R}^l$, $l \geq 2$, which does not satisfy Axiom 4, namely the *lexicographic order* defined by $(x_1,\ldots,x_l) \succsim (y_1,\ldots,y_l)$ if and only if $x = y$ or there exists $k \in \{1,\ldots,l\}$ such that: $x_j = y_j$ for $j < k$ and $x_k > y_k$. Few studies of the relationship between the order properties of Axioms 1–3 and the topological property of Axiom 4 have been made. We emphasize the following result.

Theorem (Schmeidler 1971). *Let $\succsim$ denote a transitive binary relation on a connected topological space X. Assume that there exists at least one pair $\overline{x}, \overline{y} \in X$ such that $\overline{x} \succ \overline{y}$. If for every $x \in X$, (i) $\{y \in X | y \succsim x\}$ and $\{y \in X | x \succsim y\}$ are closed and (ii) $\{y \in X | y \succ x\}$ and $\{y \in X | x \succ y\}$ are open, then $\succsim$ is complete.*

Definition Let X be a set and $\succsim$ be a preference relation on X. Then a function u from X into the real line $\mathbb{R}$ is a *(utility) representation* or a *utility function for* $\succsim$, if for all $x; y \in X : u(x) \geq u(y)$ if and only if $x \succsim y$. Clearly, if u is a utility representation for $\succsim$ and $f : \mathbb{R} \to \mathbb{R}$ is an increasing transformation, then the composition $f \to u$ is also a representation of $\succsim$. If $u : X \to \mathbb{R}$ is any function, then $\succsim$, defined by $x \succsim y$ if and only if $u(x) \geq u(y)$ for $x, y \in X$, is a preference order on X and u is a utility representation for $\succsim$.

Most utility functions used in consumer theory are continuous. If u is continuous and $\succsim$ is represented by u, then by necessity $\succsim$ is a continuous preference order. In our case where $X \subseteq \mathbb{R}^l$, the opposite implication also holds: If $\succsim$ is a continuous preference order, then it has a continuous utility representation.

Theorem (Debreu, Eilenberg, Rader) *Let X be a topological space with a countable base of open sets (or a connected, separable topological space) and $\succsim$ be a continuous preference order on X. Then $\succsim$ has a continuous utility representation.*

In our context of Euclidean commodity spaces, explicit constructions of continuous utility representations for continuous and monotonic preference orders are available. See Arrow and Hahn (1971) for the 'Euclidean distance approach' and Neuefeind (1972) for the 'Lebesgue measure approach'. For topological spaces X with a countable base of open sets, it has further been shown by Rader (1963) and Bosi and Mehta (2002) that an upper semi-continuous preference order on X has an upper semi-continuous utility representation.

As an immediate consequence of the representation theorem for preference relations, one obtains one of the standard results on the nonemptiness of the demand set $\phi(p,w)$, since any continuous function attains its maximum on a compact set (Weierstrass's th), though a direct proof is also possible.

Corollary *Let $X \subseteq \mathbb{R}^l$ be bounded below and closed, $\succsim$ be a continuous preference order on X,*

$p \in R^l_{++}$ *(that is,* $p \gg 0$*) and* $w \in \mathbb{R}$*. Then* $\beta(p, w) \neq \emptyset$ *implies* $\phi(p, w) \neq \emptyset$.

There has been a recent shift from proving existence to a more systematic study of the non-existence of utility representations. Needless to say that there are many preference orders on $\mathbb{R}^l$ or on subsets thereof with continuous utility representations. There are also total orders $\succsim$ (that is, preference orders $\succsim$ with $x \sim y \Leftrightarrow x = y$) on $\mathbb{R}^l$, $l \geq 2$, which admit utility representations, since there exist bijections $u : \mathbb{R}^l \to \mathbb{R}$. However, for $l \geq 2$, there is no total order on $\mathbb{R}^l, \mathbb{R}^l_+$ or $[0,1]^l$ which has a continuous utility representation; see Candeal and Induráin (1993). Moreover, a preference order $\succsim$ on X, which is not continuous, need not have a utility representation. For instance, the lexicographic order on $\mathbb{R}^l$, $l \geq 2$, a total order first discussed by Debreu (1954), does not have a utility representation, nor even a discontinuous one. Beardon et al. (2002) provide a classification of total orders which do not admit a utility representation. Estévez Toranzo and Hervés Beloso (1995) show that, if X is a non-separable metric space, then there exists a continuous preference order on X which cannot be represented by a utility function.

Some Properties of Preferences and Utility Functions

Some of the frequent assumptions on preference relations correspond almost by definition to analogous properties of utility functions, while other analogies need demonstration. We discuss the assumptions most commonly used.

Monotonicity A preference order $\succsim$ on $X \subseteq \mathbb{R}^l$ is *monotonic*, if $x, y \in X, x \geq y, x \neq y$ implies $x \succ y$.

This property means desirability of all commodities. If a monotonic preference order has a utility representation u, then u is an increasing function (in all arguments). Inversely, if $\succsim$ is represented by an increasing function, then $\succsim$ is monotonic.

Non-satiation Let $\succsim$ be the preference relation of a consumer over consumption bundles in X and let $x \in X$.

(i) x is a *satiation point for* $\succsim$ if $x \succsim y$ for all $y \in X$: that is, x is a best element in X.
(ii) The preference relation is *locally not satiated at* x, if for every neighbourhood U of x there exists $z \in U$ such that $z \succ x$.

Consider a utility representation u for $\succsim$. Then $x \in X$ is a satiation point if and only if u has a global maximum at x. $\succsim$ is locally not satiated at x if and only if u does not attain a local maximum at x. Local non-satiation rules out that u is constant in a neighbourhood of x. If $\succsim$ is locally not satiated at all x, then $\succsim$ cannot have thick indifference classes or satiation points.

Convexity A preference relation $\succsim$ on $X \subseteq \mathbb{R}^l$ is called

(i) *convex*, if the set $\{y \in X \,|\, y \succsim x\}$ is convex for all $x \in X$;
(ii) *strictly convex*, if X is convex and $\lambda x + (1 - \lambda)x' \succ x'$ for any two bundles x, $x' \in X$ such that $x \neq x'$; $x \succsim x'$ and for any λ such that $0 < \lambda < 1$;
(iii) *strongly convex*, if X is convex and $\lambda x + (1 - \lambda)x' \succ x''$ for any three bundles $x; x'; x'' \in X$ such that $x \neq x'$, $x \succsim x''$, $x' \succsim x''$ and for any λ such that $0 < \lambda < 1$.

Quasi-Concavity A function $u : X \to \mathbb{R}$ is called

(i) *quasi-concave*, if $u(\lambda x + (1 - \lambda)y) \geq \min \{u(x), u(y)\}$ for all $x, y \in X$ and any $\lambda \in [0,1]$;
(ii) *strictly quasi-concave*, if $u(\lambda x + (1 - \lambda)y) > \min \{u(x), u(y)\}$ for all $x, y \in X$ with $x \neq y$ and any $\lambda \in (0,1)$.

Let u be a representation of the preference order $\succsim$. Then u is (strictly) quasi-concave if and only if $\succsim$ is (strictly) convex. Quasi-concavity is preserved under increasing transformations: that

is, it is an ordinal property. In contrast, concavity is a cardinal property which can be lost under increasing transformations. With respect to the difficult problem to characterize those preference orders which have a concave representation, we refer to Kannai (1977). Clearly, if $\succsim$ is locally not satiated at all x, then $\succsim$ does not have a satiation point. In general, the inverse implication is false. If, however, $\succsim$ is strictly convex and does not have a satiation point, then $\succsim$ is locally not satiated at all x. Moreover, if h is strictly convex, then it has at most one satiation point. An immediate implication is the following lemma.

Lemma *Let $X \subseteq \mathbb{R}^l$ be bounded below, convex, and closed. Let $\succsim$ be a strictly convex, continuous preference order on $X, p \in R^l_{++}$, and $w \in \mathbb{R}$. Then $\beta(p, w) \neq \emptyset$ implies that $\phi(p,w)$ is a singleton.*

Separability Separable utility functions were used in classical consumer theory long before associated properties of preferences had been defined. All early contributions to utility theory assumed without much discussion an additive form of the utility function over different commodities. It was not until Edgeworth (1881) that utility was written as a general function of a vector of commodities. The particular consequences of separability for demand theory were discussed well after the general non-separable case in demand theory had been treated and generally accepted. Among the many contributors are Sono (1945), Leontief (1947), Samuelson (1947), Houthakker (1960), Debreu (1960), and Koopmans (1972). We follow Katzner (1970) in our presentation.

Let $N = \{N_j\}_{j=1}^k$ be a partition of the set $\{1,\ldots,l\}$ and assume that $X = S_1 \times \cdots \times S_k$ Let $J = \{1,\ldots,k\}$ and for any $j \in J$, $y \in X$, $y = (y_1, \ldots, y_k) \in \prod_{i \in J} S_i$ write $y_{-j} = (y_1,\ldots, y_{j-1}, y_{j+1},\ldots,y_k)$ for the vector of components different from j. For any y_{-j}, a preference order $\succsim$ on X induces a preference order $\succsim_{y_{-j}}$ on S_j which is defined by $x_j \succsim_{y_{-j}} x'_j$ if and only if $(y_{-j}, x_j) \succsim (y_{-j}, x'_j)$ for $x_j, x'_j \in S_j$. In general, the induced ordering $\succsim_{y_{-j}}$ will depend on y_{-j}. The first notion of separability states that for any j, the preference orders $\succsim_{y_{-j}}$ are independent of $y_{-j} \in \prod_{i \neq j} S_i$. The second notion of separability states that for any proper subset I of J, the induced preference orders $\succsim_{y_{J/I}}$ on $\prod_{i \in I}$ are independent of $y_{J/I} \in \prod_{i \notin I} S_i$.

Definition Let $\succsim$ be a preference order on $X = \prod_{j \in J} S_j$.

(i) $\succsim$ is called *weakly separable* with respect to N if $\succsim_{y_{-j}} = \succsim_{z_{-j}}$ for each $j \in J$ and any $y_{-j}, z_{-j} \in \prod_{i \neq j} S_i$.
(ii) $\succsim$ is called *strongly separable* with respect to N if $\succsim_{y_{J/I}} = \succsim_{z_{J/I}}$ for each $I \subseteq J, I \neq \emptyset, I \neq J$ and any $y_{J/I}, z_{J/I} \in \prod_{i \notin I} S_i$.

Definition Let $u : \prod_{j \in J} S_j \to \mathbb{R}$. u is called

(i) *weakly separable* with respect to N, if there exist continuous functions $v_j : S_j \to \mathbb{R}, j \in J$, and $V : \mathbb{R}^k \to \mathbb{R}$ such that $u(x) = V(v_1(x_1),\ldots, v_k(x_k))$;
(ii) *strongly separable* with respect to N, if there exist continuous functions v_j: $S_j \to \mathbb{R}$, $j \in J$, and V: $\mathbb{R} \to \mathbb{R}$ such that $u(x) = V(\sum_{j \in J} v_j(x_j))$.

The two important equivalence results on separability are due to Debreu and Katzner. The version of Debreu's theorem given here is slightly weaker than his original result.

Theorem (Katzner 1970) *Let $\succsim$ be a continuous, monotonic preference order on $X = \prod_{j \in J} S_j$ with $S_j = \mathbb{R}^{N_j}$ for all $j \in J$. Then $\succsim$ is weakly separable if and only if every continuous representation of $\succsim$ it is weakly separable.*

Theorem (Debreu 1960) *Let $\succsim$ be a continuous, monotonic preference order on $X = \prod_{j \in J} S_j$ with $S_j = \mathbb{R}^{N_j}$ for all $j \in J = \{1,\ldots, k\}$ and $k \geq 3$. Then $\succsim$ is strongly separable if and only if every continuous representation is strongly separable.*

Under the assumptions of this theorem, if $\succsim$ is strongly separable with representation $u(x) = V(\sum_{j \in J} v_j(x_j))$, then V must be increasing or decreasing.

Therefore,

$$v(x) = \begin{cases} \sum_{j \in J} v_j(x_j) & \text{for } V \text{ increasing} \\ -\sum_{j \in J} v_j(x_j) & \text{for } V \text{ decreasing} \end{cases}$$

is also a representation of $\succsim$. This is the additive form of separable utility used by early economists who thought that each commodity h had its own intrinsic utility representable by a scalar function u_h. The overall utility was then simply obtained as the sum of these functions, $u(x) = \sum_h u_h(x_h)$. Such a formulation is given by Jevons (1871) and Walras (1874) and implicitly contained in Gossen (1854).

In the case of uncertainty, with finitely many states of nature $j \in J = \{1,\ldots, k\}$, respective probabilities $\pi_j > 0$ and consumption $x_j \in S_j$ in state $j \in J$, an additively separable utility representation $u(x) = \sum_{j \in J} v_j(x_j)$ is tantamount to an expected utility representation $u(x) = \sum_{j \in J} \pi_j u_j(x_j)$ with $u_j = v_j/\pi_j$. Hence, an expected utility representation in the tradition of Savage (1954) implies separability with respect to states of nature. In contrast, the novel concept of Choquet expected utility à la Schmeidler (1986, 1989) typically violates separability with respect to states of nature.

For $k = 2$, weak and strong separability of preferences coincide. But there are separable preferences which do not admit a strongly separable utility representation, for instance $X = \mathbb{R}^2_+$, $N_j = \{j\}$ for $j = 1, 2$, $\succsim$ given by $u(x_1, x_2) = \sqrt{x_1} + \sqrt{x_1 + x_2}$. Separability of preferences imposes restrictions on demand correspondences and on demand functions (for details see Barten and Böhm 1982, Sections 9, 14, and 15).

Continuous Demand

Given any price–wealth pair $(p, w) \in \mathbb{R}^{l+1}$, the budget set of the consumer was defined as $\beta(p, w) = \{x \in X | px \leq w\}$. Let $S \subseteq \mathbb{R}^{l+1}$ denote the set of price–wealth pairs for which the budget set is non-empty. Then β describes a correspondence from S into X: that is, β associates to any $(p,w) \in S$ the non-empty subset $\beta(p,w)$ of X.

There are two standard notions of continuity of correspondences, upper hemi-continuity and lower hemi-continuity (see Hildenbrand 1974).

Definition A compact-valued correspondence Ψ from S into an arbitrary subset T of $\mathbb{R}^l$ is *upper hemi-continuous* (u.h.c.) at a point $y \in S$, if for all sequences $(y^n, z^n) \in S \times T$ such that $y^n \to y$ and $z^n \in \Psi(y^n)$ for all n, there exist $z \in \Psi(y)$ and a subsequence z^{n_k} of z^n such that $z^{n_k} \to z$.

Definition A correspondence Ψ from S into an arbitrary subset T of $\mathbb{R}^l$ is *lower hemi-continuous* (l.h.c.) at a point $y \in S$, if for any $z \in \Psi(y)$ and any sequence y^n in S with $y^n \to y$ there exists a sequence z^n in T such that $z^n \to z$ and $z^n \in \Psi(y^n)$ for all n.

Definition A correspondence is *continuous* if it is both lower and upper hemi-continuous.

For single-valued correspondences, the notions of lower and upper hemi-continuity coincide with the usual notion of continuity for functions. For proofs of the following lemmas, see Debreu (1959) or Hildenbrand (1974).

Lemma *Let $X \subseteq \mathbb{R}^l$ be a convex set. Then the budget correspondence $\beta: S \to X$ has a closed graph and is lower hemi-continuous at every point $(p,w) \in S$ for which $w > \min\{px | x \in X\}$ holds.*

Combining a previous corollary on the non-emptiness of the demand set and a fundamental theorem of Berge (1966) yields the next result.

Lemma *Let $X \subseteq \mathbb{R}^l$ be a convex set. If the preference relation has a continuous utility representation, then the demand correspondence is defined (that is, non-empty valued), compact-valued, and upper hemi-continuous at each $(p,w) \in S$ such that $\beta(p,w)$ is compact and $w > \min\{px | x \in X\}$.*

It follows immediately from the definitions that $\phi(\lambda p, \lambda w) = \phi(p,w)$ for any $\lambda > 0$ and any price–wealth pair (p,w): that is, demand is homogeneous of degree zero in prices and wealth. For convex preference orders, the demand correspondence is convexvalued. For strictly convex

preference orders, the demand correspondence is single-valued: that is, one obtains a demand function. The results of this section and of the section on continuous preference orders and utility functions are summarized in the following lemma, which uses the weakest assumptions of traditional demand theory to generate a continuous demand function.

Lemma *Let* $S' := \{(p, w) \in S | \beta(p, w)$ *is compact and* $w > \min \{px | x \in X\}\}$. *If* $\succsim$ *denotes a strictly convex and continuous preference order, then* $\phi(p,w)$ *defines a continuous demand function which satisfies: (i) homogeneity of degree zero in prices and wealth and (ii) the strong axiom of revealed preference.*

Continuous Demand Without Transitivity

Transitivity is often violated in empirical studies. This excludes utility maximization, but not necessarily preference maximization. However, as the next theorem indicates, existence and continuity of demand do not depend on transitivity as crucially as one may expect. The theorem follows from a result by Sonnenschein (1971).

Theorem *Let* $S^* = \{(p, w) \in S | \phi(p, w) \neq \varnothing\}$. *Suppose that X is compact and* $\succsim$ *is complete and has a closed graph.*

(i) *If* $\{x' \in X | x' \succ x\}$ *is convex for all* x ∈ X, *then* $\phi(p, w) \neq \varnothing$ *whenever* $\beta(p, w) \neq \varnothing$ *(that is,* $S^* = S$*).*

(ii) *If* $S^* = S$ *and* $(p^0,w^0) \in S$ *such that* β *is continuous at* (p^0,w^0), *then* ϕ *is u.h.c. at* (p^0,w^0).

The assumption that X is compact is not necessary. For case (i) it suffices that all budget sets $\beta(p,w)$ under consideration be compact. For case (ii) it is sufficient that there exist a compact subset X^0 of X and a neighbourhood S^0 of (p^0,w^0) such that $\phi(S^0) \subseteq X^0$.

To complete this section we state a lemma on the properties of a demand function obtained under preference maximization without transitivity. This contrasts with the lemma at the end of the previous section. Intransitivity essentially implies that the strong axiom of revealed preference need not hold. The lemma follows from the theorem by Sonnenschein and from the result by Shafer (1974).

Lemma *Let* $X = \mathbb{R}^l_+, B = \mathbb{R}^{l+1}_{++}$. *Suppose continuity and strong convexity of* $\succsim$ *(in addition to completeness). Then preference maximization yields a continuous demand function* $f{:}B \to X$ *which satisfies (i) homogeneity of degree zero in prices and wealth and (ii) the weak axiom of revealed preference.*

The converse statement of the lemma does not hold. For $l = 2, X = \mathbb{R}^2_+, B = \mathbb{R}^3_{++}$, there is a C^1-function $f{:}B \to X$ which fulfils (i), (ii), and (iii) $pf(p,w) = w$ for all $(p,w) \in B$, but which cannot be obtained as the demand function for a continuous, complete and strictly convex preference relation (John 1984; Kim and Richter 1986). In addition, John (1995) has shown that continuity of f, (ii) and (iii) imply (i).

Smooth Preferences and Differentiable Utility Functions

Owing to the representation theorem of Debreu, Eilenberg and Rader, continuity of a utility function and continuity of the represented preference order are identical under the perspective of demand theory. When continuous differentiability of demand is required, continuity of the preference relation will not suffice in general. The first rigorous attempt to study 'differentiable preference orders' goes back to Antonelli (1886). We follow the more direct approach of Debreu (1972) to characterize 'smooth preference orders'. Smoothness of preferences is closely related to sufficient differentiability of utility representations and the solution of the integrability problem (see Debreu 1972; also Debreu 1976; Hurwicz 1971; and the section below on integrability). For the purpose of this and subsequent sections, let $P = \mathbb{R}^l_{++}$ denote the (relative) interior of $\mathbb{R}^l_+$ and assume that $X = P$. Let $\succsim$ be a continuous and monotonic preference order on P which we may consider as a subset of $P \times P$: that is, $(x, y) \in \succsim \Leftrightarrow x \succsim y$ for $(x,y) \in P \times P$. Also, the associated indifference relation ~ will be

considered as a subset of $P \times P$. To describe a smooth preference order, differentiability assumptions will be made on the (graph of the) indifference relation in $P \times P$.

For $k \geq 1$, let C^k denote the class of functions which have continuous partial derivatives up to order k, and consider two open sets X and Y in an Euclidean space $\mathbb{R}^n$. A bijection $h{:}X \to Y$ is a *C^k-diffeomorphism* if both h and h^{-1} are of class C^k. $M \subseteq \mathbb{R}^n$ is a *C^k-hypersurface*, if for every $z \in M$, there exist an open neighbourhood U of z, an open subset V of $\mathbb{R}^n$, a hyperplane $H \subset \mathbb{R}^n$ and a C^k-diffeomorphism h: $U \to V$ such that $h(M \cap U) = V \cap H$. A C^k-hypersurface has locally the structure of a hyperplane up to a C^k-diffeomorphism. Considering the indifference relation $\sim$ as a subset of $P \times P$, the set $\tilde{I} = \{(x, y) \in P \times P | x \sim y\} \subset \mathbb{R}^{2l}$ constitutes the 'indifference surface' of the preference relation. Then $\succsim$ is called a *C^2-preference order* (or *smooth preference order*), if $\tilde{I}$ is a C^2-hypersurface.

Theorem (Debreu 1972) *Let $\succsim$ be a continuous and monotonic preference order on P and $\tilde{I}$ be its indifference surface. Then $\succsim$ is a C^2-preference order if and only if it has a monotonic utility representation of class C^2 with no critical point.*

Properties of Differentiable Utility Functions

Utility functions of class C^2 provide the truly classical approach to demand theory (see, for example, Slutsky 1915; Hicks 1939; Samuelson 1947).

Let $\succsim$ be a monotonic, strictly convex C^2-preference order on P and $u : P \to \mathbb{R}$ be a C^2-utility representation of $\succsim$ with no critical point. Then u is continuous, increasing in all arguments, and strictly quasi-concave. Moreover, all second-order partial derivatives $u_{ij}(x) = (\partial^2 u/\partial x_i \partial x_j)(x)$, $i, j = 1, \ldots, l, x \in P$, exist, all u_{ij} are continuous functions of x and $u_{ij} = u_{ji}$ for $i,j = 1,\ldots,l$. Let $D^2u = (u_{ij})$ denote the Hessian matrix of u. Then D^2u is symmetric. The first-order derivatives $u_i(x) = (\partial u/\partial x_i)(x)$, $i = 1, \ldots, l$, are continuous functions of x. Assume that $u_i(x) > 0$ for $i = 1,\ldots,l, x \in P$ and define

$$Du(x) = \begin{bmatrix} u_1(x) \\ \vdots \\ u_l(x) \end{bmatrix}$$

as the gradient of u at x. For any $m \times n$-matrix M, let M' denote the transpose of M.

Theorem *If $u : P \to \mathbb{R}$ is a strictly quasi-concave utility function of class C^2, then $z'D^2u(x)z \leq 0$ for all $x \in P$ and $z \in \{\tilde{z} \in \mathbb{R}^l \mid \tilde{z}Du(x) = 0\}$.* (For a proof, see Barten and Böhm 1982.)

It will be shown in the next section that the conclusion of this theorem does not guarantee the existence of a differentiable demand function. The following definition strengthens the property of strict quasi-concavity.

Definition u is called *strongly quasi-concave* if

$$z'D^2u(x)z < 0 \text{ for all } x \in P, z \neq 0 \text{ and } z \in \{\tilde{z} \in \mathbb{R}^l | \tilde{z}Du(x) = 0\}.$$

Consider the bordered Hessian matrix

$$H(x) = \begin{bmatrix} D^2u(x) & Du(x) \\ [Du(x)]' & 0 \end{bmatrix}.$$

Then u is strongly quasi-concave whenever u is strictly quasi-concave and $H(x)$ is non-singular. (For a proof, see Barten and Böhm 1982).

The properties of strict and strong quasi-concavity are invariant under increasing C^2-transformations. For other results and consequences of differentiable utility functions the reader may consult Barten and Böhm (1982) and the references listed there, or Debreu (1972), Mas-Colell (1974).

Differentiable Demand

The earlier section on continuous demand without transitivity provides sufficient conditions on preferences for the existence of a continuous demand function which is homogeneous of degree zero in prices and wealth and satisfies the strong axiom of revealed preference. In this section, the

implications of smooth preferences for differentiability of demand will be studied. Consider an assumption (D), consisting of the following three parts:

(D1) $X = P$.
(D2) $\succsim$ is a monotonic, strictly convex C^2-preference order on X and the closure relative to $\mathbb{R}^l_+ \times \mathbb{R}^l_+$ of its indifference surface $\tilde{I}$ is contained in $P \times P$.
(D3) The price-wealth space is $B = \mathbb{R}^{l+1}_{++}$.

Given (D), there exists a demand function $f: B \to X$ with $p\, f(p,w) = w$ for all $(p,w) \in B$. Let u be an increasing strictly quasi-concave C^2-utility representation for $\succsim$. The following key result on the differentiability of demand was first given by Katzner (1968). For a detailed proof see Barten and Böhm (1982).

Theorem *Let $(\overline{p},\overline{w}) \in B$ and $\overline{x} = f(\overline{p},\overline{w})$. Then the following assertions are equivalent:*

(i) *f is C^1 in a neighbourhood of $(\overline{p},\overline{w})$.*
(ii) $\begin{bmatrix} D^2u(\overline{x}) & \overline{p}' \\ \overline{p} & 0 \end{bmatrix}$ *is non-singular.*
(iii) *$H(\overline{x})$ is non-singular.*

Once the demand function f is continuously differentiable, it is straightforward to derive all of the well-known comparative statics properties, for the proof of which we refer again to Barten and Böhm (1982). Let $f = (f^1,\ldots,f^l)$ be a demand function of class C^1 and define the respective partial derivatives

$$\begin{aligned} f_w &= (f^1_w,\ldots,f^1_w) = \left(\frac{\partial f^1}{\partial w},\ldots,\frac{\partial f^l}{\partial w}\right), \\ f^i_j &= \frac{\partial f^i}{\partial p_j}, \qquad i,j = 1,\ldots,l; \\ s^i_j &= f^i_j + f^i_w f^i, \qquad i,j = 1,\ldots,l. \end{aligned}$$

From these we obtain the Jacobian matrix of f with respect to prices, $J = \left(f^i_j\right)$, and the so-called Slutsky matrix $S = \left(s^i_j\right)$.

Theorem

(i) $pf_w = 1$, $pJ = -f$,
(ii) $Sp' = 0$,
(iii) *S is symmetric*,
(iv) $ySy' < 0$, *if* $y \in \mathbb{R}^l$, $y \neq \alpha p$ *for all* $\alpha \in \mathbb{R}$,
(v) rank $S = l - 1$.

Property (iv) implies that all diagonal elements of S are strictly negative: that is, $s^i_i = f^i_i + f^i_w f^i < 0$. If $f^i_w > 0$, commodity i is called a *normal good* which implies that $f^i_i < 0$: that is, demand is downward sloping in its own price. On the other hand, a negative income effect $f^i_w < 0$, that is, when commodity i is an *inferior good*, is a necessary, but not a sufficient condition for a positive own price effect $f^i_i > 0$, that is, for commodity i to be a *Giffen good*.

Duality Approach to Demand Theory

With the notion of an expenditure function, an alternative approach to demand analysis is possible which was suggested by Samuelson (1947). For the further development and details, we refer to Diewert (1974, 1982).

As a matter of convenience and for ease of presentation, assumption (D) will be imposed on the preference relation $\succsim$. Let u denote a strictly quasi-concave increasing C^2-utility representation for $\succsim$ and let f: $B \to X$ be the demand function derived from preference maximization. Let us further assume that $u(X) = \mathbb{R}$. (This requirement can always be fulfilled by means of an increasing transformation.) Define the *indirect utility function* $v : B \to \mathbb{R}$ associated with u by $v(p, w) = u(f(p,w))$ for $(p,w) \in B$.

Given a price system $p \in \mathbb{R}^l_{++}$ and a utility level $c \in \mathbb{R}$, let $e(p, c) = \min \{p\, x | x \in X, u(x) \geq c\}$. Since u is strictly quasi-concave and increasing, there exists a unique minimizer $h(p,c)$ of this problem such that $e(p,c) = ph(p,c)$. $h : \mathbb{R}^l_{++} \times \mathbb{R} \to \mathbb{R}^l_{++}$ is called the *Hicksian (income-compensated) demand function* and $e : \mathbb{R}^l_{++} \times \mathbb{R} \to \mathbb{R}_{++}$ is called the *expenditure function* for u. Since assumption (D) holds, preference maximization

and expenditure minimization imply the following properties and relationships:

(i) $c = v[p,e(p,c)]$ for all (p,c).
(ii) $w = e[p,v(p,w)]$ for all (p,w).
(iii) $v(p,\cdot)$ and $e(p,\cdot)$ are inverse functions for any p.
(iv) $h(p,c) = f[p, e(p, c)]$ for all (p,c).
(v) $f(p,w) = h[p,v(p,w)]$ for all (p,w).
(vi) e is strictly increasing and continuous in c.
(vii) e is non-decreasing, positive linear homogeneous, and concave in prices.
(viii) v is strictly increasing in w, and continuous.
(ix) v is non-increasing in prices and homogeneous of degree zero in income and prices.

Moreover, some interesting and important consequences of these properties can be obtained if the functions are sufficiently differentiable.

Theorem

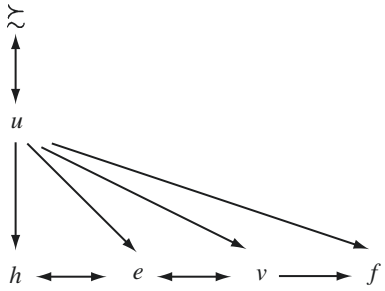
(i) *e is C^k if and only if v is C^k.* $(k = 1, 2)$.
(ii) *If e is C^1, then $\partial e/\partial p = h$.*
(iii) *If f is C^1, then: v is C^2.*
(iv) $f = -(\partial v/\partial p)/(\partial v/\partial w)$ (*Roy's identity*).
(v) *h is C^1 and e is C^2.*
(vi) *$\partial h/\partial p = S$ (Slutsky equation) with $\partial h/\partial p$ evaluated at $[p,v(p,w)]$ and S evaluated at (p,w).*

Integrability

A review of the previous discussions and analytical results involving the concepts of

$\succsim$ preference
u utility
h income-compensated demand function
e expenditure function
v indirect utility
f (direct) demand function

makes apparent their relationships which can be characterized schematically by the following diagram:



where $a \rightarrow b$ indicates that concept b can be derived from concept a under certain conditions.

The integrability problem is to establish $f \rightarrow u$: that is, to recover the utility function from the demand function f.

Two Recent Developments

Advanced microeconomic theory assumes a distribution of consumer characteristics to determine mean demand of a consumption sector. In accordance with traditional demand theory, the primitive characteristics of a consumer are his preference relation $\succsim$ and his wealth w, and possibly his consumption set X. If we disregard the latter, the corresponding distribution of consumer characteristics is a preference–wealth distribution (see Hildenbrand 1974). This approach lends itself to both positive and normative analysis. In contrast, Hildenbrand (1994) and others adopt a purely positive point of view and take pairs (f,w) as the primitive concepts, where f is a demand function not necessarily derived from preference maximization of 'rational' consumers.

Like traditional demand theory, most of theoretical and empirical economics has not distinguished between households and individual consumers. Chiappori (1988, 1992) and others have developed models of collective rationality of multi-person households where each member has his or her own preferences.

See Also

▶ Aggregation (Theory)
▶ Collective Rationality

- ▶ Correspondences
- ▶ Hicksian and Marshallian Demands
- ▶ Integrability of Demand
- ▶ Revealed Preference Theory
- ▶ Separability

Bibliography

Antonelli, G.B. 1886. *Sulla Teoria Matematica della Economia Politica*. Pisa: Nella Tipografia del Folchetto. Trans. as 'On the mathematical theory of political economy', in Chipman et al. (1971).

Arrow, K.J. 1953. Le rôle des valeurs boursières pour la répartition la meilleure des risques. In *Econométrie*, Colloques Internationaux du Centre National de la Recherche Scientifique, vol. 11, 41–47. Paris: Centre National de la Recherche Scientifique.

Arrow, K.J., and F. Hahn. 1971. *General competitive analysis*. San Francisco/Edinburgh: Holden-Day/Oliver and Boyd.

Arrow, K.J., and M.D. Intriligator, eds. 1982. *Handbook of Mathematical economics*. Vol. 2. Amsterdam: North-Holland.

Arrow, K.J., S. Karlin, and P. Suppes. 1960. *Mathematical methods in the social sciences*. Stanford: Stanford University Press.

Barten, A.P. and V. Böhm 1982. Consumer theory. In Arrow and Intriligator (1982).

Beardon, A.F., J.C. Candeal, G. Herden, E. Induráin, and G.B. Mehta. 2002. The non-existence of a utility function and the structure of non-representable preference relations. *Journal of Mathematical Economics* 37: 17–38.

Berge, C. 1966. *Espaces topologiques. Fonctions multivoques*. Paris: Dunod. Trans. as 'Topological spaces'. Edinburgh: Oliver and Boyd, 1973.

Bosi, G., and G.B. Mehta. 2002. Existence of a semicontinuous or continuous utility function: A unified approach and an elementary proof. *Journal of Mathematical Economics* 38: 311–328.

Bowen, R. 1968. A new proof of a theorem in utility theory. *International Economic Review* 9: 374.

Candeal, J.C., and E. Induráin. 1993. Utility functions on chains. *Journal of Mathematical Economics* 22: 161–168.

Chiappori, P.-A. 1988. Rational household labor supply. *Econometrica* 56: 63–89.

Chiappori, P.-A. 1992. Collective labor supply and welfare. *Journal of Political Economy* 100: 437–467.

Chipman, J.S., L. Hurwicz, M.K. Richter, and H.F. Sonnenschein. 1971. *Preferences, utility, and demand*. New York: Harcourt Brace Jovanovich.

Debreu, G. 1954. Representation of a preference ordering by a numerical function. In *Decision processes*, ed. R.-M. Thrall et al. New York: Wiley.

Debreu, G. 1959. *Theory of value*. New York: Wiley.

Debreu, G. 1960. Topological methods in cardinal utility theory. In *Mathematical methods in the social sciences*, ed. K.J. Arrow et al. Stanford: Stanford University Press.

Debreu, G. 1964. Continuity properties of Paretian utility. *International Economic Review* 5: 285–293.

Debreu, G. 1972. Smooth preferences. *Econometrica* 40: 603–615.

Debreu, G. 1976. Smooth preferences. A corrigendum. *Econometrica* 44: 831–832.

Diewert, W.E. 1974. Applications of duality theory. In *Frontiers of quantitative economics*, ed. M.-D. Intriligator and D.A. Kendrick, vol. 2. Amsterdam: North-Holland.

Diewert, W.E. 1982. Duality approaches to microeconomic analysis. In Arrow and Intriligator (1982).

Edgeworth, F.Y. 1881. *Mathematical psychics*. London: Kegan Paul.

Eilenberg, S. 1941. Ordered topological spaces. *American Journal of Mathematics* 63: 39–45.

Estévez Toranzo, M., and C. Hervés Beloso. 1995. On the existence of continuous preference orderings without utility representations. *Journal of Mathematical Economics* 24: 305–309.

Fisher, I. 1892. Mathematical investigations in the theory of value and prices. *Transactions of the Connecticut Academy of Arts and Sciences* 9: 1–124. Repr. in *The works of Irving Fisher*, vol. 1, ed. W.J. Barber. London: Pickering and Chatto, 1997.

Gossen, H.H. 1854. *Entwicklung der Gesetze des menschlichen Verkehrs und der daraus fließenden Regeln für menschliches Handeln*. Braunschweig, 2nd ed. Berlin: Prager, 1889.

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

Hildenbrand, W. 1974. *Core and equilibria of a large economy*. Princeton: Princeton University Press.

Hildenbrand, W. 1994. *Market demand: Theory and empirical evidence*. Princeton: Princeton University Press.

Houthakker, H.S. 1950. Revealed preference and the utility function. *Economica* N.S. 17: 159–174.

Houthakker, H.S. 1960. Additive preferences. *Econometrica* 28: 244–257. Errata: *Econometrica* 30 (1962), 633.

Hurwicz, L. 1971. On the problem of integrability of demand functions. In Chipman et al. (1971).

Hurwicz, L., and H. Uzawa. 1971. On the integrability of demand functions. In Chipman et al. (1971).

Jevons, W.S. 1871. *Theory of political economy*. London: Macmillan.

John, R. 1984. A counterex to a conjecture concerning the nontransitive consumer. Discussion paper no. 151. Sonderforschungsbereich 21, University of Bonn.

John, R. 1995. The weak axiom of revealed preference and homogeneity of demand functions. *Economics Letters* 47: 11–16.

Kannai, Y. 1977. Concavifiability and construction of concave utility functions. *Journal of Mathematical Economics* 4: 1–56.

Katzner, D.W. 1968. A note on the differentiability of consumer demand functions. *Econometrica* 36: 415–418.
Katzner, D.W. 1970. *Static demand theory*. New York: Macmillan.
Katzner, D.W. 1971. Demand and exchange analysis in the absence of integrability conditions. In Chipman et al. (1971).
Kihlstrom, R., A. Mas-Colell, and H. Sonnenschein. 1976. The demand theory of the weak axiom of revealed preference. *Econometrica* 44: 971–978.
Kim, T., and M.K. Richter. 1986. Nontransitive-nontotal consumer theory. *Journal of Economic Theory* 38: 324–363.
Koopmans, T. 1972. Representation of preference orderings with independent components of consumption. In *Decision and organization*, ed. C.B. McGuire and R. Radner. Amsterdam: North-Holland.
Leontief, W. 1947. Introduction to a theory of the internal structure of functional relationships. *Econometrica* 15: 361–373. Repr. in *Selected readings in economic theory*, ed. K.J. Arrow. Cambridge, MA: MIT Press, 1971.
Mas-Colell, A. 1974. Continuous and smooth consumers: Approximation theorems. *Journal of Economic Theory* 8: 305–336.
Mas-Colell, A., M.D. Whinston, and J. Green. 1995. *Microeconomic theory*. Oxford: Oxford University Press.
Neuefeind, W. 1972. On continuous utility. *Journal of Economic Theory* 5: 174–176.
Pareto, V. 1896. *Cours d'economie politique*. Lausanne: Rouge.
Pareto, V. 1906. L'ofelimità nei cicli non chiusi. *Giornale degli economisti* 33: 15–30. Trans. as 'Ophelimity in non-closed cycles', in Chipman et al. (1971).
Rader, T. 1963. The existence of a utility function to represent preferences. *Review of Economic Studies* 30: 229–232.
Richter, M.K. 1966. Revealed preference theory. *Econometrica* 34: 635–645.
Samuelson, P.A. 1947. *Foundations of economic analysis*. Cambridge, MA: Harvard University Press.
Samuelson, P.A. 1950. The problem of integrability in utility theory. *Economica* 17: 355–385.
Savage, L. 1954. *Foundations of statistics*. New York: Wiley.
Schmeidler, D. 1971. A condition for the completeness of partial preference relations. *Econometrica* 39: 403–404.
Schmeidler, D. 1986. Integral representation without additivity. *Proceedings of the American Mathematical Society* 97: 255–261.
Schmeidler, D. 1989. Subjective probability and expected utility without additivity. *Econometrica* 57: 571–587.
Shafer, W. 1974. The nontransitive consumer. *Econometrica* 42: 913–919.
Slutsky, E. 1915. Sulla teoria del bilancio del consumatore. *Giornale degli Economisti e Rivista di Statistica* 51: 1–26. Trans. as 'On the theory of the budget of the consumer', in *Readings in price theory*, ed. G.J. Stigler and K.E. Boulding. Homewood: Irwin, 1953.
Sonnenschein, H. 1971. Demand theory without transitive preferences, with applications to the theory of competitive equilibrium. In Chipman et al. (1971).
Sono, M. 1945. The effect of price changes on the demand and supply of separable goods. *Kokumni Keizai Zasski* 74: 1–51 [in Japanese]. English translation: *International Economic Review* 2 (1960): 239–271.
Uzawa, H. 1960. Preference and rational choice in the theory of consumption. In Chipman et al. (1971).
Walras, L. 1874. *Elements d'économie politique pure*. Lausanne: Corbaz. Trans. W. Jaffé as *Elements of pure economics*. London: Allen and Unwin, 1954.

Demand-Pull Inflation

George L. Perry

Abstract

The term 'demand-pull inflation' originated with the Keynesian macroeconomic model and was used to contrast price increases arising from excess demand with those arising from shocks to aggregate supply. Phillips curve models were initially amended by natural rate models and by models that appended rational expectations and flexible wages and prices to natural rate models. It is now recognized that the response of inflation and unemployment to shifts in aggregate demand itself depends on the inflation environment, and moderate inflation is the desired environment. Stabilization policy continues to distinguish between supply shocks affecting prices and the effects of aggregate demand.

Keywords

Accelerationist inflation models; Aggregate demand; Aggregate supply; Core inflation; Cost-push inflation; Demand-pull inflation; Excess demand; Federal Reserve System; Friedman, M.; Full employment; Incomes policies; Inflation; Inflation targeting; Inflationary expectations; Keynesianism; Monetary policy;

Natural rate of unemployment; Neo-Keynesian models; Organization of Petroleum Exporting Countries; Phelps, E.; Phillips curve; Price control; Rational expectations; Stabilization policies; Sticky prices; Sticky wages; Tobin, J.; Unemployment; Volcker, P.; Wage-price spiral

JEL Classifications
E3

The term 'demand-pull' inflation originated with the simple Keynesian model of the macroeconomy and was used as a contrast to price increases arising from shocks to aggregate supply. In the Keynesian model, there is a well-defined level of potential GDP corresponding to full employment levels of employment and unemployment. Nominal wages are downwardly rigid, so that below full employment aggregate supply increases with prices while aggregate demand decreases. The difference between potential and actual GDP is the output gap, and there is an asymmetry in the economy's response to shifts in demand when output gaps are positive and when they are not. With a positive gap – that is, in the operating region below full employment – an expansion of aggregate demand mainly raises employment and output and only moderately raises prices. But at full employment the aggregate supply curve is vertical, and an expansion of demand only pulls up wages and prices. Hence the term 'demand-pull inflation'.

Macro models of fluctuations have evolved in important ways from this simple Keynesian case. The early empirical Phillips curves described an empirical relation between the level of unemployment and rates of change, rather than levels, of prices. Such relations were estimated from periods characterized by frequent cycles in activity. They did not control for expected or ongoing rates of inflation, so did not directly address the consequences of maintaining real aggregate demand at levels that raised prices. James Tobin (1972), among others, reasoned that the average wage and price increases associated with approaching full employment in the empirical Phillips curves came from the operation of a heterogeneous labour market in which demand constantly shifted among sectors. In his model, the short-run inflation that was observed in the typical cyclical episode reflected wage and price changes that reduced wasteful search unemployment, rather than a misguided attempt to sustain employment above the full employment level.

The first important departure came from theoretical models based on representative agents and firms that examined the consequences of permanently maintaining demand at levels that raised wages and prices in the short run. In the late 1960s Milton Friedman (1968) and Edmund Phelps (1969) independently formulated models of a natural rate of unemployment in which inflation fed back fully into wages and hence prices, so that an unemployment rate below the natural rate could be sustained only by ever-higher inflation rates. In effect, these accelerationist price models resurrected the vertical Keynesian supply curve at full employment for the long run, but allowed demand policies that raised the inflation rate in the short run to achieve lower levels of unemployment, but only temporarily. Since the higher employment associated with price increases could not be sustained, a corollary was that zero inflation was the appropriate target for policy. Tobin's model, with its heterogenous economy, denied that a natural rate identified by prices rising faster corresponded to full employment. However, the natural rate model became widely accepted as a theoretical construct, especially after the introduction of rational expectations models in which anticipation of faster or slower price increases would speed up the process of price acceleration or deceleration. Some theoretical models also assumed price and wage flexibility rather than stickiness. And some even rejected the idea that aggregate demand could leave the economy below full employment, modelling all cyclical variations in output and employment as shocks to aggregate supply. Modern neo-Keynesian models retain both the assumption of price and wage stickiness, which is supported by empirical research, and the implication that output can depart from its potential level. But they attach

a more central role to expectations than do early Keynesian models.

All these models share the original idea of demand-pull inflation in that inflation arises when aggregate demand is excessive. They differ in their description of how the process works out over different time horizons and empirically in how the region of excess demand can be identified for informing forecasters and policymakers. Empirical implementation of rational expectations models continues to be elusive, and most empirical work has used adaptive expectations with accelerationist models to estimate the natural rate and the level of potential output. These estimates proved to be unreliable in the 1990s when economic expansion steadily reduced unemployment rates well below those predicted to cause accelerating inflation in those models. Some recent research has supported the idea that a modest rate of inflation, rather than complete price stability, is necessary to maintain the fullest utilization of resources. This can be so for a variety of reasons. With downward wage rigidity, price stability will keep real wages above their efficient level in a noticeable fraction of firms. Moderate inflation will minimize this problem, permitting the economy to achieve optimal employment (Akerlof et al. 1996). Furthermore, very low inflation rates will be ignored by many economic agents, leading firms to sustain output and employment at levels above those of a full expectational equilibrium (Akerlof et al. 2000). And on the demand side, with very low or zero inflation, the zero floor on nominal interest rates may prevent monetary policy from getting real interest rates low enough to achieve full employment. The experience of Japan after its financial bubble burst is an example (Krugman 1998).

Originally, the explicit modelling of demand-pull inflation was important because of the distinction it drew between price increases arising from excess demand and price increases originating in shifts up in the aggregate supply schedule, also referred to as cost-push. The sharp increases in wage costs that occurred in the heyday of union strength in industrialized economies are important historical examples of shifts in aggregate supply schedules. In the 1960s and 1970s, the experience with such cost-push shocks motivated the attempts to impose wage-price guideposts in the United States, and similar incomes policies in the United Kingdom and elsewhere. Such incomes policies were seen as a way to contain excessive wage and price increases that arose when the economy was operating below its full employment level.

Although there has been no recent interest in incomes policies, the distinction between price increases originating in excess aggregate demand and those originating from shifts in important supply schedules continues to be a feature of policy deliberations and of empirical work today. Core inflation rates, which omit the impact effect of energy and food prices on aggregate price indices, are routinely reported in monthly statistical releases, reflecting a distinction most analysts find useful. Core inflation rates are seen as more likely to feed back into wage increases, and are a better indicator of demand-pull effects on prices. And policymakers regularly make allowances for the effect of supply shocks in considering their stabilization response to changes in reported inflation rates.

History provides examples of significant inflation in which excess demand or major supply shocks or both were important. In the United States, during the Second World War and the Korean War maximizing output was the paramount goal of government even though it meant expanding demand well beyond the normal full-employment point. The potential inflation generated by operating in this excess-demand region was moderated, if not completely suppressed, by rationing and price controls. Demand-pull inflation was also a feature of the industrial economies in the late 1960s, when US military spending was greatly enlarged and labour and product markets became tight for an extended period throughout the industrial world. An abrupt explosion of wage increases at the end of the 1960s and in the early 1970s in most industrialized countries suggests that cost-push contributed importantly to the inflation of that period. The rise in food prices in 1973 and the oil supply shocks of 1973 and 1979 added further to the ongoing inflation of that decade and doubtless contributed to an increase in

inflationary expectations and to the response of unions and firms to those expectations.

It was particularly striking that inflation was so little affected by the very deep recessions of the mid-1970s in the advanced economies. That episode convinced most economists of the shortcomings of the simple short-run Phillips curve model, which predicted that inflation would slow cyclically in the mid-1970s. But it was also not consistent with flexible price accelerationist models which predict that prices and wages will fall when the economy is operating below its natural rate. It did support the pessimistic verdict that a well-established inflation can persist long after the initiating shocks have disappeared and long after a reduction of demand has eliminated any excess demand from the economy.

The stabilization challenge confronting policymakers in that period was seen not merely as avoiding excess aggregate demand, but also as choosing how much to accommodate inflation in order to maintain real growth and how much to give up in output and employment in order to suppress inflation. After the second OPEC oil price shock in 1979, Paul Volcker was appointed Chairman of the US Federal Reserve and, under his leadership, the Fed chose to strongly suppress demand until inflation receded sharply. The lower inflation that ensued is consistent with the predictions of some conventional cyclical models. The severity of the policy used, as reflected in the record high interest rates it produced and the very deep recession that policymakers tolerated, can also be interpreted as evidence that policymakers can shape expectations and that doing so affects how promptly the inflation rate changes.

In the United States, the period that began in the 1990s was a sharp contrast to the 1970s in that inflation had been moderate for many years. As noted above, by the end of the decade the unemployment rate had fallen well below existing empirical estimates from natural-rate models. Yet inflation remained very low, both before the modest recession of the early 2000s and in the several years after it, even after a new oil price shock. Most European economies experienced similarly low inflation in this period. However, several suffered from chronically high rates of unemployment. While considerable controversy surrounds the reasons for this persistence of unemployment, some analysts believe inadequate aggregate demand over an extended period is partly to blame. There are several implications for stabilization policies aimed at avoiding inflation from all this experience: While empirical estimates from the 1970s suggested inflation was prone to quicken through a wage-price spiral, the recent period suggests no such tendency so long as inflation rates are modest (Brainard and Perry 2000). Furthermore, the economy's potential output and the attainable unemployment rate – the thresholds of the demand-pull region of resource utilization – cannot be adequately estimated using typical accelerationist models. Finally, the contrasting experiences across the United States and European economies show that policies targeting inflation alone are not sufficient to assure full employment.

See Also

- ▶ Cost-Push Inflation
- ▶ Inflation
- ▶ Monetary and Fiscal Policy Overview
- ▶ Monetary Business Cycle Models (Sticky Prices and Wages)

Bibliography

Akerlof, G., W. Dickens, and G. Perry. 1996. The macroeconomics of low inflation. *Brookings Papers on Economic Activity* 1: 1–59.

Akerlof, G., W. Dickens, and G. Perry. 2000. Near-rational wage and price setting and the long-run Phillips curve. *Brookings Papers on Economic Activity* 1: 1–44.

Blanchard, O., and L. Summers. 1986. Hysteresis and the European unemployment problem. In *NBER macroeconomics annual*, ed. S. Fischer. Cambridge, MA: MIT Press.

Brainard, W., and G. Perry. 2000. Making policy in a changing world. In *Economic events, ideas, and policies*, ed. G. Perry and J. Tobin. Washington, DC: Brookings Institution.

Calvo, G. 1983. Staggered prices in a utility maximizing framework. *Journal of Monetary Economics* 12: 383–398.

Friedman, M. 1968. The role of monetary policy. *American Economic Review* 58: 1–17.

Gordon, R. 1981. Output fluctuations and gradual price adjustment. *Journal of Economic Literature* 19: 493–530.
Hall, R. (ed.). 1982. *Inflation: Causes and effects*. Chicago: University of Chicago Press for the NBER.
Krugman, P. 1998. It's baaack: Japan's slump and the return of the liquidity trap. *Brookings Papers on Economic Activity* 2: 137–187.
Lucas, R. 1972. Econometric testing of the natural rate hypothesis. In *The econometrics of price determination*, ed. O. Eckstein. Washington, DC: Board of Governors of the Federal Reserve System.
Nickell, S. 1997. Unemployment and labor market rigidities: Europe versus North America. *Journal of Economic Perspectives* 11(3): 55–74.
Nordhaus, W. 1981. Macroconfusion: The dilemma of economic policy. In *Macroeconomics, prices, and quantities*, ed. J. Tobin. Washington, DC: Brookings Institution.
Okun, A. 1975. Inflation: Its mechanics and welfare costs. *Brookings Papers on Economic Activity* 2: 351–390.
Okun, A. 1981. *Prices and quantities: A macroeconomic analysis*. Washington, DC: Brookings Institution.
Perry, G. 1980. Inflation in theory and practice. *Brookings Papers on Economic Activity* 1: 207–241.
Phelps, E. 1969. The new microeconomics in inflation and employment theory. *American Economic Review* 59: 147–160.
Sargent, T. 1982. The ends of four big inflations. In *Inflation: Causes and consequences*, ed. R. Hall. Chicago: University of Chicago Press.
Schultze, C. 1981. Some macro foundations for micro theory. *Brookings Papers on Economic Activity* 2: 521–576.
Taylor, J. 1980. Aggregate dynamics and staggered contracts. *Journal of Political Economy* 88: 1–24.
Tobin, J. 1972. Inflation and unemployment. *American Economic Review* 62: 1–18.

Democratic Paradoxes

Norman Schofield

Abstract

Formal models of voting have emphasized the *mean voter theorem,* namely, that all parties should rationally adopt identical positions at the electoral mean. The lack of evidence for this assertion is a *paradox* which this article attempts to resolve by considering an electoral model that includes 'valence' or non-policy judgements by voters of party leaders. In a polity such as Israel, based on proportional electoral rule, low-valence parties would adopt positions far from the centre, making coalition formation unstable. In Britain, by contrast, a party with a low-valence leader would be subject to the demands of non-centrist activists.

Keywords

Condorcet, Marquis de; Democratic paradoxes; Downs, A.; Hotelling, H.; Local Nash equilibrium; Madison, J.; Median voter theorem; Mixed strategy Nash equilibrium; Plurality electoral rule; Political competition; Proportional representation; Pure strategy Nash equilibrium; Valence; Vote maximizing strategies; Voting

JEL Classifications

D7

Models of elections tend to give two quite contradictory predictions about the result of political competition. In two-party competition, if the 'policy space' involves two or more independent issues, then 'pure strategy Nash equilibria' generally do not exist and instability or *chaos* may occur (see Plott 1967; McKelvey 1976, 1979; Schofield 1978, 1983, 1985; McKelvey and Schofield 1986, 1987; Saari 1997; Austen-Smith and Banks 1999). That is to say, whatever position is picked by one party, there always exists another policy point which will give the second party a majority over the other. Moreover, vote maximizing strategies could lead political candidates to wander all over the policy space.

On the other hand, the earlier electoral models based on the work of Hotelling (1929) and Downs (1957) suggest that parties will converge to an electoral centre (at the electoral *median*) when the policy space has a single dimension. (An equilibrium can also be guaranteed as long as the decision rule requires a sufficiently large majority – Schofield 1984; Strnad 1985; Caplin and Nalebuff 1988 – or when the electoral distribution has a certain concavity property – Caplin and Nalebuff 1991.) Although a pure strategy Nash equilibrium generically fails to exist in competition between two agents under majority rule in

high enough dimension, there will exist mixed strategy Nash equilibria (Kramer 1978) whose support lies within a subset of the policy space known as the 'uncovered set' (see McKelvey 1986; Banks et al. 2002). These various and contrasting theoretical results can be seen as a paradox: will democracy tend to generate centrist compromises, or can it lead to chaos? This question is of fundamental importance in a world in which many countries are experimenting with democracy for the first time.

Partly as a result of these theoretical difficulties with the 'deterministic' electoral model, and also because of the need to develop empirical models of voter choice (Poole and Rosenthal 1984), attention has focused on 'stochastic' vote models. A formal basis for such models is provided by the notion of 'quantal response equilibria' (McKelvey and Palfrey 1995). In such models, the behaviour of each voter is modelled by a vector of choice probabilities (Lin et al. 1999). A standard result in this class of models is that all parties converge to the electoral origin when the parties are motivated to maximize vote share or plurality (in the two-party case) (see McKelvey and Patty 2006; Banks and Duggan 2005). The predictions concerning convergence are at odds with empirical evidence that parties appear to diverge from the electoral centre (Merrill and Grofman 1999; Adams 2001; Schofield and Sened 2006).

The *paradox* that actual political systems display neither *chaos* nor *convergence* is the subject of this article. The key idea is that the convergence result need not hold if there is an asymmetry in the electoral perception of the 'quality' of party leaders (Stokes 1992). The average weight given to the perceived quality of the leader of the j^{th} party is called the party's 'valence'. In empirical models this valence is independent of the party's position, and adds to the statistical significance of the model. In general, valence reflects the overall degree to which the party is perceived to have shown itself able to govern effectively in the past, or is likely to be able to govern well in the future (Penn 2003). The early empirical model of US presidential elections by Poole and Rosenthal (1984) included these valence terms. The authors noted that there was no evidence of candidate convergence.

Formal models of elections incorporating valence have been developed (Ansolabehere and Snyder 2000; Groseclose 2001; Aragones and Palfrey 2002), but the theoretical results to date have been somewhat inconclusive. Extension to the multiparty case is of interest because of recent empirical models of voting in the Netherlands and Germany (Schofield et al. 1998a, b; Quinn et al.1999; Quinn, and Martin 2002), Britain (Schofield 2005a, b), Israel (Schofield et al. 1998a, b; Schofield and Sened 2002, 2005, 2006) and Italy (Giannetti and Sened 2004). All these empirical models have suggested that divergence is generic. Most of these empirical models have been based on the 'multinomial logit' assumption that the stochastic errors had a 'Type I extreme value distribution' (Dow and Endersby 2004).

Schofield (2007) provides a 'classification theorem' for the formal vote model based on the same stochastic distribution assumption. The 'policy space' is assumed to be of dimension w, and there is an arbitrary number, p, of parties. The party leaders exhibit differing valence. A 'convergence coefficient' incorporating all the parameters of the model can be defined. Instead of using the notion of a Nash equilibrium, the result is given in terms of the existence of a 'local Nash equilibrium'. It is shown that there are necessary and sufficient conditions for the existence of a 'pure strategy vote maximizing local Nash equilibrium' (LNE) at the mean of the voter distribution. When the necessary condition fails, then parties, in equilibrium, will adopt divergent positions. In general, parties whose leaders have the lowest valence will take up positions furthest from the electoral mean. Moreover, because a pure strategy Nash equilibrium (PNE) must be a local equilibrium, the failure of existence of the LNE at the electoral mean implies non-existence of such a centrist PNE. The failure of the necessary condition for convergence has a simple explanation: if the variance of the electoral distribution is sufficiently large in contrast to the expected vote share

of the lowest-valence party at the electoral mean, then this party has an incentive to move away from the origin towards the electoral periphery. Other low-valence parties will follow suit, and the local equilibrium will result with parties distributed along a 'principal electoral axis'.

An empirical study of voter behaviour for Israel for the election of 1996 (based on Schofield and Sened 2005) is used to show that the necessary condition for party convergence failed for this election. The equilibrium positions obtained from the formal result, under vote maximization, are in general comparable with, but not identical to, the estimated positions. The two highest-valence parties (Labour and Likud) were symmetrically located on either side of the electoral origin, while the lowest-valence parties were located far from the origin. In such a polity, based on a proportional electoral system, it is generally necessary to form coalition governments. The existence of small, low-valence, radical parties on the electoral periphery may create serious difficulties in the formation of majority government. It is possibly for this reason that Ariel Sharon, formerly leader of Likud, and Shimon Peres, formerly leader of Labour, in 2005 formed Kadima, a new centrist party.

This article also presents results from analysis of the 1997 election in Britain (Schofield 2005a, b). In this case the empirical estimates of the parameters of the model, taken together with the formal analysis, suggest that convergence should have occurred. Instead the Conservative Party was estimated to be at a position far from the electoral centre. It is suggested that the discrepancy between the formal and the empirical models can be accommodated by considering the effect of activists on the optimal party position. Since concerned activists will raise funds for the party, but only if the party adopts a policy position that accords with activists' concerns, there is a tension between activist demands and the electoral concerns of the party leadership. The model based on activist support estimates the marginal trade-off generated by opposed activist groups within a party. It is suggested that the low valence of recent Conservative leaders obliged them to seek support from activists supporting British sovereignty against the European Union, and thus to take up radical positions on the second, 'European' axis.

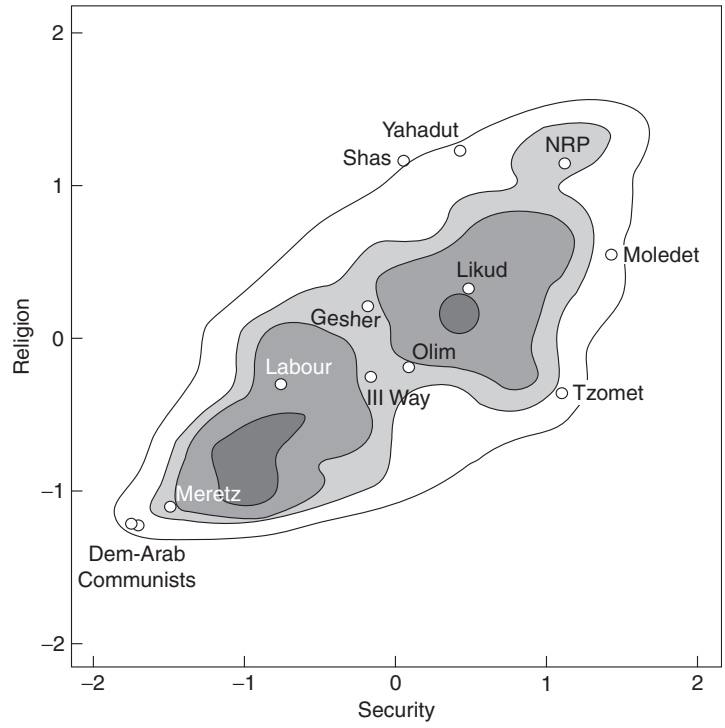
In contrast, the apparent move by the Labour Party towards the electoral centre between 1992 and 1997 was a consequence of the increase of the electoral valence of Tony Blair, the leader of the party, rather than a cause of this increase.

Recent work by Miller and Schofield (2003) using this model suggests that, in the United States, the movement of presidential candidates in a two-dimensional policy space generated by economic and social dimensions is the result of contending and opposed activist groups.

The underlying premise of the notion of the *local Nash equilibrium,* used in these models, is that party leaders will not consider 'global' changes in party policies, but will instead propose small changes in the party position in response to changes in beliefs about electoral response. These models regard elections as the aggregation of both electoral evaluation or 'valence' and electoral preferences. Valence can be regarded as that element of a voter's choice which is determined by judgement rather than preference. This accords well with the arguments of James Madison in *Federalist 10* of 1787 (Rakove 1999) and of Condorcet (1785) in his treatise on social choice theory. Schofield (2005c, 2006) provides a discussion of the relevance of these valence models for the constitutional basis of the US polity.

Empirical Analysis for Israel

Figure 1 shows the estimated positions of the parties at the time of the 1996 Israeli election. Fig. 1 also gives the estimated distribution of voter ideal points for the 1996 election, based on a factor analysis of the survey responses derived from the survey of Arian and Shamir (1999). The two dimensions of policy deal with attitudes to the Palestine Liberation Organization (PLO) (the horizontal axis) and religion (the vertical). The party positions were obtained from analysis of party manifestos (Schofield et al. 1998a, b; Schofield



Democratic Paradoxes, Fig. 1 Voter distribution and estimated party positions in the Knesset at the 1996 election

and Sened 2005, 2006). With the use of information on the individual voter intentions, it is possible to construct a multinomial logit model (based on the Type I extreme value distribution).

The model assumes that the voter utility vector has the form $\mathbf{u}_i(xi, \mathbf{z}) = (u_{i1}(x_i, z_1), \ldots, u_{ip}(x_i, z_p))$ where

$$u_{ij}(x_i, z_j) = u^*_{ij}(x_i, z_j) + \varepsilon_j$$
$$and\ u^*_{ij}(x_i, z_j) = \lambda_j - \beta||x_i - z_j||^2.$$

Here the position of voter i is x_i while the position of party j is z_j. The term $||x_i - z_j||$ is the distance between these two points. The *valences* of the p parties are given by the vector $\lambda = (\lambda_p, \lambda_{p-1}, \ldots, \lambda_2, \lambda_1)$ and are ranked

$$\lambda_p \geq \lambda_{p-1} \geq \ldots \geq \lambda_2 \geq \lambda_1.$$

The error terms $\{\varepsilon_j\}$ have the *Type I extreme value distribution,* Ψ.

(The cumulative distribution, Ψ, takes the closed form $\Psi(h) = \exp[-\exp[-h]]$.)

The probability that a voter i chooses party j is

$$\rho_{ij}(\mathbf{z}) = \Pr\left[\left[u_{ij}(x_i, z_j) > u_{il}(x_i, z_l)\right],\ \text{for all } l \neq j\right].$$

Here Pr stands for the probability operator associated with Ψ. The expected vote share of agent j is

$$V_j(\mathbf{z}) = \frac{1}{n_i}\sum_{\in N} \rho_{ij}(\mathbf{z}).$$

This model is denoted M (λ, β; Ψ). A *local pure strategy Nash equilibrium* (LNE) is simply a vector $\mathbf{z} = (z_1,..., z_p)$ of party positions with the property that each z_j locally maximizes $V_j(\mathbf{z})$, taking the other party positions A necessary condition for $\mathbf{z}^* = (\mathbf{0},\ldots, \mathbf{0})$ to be *pure strategy Nash equilibrium* (PNE) is that it be a LNE and thus that all Hessians have eigenvalues at $\mathbf{z}^*$ that are non-positive. This can be expressed as a single necessary condition on a 'convergence coefficient' defined terms of the Hessian of the vote

share function of the party with the lowest valence (Schofield 2006b). Since the lowest-valence party is the National Religious Party (NRP) (for the 1996 model for Israel), a *necessary* condition for the NRP vote share to be maximized at the origin is that *both* eigenvalues of this Hessian be nonpositive. However, the calculation given below shows that that one of the eigenvalues was positive. It follows that the NRP position that maximizes its vote share is *not* at the origin. Thus the convergent position (**0**,. . ., **0**) cannot be a Nash equilibrium to the vote maximizing game.

Indeed it is obvious that there is a principal component of the electoral distribution, and this axis is the eigenspace of the positive eigenvalue. It follows that low-valence parties should then position themselves on this eigenspace, as illustrated in the simulation given in Fig. 2.

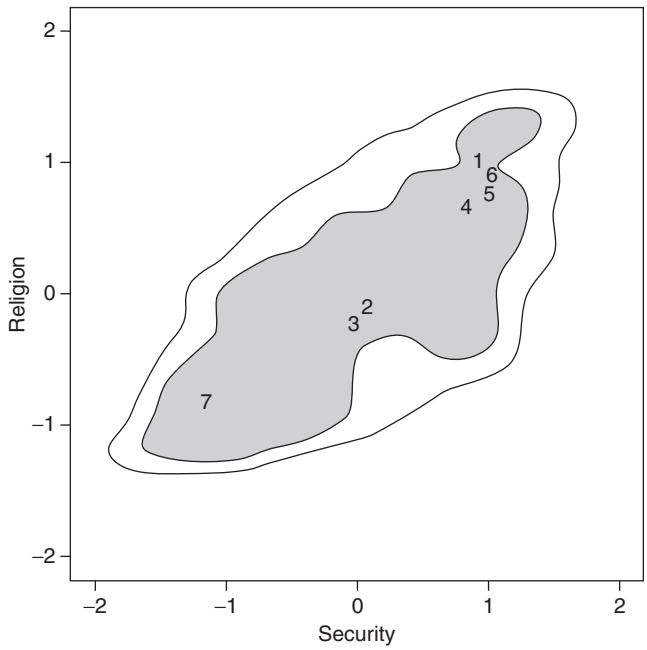
To present the calculation, we use the fact that the valence of the NRP was – 4.52. The spatial coefficient is $\beta = 1.12$. Because the valences of the major parties are 4.15 and 3.14, the formal analysis implies that, when all parties are at the origin, the vote share, ρ_{NRP} can be computed to be

$$\rho_{NRP} \simeq \frac{1}{1 + e^{4.15+4.52} + e^{3.14+4.52}} \simeq 0.$$

Moreover, the Hessian of the NRP at the origin depends on the electoral variance and this is

$$C_{NRP} = 2(1.12)\begin{pmatrix} 1.0 & 0.591 \\ 0.591 & 0.732 \end{pmatrix} - I$$
$$= \begin{pmatrix} 1.24 & 1.32 \\ 1.34 & 0.64 \end{pmatrix}.$$

The eigenvalues of the NRP Hessian at the origin are 2.28 and –0.40, giving a saddle point. Thus, the origin cannot be a Nash equilibrium. The 'convergence coefficient' can be calculated to be 3.88, larger than the necessary upper bound of 2.0. The major eigenvector for the NRP is (1.0,0.8), and along this axis the NRP vote share function increases as the party moves away from the origin. The minor, perpendicular axis is given by the vector (1, – 1.25) and on this axis the NRP vote share decreases. Figure 2 gives one of the local equilibria in 1996, obtained by simulation of the model. The figure makes it clear that the vote maximizing positions lie on the principal axis



Democratic Paradoxes, Fig. 2 A simulated local Nash equilibrium in the vote maximizing game in Israel in 1996. *Note:* 1: Shas; 2: Likud; 3: Labour; 4: NRP; 5: Molodet; 6: III Way; 7: Meretz

through the origin and the point (1.0, 0.8). In all, five different LNE were located. However, in all the equilibria the two high-valence parties, Labour and Likud, were located at precisely the same positions, as shown in Fig. 2. The only difference between the various equilibria was that the positions of the low-valence parties were perturbations of each other.

Figure 2 suggests that the simulation was compatible with the predictions of the formal model based on the extreme value distribution. All parties were able to increase vote shares by moving away from the origin, along the principal axis, as determined by the large, positive principal eigenvalue. In particular, the simulation confirms the logic of the above analysis. Low-valence parties, such as NRP and Shas, in order to maximize vote shares must move far from the electoral centre. Their optimal positions will lie in either the north-east quadrant or the south-west quadrant. The vote maximizing model, without any additional information, cannot determine which way the low-valence parties should move. As noted above, the simulations of the empirical models found multiple LNE essentially differing only in permutations of the low-valence party positions.

In contrast, since the valence difference between Labour and Likud was relatively low, their optimal positions would be relatively close to, but not identical to, the electoral mean. The simulation for the elections of 1988 and 1992 are also compatible with this theoretical inference. Figure 2 also suggests that every party, in local equilibrium, should adopt a position that maintains a minimum distance from every other party. The formal analysis, as well as the simulation exercise, suggests that this minimum distance depends on the valences of the neighbouring parties. Intuitively it is clear that, once the low-valence parties vacate the origin, then high-valence parties like Likud and Labour will position themselves almost symmetrically about the origin, and along the major axis.

Comparison between Fig. 1, of the estimated party positions, and Fig. 2, of simulated equilibrium positions, reveals a notable disparity particularly in the position of Shas. In 1996 Shas was pivotal between Labour and Likud, in the sense that, to form a winning coalition government, either of the two larger parties required the support of Shas. It is obvious that the location of Shas in Fig. 1 suggests that it was able to bargain effectively over policy and, presumably, perquisites. Indeed, it is plausible that the leader of Shas was aware of this situation, and incorporated this awareness in the utility function of the party.

The close correspondence between the simulated LNE based on the empirical analysis and the estimated actual political consuggests that the true utility function for each party j has the form $U_j(\mathbf{z}) = V_j(\mathbf{z}) + \delta_j(\mathbf{z})$, where $\delta_j(\mathbf{z})$ may depend on the beliefs of party leaders about the post-election coalition possibilities, as well as the effect of activist support for the party.

This hypothesis leads to the further hypothesis that, for the set of feasible strategy profiles in the Israel polity, $\delta_j(\mathbf{z})$ is 'small' relative to $V_j(\mathbf{z})$. A formal model to this effect could indicate that the LNE for $\{U_j\}$ would be close to the LNE for $\{V_j\}$. Note, however, that this perturbation of the party utility function causes parties to leave the main electoral axis. It is possibly for this reason that coalition politics in Israel has been very complex.

The Likud Party, under Ariel Sharon, was constrained by the religious parties in its governing coalition. This apparently caused Sharon to leave Likud to set up a new centrist party, Kadima ('Forward') with Shimon Peres, previously leader of Labour. The reason for this reconfiguration was the victory on 10 November 2005 of Amir Peretz over Peres for leadership of the Labour Party, and Peretz's move to the left along the principal electoral axis.

Consistent with the model presented here, Sharon's intention was to position Kadima very near the electoral centre on both dimensions, to take advantage of his high valence among the electorate. Sharon's subsequent hospitalization had an adverse effect on the valence of Kadima, under its new leader, Ehud Olmert. Even so, in the election of 28 March 2006 Kadima took 29 seats, against 19 seats for Labour, and only 12 for Likud. One surprise was a new centrist pensioners' party with 7 seats. Because Kadima with Labour and the other parties of the left had 70 seats, Olmert was able to put together a majority coalition on

28 April, including the Orthodox party Shas. As Fig. 1 illustrates, Shas is centrist on the security dimension, indicating that this was the key issue of the election.

Empirical Analysis for Britain

This section analyses the general election in Britain in 1997 in order to suggest how activists for the parties may influence party positioning. The analysis shows that the valence model as presented above cannot always explain divergence of party positions. For example, Fig. 3 shows the estimated positions of the party leaders, based on a survey of party MPs in 1997 (Schofield 2005a, b). In addition to the Conservative Party, Labour Party, and Liberal Democrats, responses were obtained from Ulster Unionists, Scottish Nationalists and Plaid Cymru (Welsh Nationalists). The axis is economic, the second pro or anti the European Union. The electoral model was estimated for the election in 1997, using only the economic dimension.

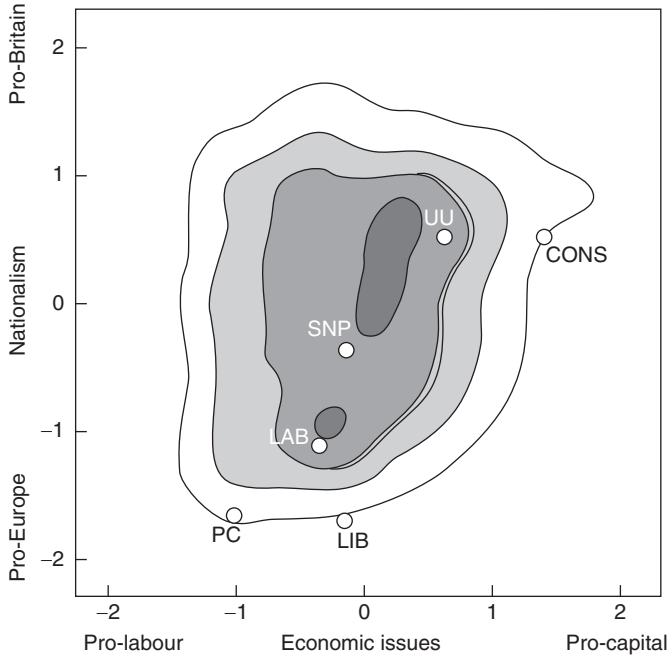
For this election, we $(\lambda_{con}, \lambda_{lab}, \lambda_{lib}, \beta)_{1997} = (+1.24, 0.97, 0.0, 0.5)$ so the probability ρ_{lib}, that a voter chooses the Liberal Democrats is

$$\rho_{lib} = \frac{e^0}{e^0 + e^{1.24} + e^{0.97}} = \frac{1}{7.08} = 0.14.$$

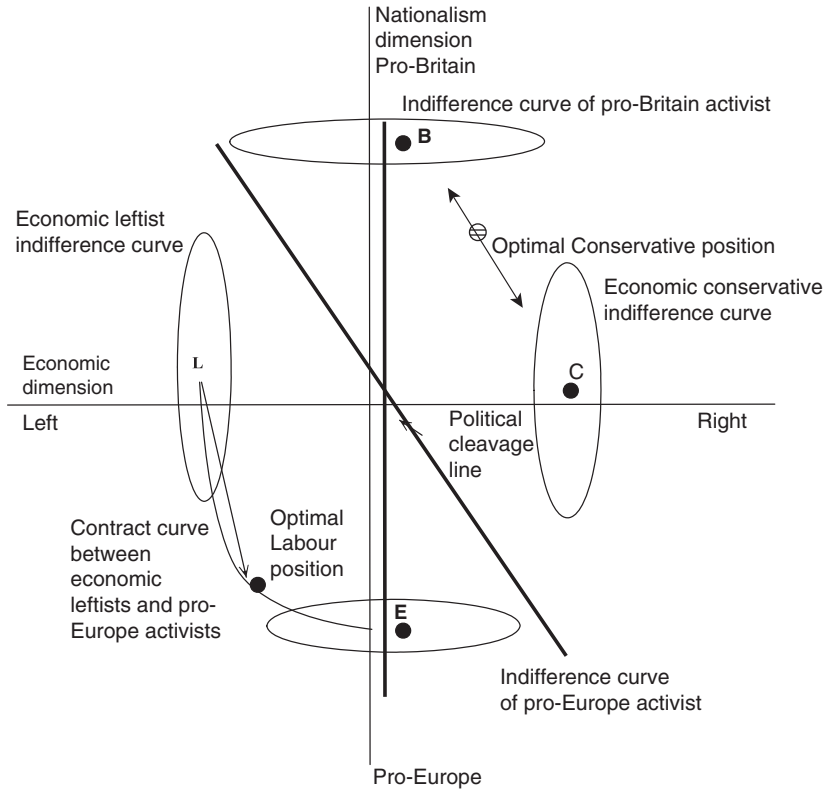
The Hessian for this party at the origin is $C_{lib} = -0.28$, which is compatible with a Nash equilibrium at the origin. Extending the model to two dimensions gives a Hessian

$$C_{lib} = (0.72)\begin{pmatrix} 1.0 & 0 \\ 0 & 1.5 \end{pmatrix} - I = \begin{pmatrix} -0.28 & 0 \\ 0 & +0.8 \end{pmatrix}.$$

According to the formal model, all parties should have converged to the origin on the first axis. Because the eigenvalue for the Liberal Democrats is positive on the second axis, we have an explanation for its position away from the origin on the Europe axis. However, there is no explanation for the location of the Conservative Party so far from the origin on both axes. Schofield (2005a, b) adapts the activist model of Aldrich (1983a, b) wherein the falling exogeneous valence of the Conservative Party leader increases the marginal importance of two opposed activist groups in the party:



Democratic Paradoxes, Fig. 3 Estimated party positions in the British Parliament for a two-dimensional model. *Notes:* Highest-density contours of the voter sample distribution at the 95%, 75%, 50% and 10% levels. CONS: Conservative Party; LAB: Labour Party; LIB: Liberal Democrats; PC: Plaid Cymru (Welsh Nationalists); SNP: Scottish National Party; UU: Ulster Unionist Party (*Source:* MP survey data and a National Election Survey)



Democratic Paradoxes, Fig. 4 Illustration of vote maximizing positions of Conservative and Labour Party leaders in a two-dimensional policy space

one group 'pro-capital' and one group 'anti-Europe', as in Fig. 4.

The optimal Conservative position will be determined by balancing the electoral effects of these two groups. The optimal position for this party will be one which is 'closer' to the locus of points that generates the greatest activist support. This locus is where the joint marginal activist pull is zero. This locus of points can be called the 'activist contract curve' for the Conservative Party.

Note that in Fig. 4 the indifference curves of representative activists for the parties are described by ellipses. This is meant to indicate that preferences of different activists on the two dimensions may accord different saliences to the policy axes. The 'activist contract curve' given in the figure, for Labour, say, is the locus of points satisfying the first order. This curve represents the balance of power between Labour supporters more interested in economic issues (centred at L in the figure and those more interested in Europe (centred at E). The optimal positions for the two parties will be at appropriate positions that satisfy the optimality condition.

According to this model, a party's optimal position will tend to be nearer to the electoral origin when the valence of the party leader is higher. In contrast, a party whose leader has low valence will be more influenced by activist groups, and will tend to adopt a position further from the electoral centre and nearer to the position preferred by the dominant activist group. This model has been applied to the US polity by Miller and Schofield (2003) and Schofield et al. (2003).

Proportional Representation and Plurality Rule

Most of the early work in formal political theory focused on two-party competition, and generally

concluded that there would be strong centripetal electoral forces causing parties to converge to the electoral centre. The extension of this theory to the multiparty context, common in European polities, has proved very difficult because of the necessity of dealing with coalition governments (Riker 1962). However, the symmetry conditions developed by McKelvey and Schofield (1987) showed that a large, centrally located party could dominate policy if it occupied what is known as a 'core position'. Thus, in situations where there is a stable policy core there would be certainty over the post-election policy outcome of coalition negotiation (Laver and Schofield 1998). Absent a policy core, the post-election outcome will be a lottery across various possible coalitions, all of which are associated with differing policy outcomes and cabinet allocations. Modelling this post-election 'committee game' can be done with cooperative game theoretical concepts (Banks and Duggan 2000).

Although the non-cooperative stochastic electoral model presented here can give insight into the relationship between electoral preferences and beliefs (regarding the valences of party leaders), it is still incomplete. The evidence suggests that party leaders pay attention not only to electoral responses but also to the post-election coalition consequences of their choices of policy positions. Nonetheless, the combination of the electoral model and post-election bargaining theory (Schofield and Sened 2002) suggests the following.

In a polity based on a proportional electoral rule, the high-valence parties will be attracted towards the electoral centre. However, if there are two such competing parties of similar valence neither will locate quite at the centre. There may be many low-valence parties, whose equilibrium, vote maximizing positions will be far from the electoral centre. In order to construct winning coalitions, one or other of the high-valence parties must bargain with more 'radical' low-valence parties, and this could induce a degree of coalitional instability. However, it is possible that a charismatic leader, such as Sharon in Israel, can adopt a centrist position and dominate politics by controlling the policy core.

In a polity based on a plurality electoral rule, the disproportionality between votes and seats may increase the importance of activist groups. A party with a relatively low-valence leader will be forced to depend on activist support. Consequently, the party will be obliged to move to a more radical position so to attract activist support.

This may provide a reason why Britain's Labour Party appeared to acquiesce to the demands of its left-wing supporters during the leadership of Michael Foot in 1980–3 and of Neil Kinnock in 1983–92. This led to Labour defeats in the elections between 1983 and 1992. Tony Blair became Labour leader following the death of John Smith in 1994 and his high valence allowed him to overcome union opposition and to craft the centrist 'New Labour' policies that led to Labour victories in the elections of 1997, 2001 and 2005.

Concluding Remarks

To sum up, these models suggest how the democrat paradox can be resolved: convergence to an electoral centre is not a generic phenomenon, but can occur when a party leader is generally regarded by the electorate to be of superior quality or valence. Chaos does not occur in these models, though a degree of coalitional instability is possible under proportional electoral rule when there is no highly regarded political leader at the policy core.

See Also

- ▶ Political Competition
- ▶ Rational Behaviour
- ▶ Rational Choice and Political Science

Acknowledgment This article is based on research supported by NSF Grant SES 024173. The table and figures are reproduced from Schofield and Sened (2006) by permission of Cambridge University Press.

Bibliography

Adams, J. 2001. *Party competition and responsible party government*. Ann Arbor: University of Michigan Press.
Adams, J., and S. Merrill III. 1999. Modeling party strategies and policy representation in multiparty elections:

Why are strategies so extreme? *American Journal of Political Science* 43: 765–791.

Aldrich, J. 1983a. A spatial model with party activists: Implications for electoral dynamics. *Public Choice* 41: 63–100.

Aldrich, J. 1983b. A Downsian spatial model with party activists. *American Political Science Review* 77: 974–990.

Ansolabehere, S., and J. Snyder. 2000. Valence politics and equilibrium in spatial election models. *Public Choice* 103: 327–336.

Aragones, E., and T. Palfrey. 2002. Mixed equilibrium in a Downsian model with a favored candidate. *Journal of Economic Theory* 103: 131–161.

Aragones, E., and T. Palfrey. 2005. Spatial competition between two candidates of different quality: The effects of candidate ideology and private information. In *Social choice and strategic decisions*, ed. D. Austen-Smith and J. Duggan. Heidelberg: Springer.

Arian, A., and M. Shamir. 1999. *The election in Israel: 1996*. Albany: SUNY Press.

Austen-Smith, D., and J. Banks. 1999. *Positive political theory I*. Ann Arbor: University of Michigan Press.

Banks, J., and J. Duggan. 2000. A bargaining model of collective choice. *American Political Science Review* 94: 73–88.

Banks, J., and J. Duggan. 2005. The theory of probabilistic voting in the spatial model of elections. In *Social choice and strategic decisions*, ed. D. Austen-Smith and J. Duggan. Heidelberg: Springer.

Banks, J., J. Duggan, and M. Le Breton. 2002. Bounds for mixed strategy equilibria and the spatial model of elections. *Journal of Economic Theory* 103: 88–105.

Caplin, A., and B. Nalebuff. 1988. On 64% majority rule. *Econometrica* 56: 787–814.

Caplin, A., and B. Nalebuff. 1991. Aggregation and social choice: A mean voter theorem. *Econometrica* 59: 1–23.

Condorcet, N. 1785. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix.* Paris*:* Imprimerie Royale*. Trans. in part in I. McLean and F. Hewitt. Condorcet: Foundations of social choice and political theory.* Aldershot*:* Edward Elgar*.* 1994*.*

Coughlin, P. 1992. *Probabilistic voting theory*. Cambridge: Cambridge University Press.

Dow, J., and J. Endersby. 2004. Multinomial probit and multinomial logit: A comparison of choice models for voting research. *Electoral Studies* 23: 107–122.

Downs, A. 1957. *An economic theory of democracy*. New York: Harper and Row.

Enelow, J., and M. Hinich. 1984. *The spatial theory of voting*. Cambridge: Cambridge University Press.

Giannetti, D., and I. Sened. 2004. Party competition and coalition formation: Italy 1994–1996. *Journal of Theoretical Politics* 16: 483–515.

Groseclose, T. 2001. A model of candidate location when one candidate has a valance advantage. *American Journal of Political Science* 45: 862–886.

Hinich, M. 1977. Equilibrium in spatial voting: The median voter result is an artifact. *Journal of Economic Theory* 16: 208–219.

Hotelling, H. 1929. Stability in competition. *Economic Journal* 39: 41–57.

Kramer, G. 1978. Existence of electoral equilibrium. In *Game theory and political science*, ed. P. Ordeshook. New York: New York University Press.

Laver, M., and N. Schofield. 1998. *Multiparty government: The politics of coalition in Europe*. Ann Arbor: Michigan University Press.

Lin, T.-M., J. Enelow, and H. Dorussen. 1999. Equilibrium in multicandidate probabilistic spatial voting. *Public Choice* 98: 59–82.

McKelvey, R. 1976. Intransitivities in multidimensional voting models and some implications for agenda control. *Journal of Economic Theory* 12: 472–482.

McKelvey, R. 1979. General conditions for global intransitivities in formal voting models. *Econometrica* 47: 1085–1111.

McKelvey, R. 1986. Covering, dominance and institution-free properties of social choice. *American Journal of Political Science* 30: 283–314.

McKelvey, R., and T. Palfrey. 1995. Quantal response equilibria for normal form games. *Games and Economic Behavior* 10: 6–38.

McKelvey, R., and J. Patty. 2006. A theory of voting in large elections. *Games and Economic Behavior* 57: 155–180.

McKelvey, R., and N. Schofield. 1986. Structural instability of the core. *Journal of Mathematical Economics* 15: 179–198.

McKelvey, R., and N. Schofield. 1987. Generalized symmetry conditions at a core point. *Econometrica* 55: 923–933.

Merrill, S. III, and B. Grofman. 1999. *A Unified Theory of Voting*. Cambridge: Cambridge University Press.

Miller, G., and N. Schofield. 2003. Activists and partisan realignment in the U.S. *American Political Science Review* 97: 245–260.

Penn, E. 2003. A model of far-sighted voting. Working paper, Institute of Quantitative Social Science, Harvard University.

Plott, C. 1967. A notion of equilibrium and its possibility under majority rule. *American Economic Review* 57: 787–806.

Poole, K., and H. Rosenthal. 1984. U.S. presidential elections 1968–1980: A spatial analysis. *American Journal of Political Science* 28: 283–312.

Quinn, K., and A. Martin. 2002. An integrated computational model of multiparty electoral competition. *Statistical Science* 17: 405–419.

Quinn, K., A. Martin, and A. Whitford. 1999. Voter choice in multiparty democracies. *American Journal of Political Science* 43: 1231–1247.

Rakove, J., ed. 1999. *James Madison: Writings*. New York: Library of America.
Riker, W. 1962. *The theory of political coalitions*. New Haven: Yale University Press.
Saari, D. 1997. The generic existence of a core for q-rules. *Economic Theory* 9: 219–260.
Schofield, N. 1978. Instability of simple dynamic games. *Review of Economic Studies* 45: 575–594.
Schofield, N. 1983. Generic instability of majority rule. *Review of Economic Studies* 50: 695–705.
Schofield, N. 1984. Social equilibrium and cycles on compact sets. *Journal of Economic Theory* 33: 59–71.
Schofield, N. 1985. *Social choice and democracy.* Heidelberg: Springer.
Schofield, N. 2005a. A valence model of political competition in Britain: 1992–1997. *Electoral Studies* 24: 347–370.
Schofield, N. 2005b. Local political equilibria. In *Social choice and strategic decisions: Essays in honor of jeffrey S. banks*, ed. D. Austen-Smith and J. Duggan. Heidelberg: Springer.
Schofield, N. 2005c. The intellectual contribution of Condorcet to the founding of the US republic. *Social Choice and Welfare* 25: 303–318.
Schofield, N. 2006. *Architects of political change: Constitutional quandaries and social choice theory.* Cambridge: Cambridge University Press.
Schofield, N. 2007. The mean voter theorem: Necessary and sufficient conditions for convergent equilibrium. *Review of Economic Studies* 74: 965–980.
Schofield, N., A. Martin, K. Quinn, and A. Whitford. 1998a. Multiparty electoral competition in the Netherlands and Germany: A model based on multinomial probit. *Public Choice* 97: 257–293.
Schofield, N., G. Miller, and A. Martin. 2003. Critical elections and political realignment in the U.S.: 1860–2000. *Political Studies* 51: 217–240.
Schofield, N., and I. Sened. 2002. Local Nash equilibrium in multiparty politics. *Annals of Operations Research* 109: 193–211.
Schofield, N., and I. Sened. 2005. Multiparty competition in Israel: 1988–1996. *British Journal of Political Science* 35: 635–663.
Schofield, N., and I. Sened. 2006. *Multiparty government: Elections and legislative politics*. Cambridge: Cambridge University Press.
Schofield, N., I. Sened, and D. Nixon. 1998b. Nash equilibrium in multiparty competition with stochastic voters. *Annals of Operations Research* 84: 3–27.
Stokes, D. 1963. Spatial models and party competition. *American Political Science Review* 57: 368–377.
Stokes, D. 1992. Valence politics. In *Electoral politics*, ed. D. Kavanagh. Oxford: Clarendon Press.
Strnad, J. 1985. The structure of continuous-valued neutral monotonic social functions. *Social Choice and Welfare* 2: 181–195.
Train, K. 2003. *Discrete choice methods for simulation*. Cambridge: Cambridge University Press.

Demographic Transition

Ronald D. Lee

Abstract

The ‘demographic transition’ refers to the fall of fertility and mortality from initially high to subsequent low levels and accompanying changes in the population. It began around 1800 with declining mortality in Europe, and is expected to be complete worldwide by 2100. In that time the global population will have risen tenfold, the ratio of elders to children will have risen by a factor of ten, longevity will have tripled, and fertility fallen from six births per woman to two. Individual and population ageing will pose many challenges, from life-cycle planning to the rising costs of health care and retirement.

Keywords

Ageing populations; Capital–labour ratio; Demographic transition; Dependency ratio; Fertility in developed countries; Fertility in developing countries; Health care; International capital flows; International migration; Life expectantcy; Malthus, T.; Mortality; Nutrition and development

JEL Classifications

J11

The demographic transition is the process whereby fertility and mortality move from initially high levels to subsequent low levels, with accompanying changes in the size, growth rate and age distribution of the population.

Before the start of the demographic transition, life was short, fertility was high, growth was slow, and the population was young. Declining mortality starts the typical transition, followed after a considerable lag by fertility decline (France and

Demographic Transition, Table 1 Global population trends over the transition: estimates, guesstimates and forecasts, 1700–2100

Year	Life expectancy (*y*ears at birth)	Total fertility rate (births per woman)	Pop. size (billions)	Pop. growth rate (%/year)	Pop. < 15 (% total pop.)	Pop > 65 (% total pop)
1700	27	6.0	.68	0.50	36	4
1800	27	6.0	.98	0.51	36	4
1900	30	5.2	1.65	0.56	35	4
1950	47	5.0	2.52	1.80	34	5
2000	65	2.7	6.07	1.22	30	7
2050	74	2.0	9.08	0.33	20	16
2100	81	2.0	9.46	0.04	18	21?

Sources: United Nations estimates and projections, 1900–2100; other sources for earlier years (see Lee 2003, for details)

the United States were important exceptions to this ordering). This pattern of change causes growth rates first to accelerate and then to slow again, as population moves towards low fertility, long life and an old age structure.

The transition began around 1800 with declining mortality in Europe. It has now spread to all parts of the world and is projected to be completed by 2100. This global demographic transition has brought momentous changes, reshaping the economic and demographic life cycles of individuals and restructuring populations. Global population size increased by a factor of 6.5 between 1800 and 2000, and by 2100 will have risen by a factor of ten. There will then be 50 times as many elderly but only five times as many children: the ratio of elders to children will have risen by a factor of ten. The length of life, which has already more than doubled, will have tripled, while births per woman will have dropped from six to two. In 1800, women spent about 70 per cent of their adult years bearing and rearing young children, but that fraction has decreased in many parts of the world to only about 14 per cent due to lower fertility and longer life (Lee 2003). These changes are sketched in Table 1.

Before the Demographic Transition

According to Thomas Malthus (1798), slow population growth in the pre-industrial past was no accident. Faster population growth would depress wages, causing fertility to fall and mortality to rise due to famine, war or disease. Thus, population size was held in equilibrium with the slowly growing economy. The need to establish a separate household at marriage kept mean age at first marriage high, averaging around 25 years for women, and overall fertility low, at four to five births per woman. Mortality was moderately high, with life expectancy between 25 and 35 years. Outside of Europe and its offshoots, fertility and mortality were higher in the pre-transitional period. In India in the late 19th century, life expectancy was in the low twenties, while fertility was six or seven births per woman (Bhat 1989). In Taiwan, the picture was similar around 1900. In the 1950s and 1960s, fertility in the less developed countries (LDCs, see UNPD 2005, for definition) was typically six or higher.

Declining Mortality

The demographic transition began first in northwest Europe, where mortality started its secular decline around 1800. In many low-income countries, the decline in mortality began in the early 20th century and then accelerated dramatically after the Second World War. The first stage of mortality decline is due to reductions in contagious and infectious diseases. Starting with the development of smallpox vaccine in the late 18th century, preventive medicine played a role in mortality decline in Europe. Public health measures were important from the late 19th century, and some quarantine measures may have been effective in earlier centuries. Improved personal hygiene also helped as the germ theory of disease

became more widely known and accepted. Improving nutrition was also important in the early phases of mortality decline. Famine mortality was reduced by improvements in storage and transportation that permitted integration of regional and international food markets. Secular increases in incomes led to improved nutrition in childhood and throughout life. Better-nourished populations with stronger organ systems were better able to resist disease.

Today, the high-income countries have already largely achieved the potential mortality reductions through control of contagious disease and improved nutrition. For them, further reduction in mortality must continue to come from reductions in chronic and degenerative diseases, notably heart disease and cancer (Riley 2001).

Most LDCs did not begin the mortality transition until the 20th century but then made rapid gains. Between 1950–1954 and 2000–2004, life expectancy in LDCs has increased from 41.1 years to 63.4, with average gains of 0.45 years of life per calendar year. Such rapid rates of increase in low-income countries will surely taper off as mortality levels approach those of the more developed countries (MDCs), whose gains have been less than half as rapid at 0.19 years per year. There are notable exceptions to this generally favourable picture. In sub-Saharan Africa, life expectancy has been declining since the early 1980s, largely due to HIV/AIDS. In the southern African region, life expectancy dropped from 62 to 48 between the early 1990s and the early 2000s. On average, eastern European (including the former USSR) life expectancy is lower now than it was in the late 1960s (UNPD 2005).

How far and how fast will mortality fall and life expectancy rise in the 21st century? Methods that extrapolate historical trends in mortality by age suggest greater longevity gains than MDC government actuaries typically project, but past official projections have under-predicted subsequent gains, particularly at the older ages. Some experts argue that we are approaching biological limits and that these historical trends cannot be expected to continue; they foresee an upper limit of 85 years for life expectancy. Others, impressed by advances in genetic and stem cell research, foresee much more rapid gains for the future than occurred in the past.

Fertility Transition

Most economic theories of fertility start with the idea that couples wish to have some number of surviving children rather than a number of births per se. On this assumption, once potential parents recognize an exogenous increase in child survival, fertility should decline. However, mortality and fertility interact in complicated ways. For example, increased survival raises the return on post-birth investments in children, while some of the improvement in child survival is itself a response to parental decisions to invest more in the health and welfare of a smaller number of children. Nonetheless, it is very likely that mortality decline has exerted an important independent influence on fertility decline.

Economic change also influences fertility by altering the costs and benefits of childbearing and rearing, which are time-intensive. Technological progress and increasing physical and human capital make labour more productive, raising the value of time in all activities and making children increasingly costly relative to consumption goods. Since women have had primary responsibility for childbearing and rearing, variations in the productivity of women have been particularly important. For example, physical capital may substitute for human strength, reducing or eliminating the productivity differential between male and female labour, and thus raising the opportunity cost of children. Rising incomes have shifted consumption demand towards non-agricultural goods and services, for which educated labour is a more important input. A rise in the rate of return to education then leads to increased investments in education. Overall, these patterns have several effects: children become more expensive, their economic contributions are diminished by school time, and educated parents have higher value of time, which raises the opportunity costs of childrearing. Furthermore, parents with higher incomes choose to devote more resources to

each child, and, since this raises the cost of each child, it also leads to fewer children. Developing markets and governments replace many economic functions of the traditional family and household, to which children contributed, further weakening the value of children.

The importance of contraceptive technology for fertility decline in the past and future is hotly debated, with many economists viewing it as of relatively little importance (Pritchett 1994). The European fertility transition, for example, was achieved using *coitus interruptus,* a widely known traditional method requiring no modern technology.

Between 1890 and 1920, fertility within marriage began to decline in most European provinces, with a median decline of about 40 per cent from 1870 to 1930 (Coale and Watkins 1986). The fertility transition in the MDCs largely occurred before the Second World War. After the war, many of these countries experienced baby booms and busts, followed by the 'second fertility transition' as fertility fell far below replacement level, marriage rates fell, and increasing proportions of births occurred outside marriage. Many LDCs began the fertility transition in the mid- 1960s, and these later transitions have typically been more rapid than earlier ones, with fertility reaching replacement level (around 2.1 births per woman) within 20 to 30 years after onset. Fertility transitions in East Asia have been particularly early and rapid, while those in South Asia and Latin America have been slower (Bulatao and Casterline 2001). The transition in sub-Saharan Africa started from a higher initial level of fertility and began later. By now, almost all countries have begun the fertility transition (UNPD 2005; Bulatao and Casterline 2001).

Currently, 66 countries with 44 per cent of the world's population have fertility at or below replacement level. Of these, 43 are MDCs, but 23 are LDCs. Average fertility in the MDCs is 1.56 births per woman, and in many it has fallen below 1.3. Many LDCs, particularly in East Asia, also have fertility far below replacement. It is not yet clear whether fertility will fall farther, rebound towards replacement, or stay at current levels.

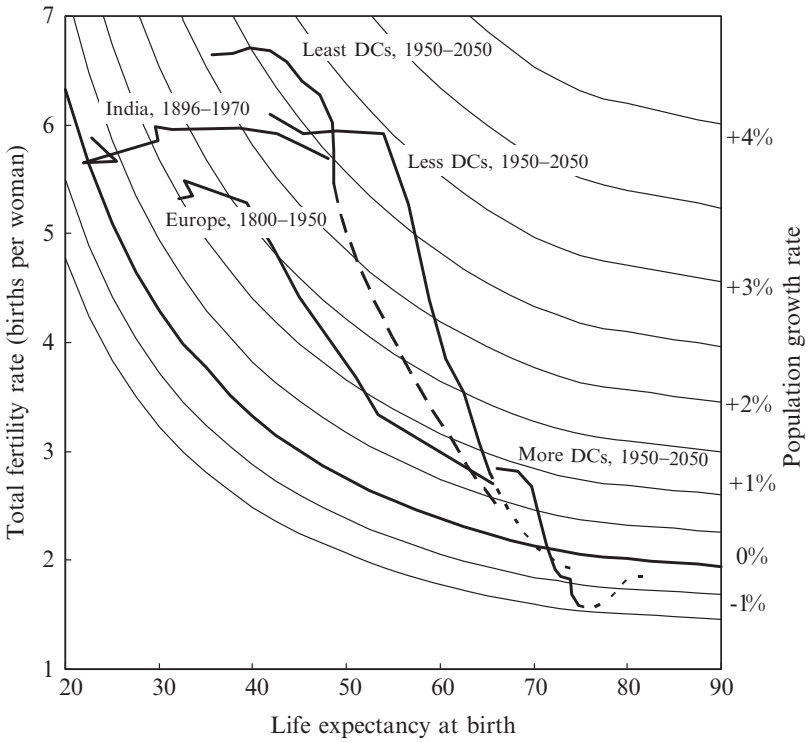
Age at first marriage and first birth are generally moving to older ages throughout the industrial world and much of the developing world as well. This depresses the total fertility rate, which is a synthetic cohort measure, by 10–40 per cent below the underlying completed fertilities of generations. When the average age of childbearing stops rising, the total fertility rate should increase to this underlying level.

Population Growth

A steady state population growth rate for a hypothetical zero migration population can be associated with each level of fertility and life expectancy, as depicted in Fig. 1.

Figure 1 plots growth rate contours or isoquants. These differ from actual growth rates due to net migration and the transitory influence of age distribution. In this figure, a demographic transition begins as a move to the right, representing a gain in life expectancy with little change in fertility and therefore movement to a higher population growth contour. Next, a diagonal downward movement to the right reflects a simultaneous decline in fertility and mortality, recrossing contours towards lower rates of growth and perhaps going negative, as do the MDCs. Historical data are extended using UNPD (2005) data and projections, by development status.

India, shown separately, had higher initial fertility and mortality than Europe, as did the least developed countries relative to the LDCs in 1950, which in turn had far higher mortality and fertility than the MDCs in that year. In all cases, the initial path is horizontally to the right, indicating that mortality decline preceded fertility decline, causing accelerating population growth approaching three per cent for the LDCs and least developed countries. Europe briefly attains 1.5 per cent steady state population growth, but then fertility plunges, a decline picked up after 1950 by the



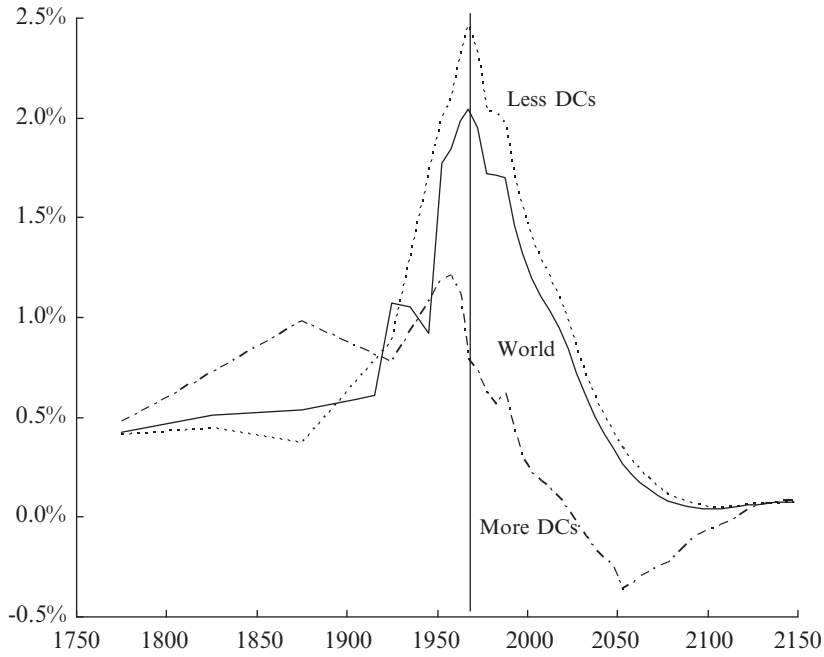
Demographic Transition, Fig. 1 Life expectancy and total fertility rate with population growth isoquants: past and projected trajectories for more, less, and least developed countries. (Source: Bhat (1989); UNPD (2005); see Lee (2003) for further details)

group of LDCs, ending with population decline at 1 per cent annually (the actual European population growth rate is slightly higher than this hypothetical steady state one due to age distribution and immigration). All three groups are projected by the UN to approach the zero-growth contour by 2050.

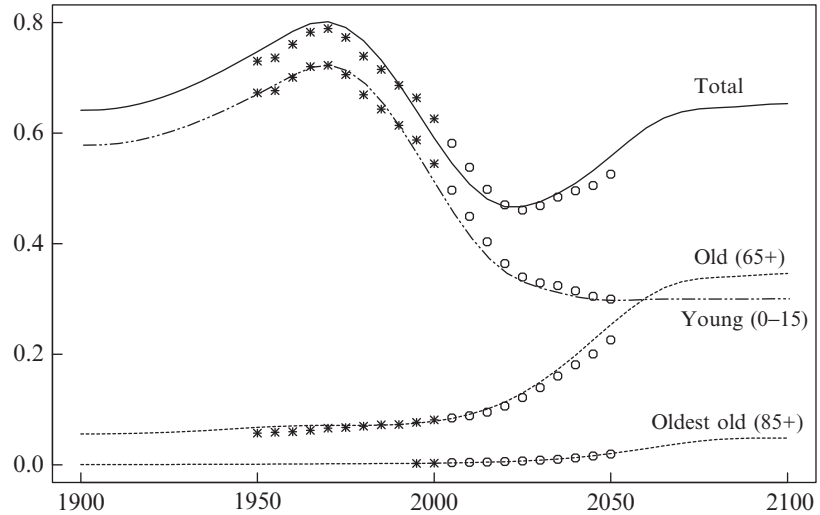
Historical and projected population growth rates, as opposed to hypothetical steady state ones, can be seen over a longer time period in Fig. 2. Growth rates in the MDCs rose about a half of one per cent above those in the LDCs in the century before 1950. But after the Second World War, population growth surged in the LDCs, with the growth rate peaking at 2.5 per cent in the mid-1960s, then dropping rapidly. The population share of the MDCs is projected to drop from its current 20 per cent to only 13 per cent in 2050. Long-term United Nations projections suggest that global population growth will be close to zero by about 2100. The projection for the MDC population is nearly flat, with population decrease in Europe and Japan offset by population increase in the United States and other areas.

Changing Age Distribution over the Demographic Transition

Figure 3 plots the changes in age distribution that accompany a classic demographic transition, using historical data from India from 1896 to 2000 (stars) and United Nations projections through 2050 (hollow circles). These data are superimposed on a stylized transition that was simulated with the use of mathematical functions for the trajectories of fertility and life expectancy. Simulated fertility starts close to six births per woman and ends at 2.1. Life expectancy starts at 24 years and ends at 80. Mortality decline starts in 1900, 50 years before the fertility decline begins in 1950. The Indian fertility transition is slower than that of East Asia but similar to that of Latin America.



Demographic Transition, Fig. 2 Population growth rates, 1750–2150. (Source: UNPD (2005); see Lee (2003) for further details)



Demographic Transition, Fig. 3 Changing age distribution over a classic demographic transition: actual and projected dependency ratios for India and simulations, 1900–2100. (Source: Actual India data for the period 1891–1901 to 1941–51 are taken from Bhat (1989). Other actual and projected data are taken from UNPD (2003). Note: Lines indicate a simulated demographic transition superimposed over actual (*) and projected (o) dependency ratios for India)

The distinctive changes in the age distribution can be seen in the 'dependency ratios', which take either the younger or the older population and divide it by the working-age population. The initial mortality decline, while fertility remains high, raises the proportion of surviving children in the

population, as reflected in the increasing child dependency ratios. Counter-intuitively, mortality decline initially makes populations younger rather than older in a phase which here lasts 70 years. Families find themselves with increasing numbers of surviving children, and both families and governments may struggle to achieve human capital investment goals for the unexpectedly high number of children.

Next, as fertility declines, child dependency ratios decline and soon fall below their pre-transition levels. The working age population grows faster than the population as a whole, so the total dependency ratio declines. This second phase may last 40 or 50 years. Some analysts have worried that the rapidly growing labour force in this phase might cause rising unemployment and falling capital–labour ratios, while others have stressed the advantages of this phase, calling these a demographic gift or bonus. Figure 3 shows that in India the bonus occurs between 1970 and 2015, when the total dependency rate is declining. The decline in dependents per worker would by itself raise per capita income by 22 per cent, other things equal, adding 0.5 per cent per year to per capita income growth over the 45-year span.

In a third phase, increasing longevity leads to a rapid increase in the elderly population while low fertility slows the growth of the working age population. The old-age dependency ratio rises rapidly, as does the total dependency ratio. In India, population ageing will occur between 2015 and 2060. If the elderly are supported by transfers, either from their adult children or a public-sector pension system supported by current tax revenues, then the higher total dependency ratio means a greater burden on the working-age population. However, to the extent that the elderly prepare for their retirement by saving and accumulating assets earlier in their lives and then dissave in retirement, population ageing may cause lower aggregate saving rates, as life-cycle savings models and some empirical analyses suggest. But even with lower savings rates the capital–labour ratio may rise, since the labour force is growing more slowly. The net effect would then be to stimulate growth in labour productivity due to capital deepening.

At the end of the full transitional process, the total dependency ratio is back near its level before the transition began, but now child dependency is low and old-age dependency is high. Presumably mortality will continue to decline in the 21st century, and the process of individual and population ageing will continue. No country in the world has yet completed this phase of population ageing, since even the industrial countries are projected to age rapidly over the next three or four decades. In this sense, no country has yet completed its demographic transition.

Population ageing is due both to low fertility and to long life. Low fertility raises the ratio of elderly to working-age people, with no corresponding improvement in health to facilitate a prolongation of working years. For this reason, it imposes important resource costs on the population, regardless of institutional arrangements for old-age support. Lower total expenditures on children and increased capital per worker will offset these costs.

By contrast, population ageing due to declining mortality is generally associated with increasing health and vitality of the elderly. Such ageing may put pressure on pension programmes that have rigid retirement ages, but this problem is a curable institutional one, since the ratio of healthy, vigorous years over the life cycle to frail or disabled years has not changed, and individuals can adjust by keeping the fractions of their adult life spent working and retired constant, for example.

Some Consequences of the Demographic Transition

The three centuries of demographic transition from 1800 to 2100 will reshape the world's population in a number of ways. Population will rise from 1 billion in 1800 to 9.5 billion in 2100. The average length of life will increase by a factor of two or three, fertility will have declined by two-thirds, and the median age of the population will double from the low twenties to the low forties. The population of Europe will decline by ten per cent between 2005 and 2050, and its share of world population will have declined by two-thirds

since 1950. But many other changes will also have been set in train in family structure, health, institutions for saving and supporting retirement, and even in international flows of people and capital.

At the level of families, as the number of children born declines sharply, childbearing becomes concentrated into only a few years of a woman's life; combined with greater longevity, this means that many more adult years are available for other activities. Parents with fewer children are able to invest more in each child, reflecting the quality–quantity trade-off, which may also be one of the reasons parents reduced their fertility.

The processes which lead to longer life also alter the health status of the surviving population. For the United States, it appears that years of healthy life are growing roughly as fast as total life expectancy. In other industrial populations the story is more mixed. Trends in health, vitality and disability are of enormous importance for human welfare.

The economic pressures on pension programmes caused by the increasing proportion of elderly are exacerbated in the MDCs by dramatic declines in the age at retirement, which for US men fell from 74 in 1910 to 63 in 2000. Population ageing will also generate intense financial pressures on publicly funded systems for health care and for long-term care.

At the international level, the flow of people and capital across borders may offset these demographic pressures. As population growth has slowed or even turned negative in the MDCs, it is not surprising that international migration from Third World countries has accelerated. Net international migration to the MDCs experienced a roughly linear increase from near-zero in the early 1950s to around 2.6 million per year in the 1990s. Of course, these net numbers for large population aggregates conceal a great deal of offsetting international gross migration flows within and between regions (UNPD 2005). For example, prior to 1970 Europe was a net sending region, but from 1970 to 2000 it received 18 million net immigrants. During the 1990s, repatriation of African refugees reversed the net flows from the least developed countries. But overall, while MDCs may seek to alleviate their population ageing through immigration, United Nations simulations indicate that the effect will be only modest, since immigrants also grow old, and their fertility converges to levels in the receiving country.

Might international flows of capital cushion the financial effects of population ageing? Population ageing may cause declining aggregate saving rates, but, with slowing labour force growth, capital–labour ratios will probably rise and profit rates fall, particularly if there is a move towards funded pensions. Capital flows from the MDCs into the LDCs might help keep the rate of return on investments from falling, but the possibilities are limited by the much smaller size of Third World economies.

Dramatic population ageing is the inevitable final stage of the global demographic transition, and it will bring serious economic and political challenges. Meeting these challenges will require flexible institutional structures, adjustments in life-cycle planning, and a willingness to pay for rising costs of health care and retirement.

See Also

- ▶ Fertility in Developed Countries
- ▶ Historical Demography
- ▶ International Migration
- ▶ Mortality
- ▶ Population Ageing

Bibliography

Bhat, P. 1989. Mortality and fertility in India, 1881–1961: A reassessment. In *India's historical demography: Studies in famine, disease and society*, ed. T. Dyson. London: Curzon Press.

Bulatao, R. and Casterline, J., eds. 2001. *Global fertility transition*, a supplement to *population and development review* 27.

Coale, A., and S. Watkins, eds. 1986. *The decline of fertility in Europe*. Princeton: Princeton University Press.

Lee, R. 2003. The demographic transition: Three centuries of fundamental change. *Journal of Economic Perspectives* 17(4): 167–190.

Malthus, T. 1798. *An essay on the principle of population*, ed. D. Winch and P. James. Cambridge: Cambridge University Press, 1992.

Pritchett, L. 1994. Desired fertility and the impact of population policies. *Population and Development Review* 20: 1–55.
Riley, J. 2001. *Rising life expectancy: A global history*. Cambridge: Cambridge University Press.
UNPD (United Nations Population Division). 2003. *World population prospects: The 2002 revision*. New York: United Nations.
UNPD. 2005. *World population prospects: The 2004 revision*. New York: United Nations.

Demography

Nathan Keyfitz

Demography is the analysis of population, including both techniques and substance. It is applied most often to human populations, and includes the gathering of data, the construction of models, interpretation of population changes, policy recommendations. The *data* used by demographers are partly cross-sectional in the form of censuses and sample surveys, partly flow data consisting of time series of births and deaths. *Models* that express the relation between the flow series of births, deaths and migration on the one side and the cross sections on the other are a main tradition of demography, running through the work of Lotka, Leslie and many others. *Interpretation* includes tracing causes of changes, and assessing their future consequences. *Policy* recommendations aim at lowering birth rates in countries of rapid growth, and raising it in countries below replacement.

Demography on the whole belongs to social science, though part of it (some of the analysis of mortality, was well as questions of fecundity) falls within the field of biology. It draws from and overlaps with other social sciences, especially sociology and economics. Reliability engineers deal with the life and demise of equipment and face problems analogous to those of human mortality; the mathematics they use is in many respects the same as that of demography, with superficial differences of notation. Epidemiology deals with some of the same problems as demography, though it too has developed a different tradition of exposition and notation. In so far as demographers collect and interpret data they necessarily borrow the techniques of statistics, including probability and stochastic processes. Ecology, a branch of biology, makes use of demographic techniques and results (Sauvy 1954; Scudo 1984).

For the more numerically minded demographer the subject begins with John Graunt (1662), who published his *Observations on the Bills of Mortality* more than three centuries ago. Yet Graunt's close study of the primitive death certificates of his day is not often referred to by working demographers now More often mentioned as a predecessor is Lotka, who applied the renewal equation, developed in mathematical physics about the beginning of the century, to the renewal of a human population. The part of his long career, with publications dating all the way from 1907 to 1948, that is most remembered was devoted to developing the consequences of that one equation. Those who see demography as emphasizing forecasting are likely to think of Cannan (1895), Bowley (1924), and Whelpton (1936), whose components method was put into convenient matrix form by Leslie (1945).

Data

The most fundamental of all demographic data is the Census. Census taking is by no means novel. Ten cases of enumeration of the whole people (the earliest under Moses (Exodus xxxviii) and the last under Ezra (Ezra ii, 64)) are reported in the Old Testament, and one very famous occasion by the Romans is reported in the New Testament (Luke ii, 2). For a time the Romans took a census every five years. Classical Chinese literature contains innumerable references to counts in one part of the country or another. Premodern censuses were taken primarily to establish obligations on payment of taxes and military service, and they were correspondingly subject to evasion.

Modern censuses have been associated with the national state, as were other kinds of statistics: the word statistics itself itself reminds us of the

association. Among the early acts of the revolutionary government of France was legislation providing for collection of data, including the taking of censuses. This was anticipated by Sweden, whose series of censuses goes back to the 18th century. Depending on the definition, the first census of modern times was taken in Sweden, Canada, or Virginia.

The association of the census with the national state has been seen in many of the new countries established after World War II. Countries seized on censuses to legitimate their nationhood, just as did France two centuries ago.

What characterizes modern censuses is (a) that they take place periodically, (b) that the enumeration is name-by-name, (c) that they seek to include all the persons belonging in a given area, (d) that they ask questions on age, sex, activity, etc., some of the questions often being on a sample, (e) that they recognize the problem of error and omission.

Geographic preparation is a major part of the effort to attain accuracy and completeness. The country is divided into enumeration areas on maps, with boundaries indentifiable on the ground, and each such area is assigned to an enumerator to be held responsible for its coverage. This principle of a division, first on maps and then on the ground, into an exhaustive set of non- overlapping areas is the essential principle of censustaking. It was apparently Morris H. Hansen who first applied the fact that every such area need not be covered for surveys (for example, population surveys taken between censuses). In area sampling the identification of individuals with a point on the map constitutes an implicit listing; the sample is specified in such a way that all individuals, including those unknown to the sample designers, have a prescribed chance of inclusion.

Equally valued with censuses for demographic calculations, though much less widely available, are accurate vital statistics. Partial records of births and deaths are to be found in many places and in many historical epochs, but effectively complete registration was largely a 19th-century innovation; the Swedish series going back to the 1700s is virtually unique.

Only under modern conditions do citizens need passports and other identification that depend on birth registration, and the citizen co-operation that is a condition for good vital statistics comes only with modernization. Censuses have now been taken in most countries of the world, but accurate vital statistics, covering current births and deaths, are to be had for countries including no more than about 30 per cent of the world's population. If we had to wait for the general awakening of public statistical consciousness that is required for a complete vital statistics system the population problem of the world would be solved before it could be measured.

Comparison

One of the oldest demographic problems is the simple comparison of mortality level as between two populations, or one population between two points of time. US advances in longevity were slow and uncertain in the 1950s and 1960s; it is a statistically delicate question whether mortality was lower in the United States in 1980 than it was in 1950 and by how much. A first attempt to answer it is comparison of crude rates, and we find that for white females the crude rate, deaths D divided by population P, was the same in both years. But this is not a pure comparison of mortality. If the populations number p_x^1 and p_x^2 at age x, and their death rates are μ_x^1 and μ_x^2, then the comparison of crude rates is D^1/P^1 versus D^2/P^2 or

$$\frac{\sum p_x^1 \mu_x^1}{\sum p_x^1} \text{ versus } \frac{\sum p_x^2 \mu_x^2}{\sum p_x^2}$$

whose sole advantage as an index is that it may be calculated from the number of deaths and the number of exposed population at each of the two times, without any breakdown of the data by age. The p_x^1 and the p_x^2 confound the comparison, and if they are systematically different then the comparison of crude rates tells little about relative mortality. In particular one population having a larger proportion of old people than the other badly distorts the comparison.

To meet this difficulty, basic information was collected by age as far back as the 18th century in Sweden. To eliminate the different age weighting of the two populations from the comparison, it is common to use the directly standardized index with fixed p_x^1.

$$\frac{\sum p_x^1 \mu_x^2}{\sum p_x^1 \mu_x^1},$$

whose analogue in economics is the base-weighted aggregative price index. (The μs are similar to prices, and the ps to quantities used.) This formula gives for white females 6.5 in 1950 and 4.1 in 1980, a major difference from the crude rates, that were unchanged. Other formulas, for instance that obtained by replacing p_x^1 by p_x^2, give different answers, and the choice among them is difficult to make on logical grounds. Thus the famous price index number problem carries over to demographic comparison, though not the difficulty that rising or falling prices by themselves affect the amounts purchased. (Kitagawa and Hauser 1973).

Demography has a resource not available to the study of price changes: the life table model. If the death rates of this year, including all ages at which anyone is living, can be seen as the successive ages in the life of an individual, then the individual subject to those rates would have a certain expectation of life. No real person will have such an expectation, but the model provides what is the most common means of interpreting a current pattern of mortality.

If $\mu(x)$ is the age-specific death rate at age μ to $\mu + \mathrm{d}x$ then the chance of a baby living to age α is $l(\alpha) = \exp\left[-\int_0^a \mu(x)\mathrm{d}x\right]$, this being the solution of the differential equation defining the death rate,

$$\mu_x = \frac{1}{l(x)} \frac{\mathrm{d}l(x)}{\mathrm{d}x}.$$

The expectation of life at age x is then

$$e_x^0 = \frac{\int_x^\omega l(a)\mathrm{d}a}{l_x},$$

where ω is the highest age to which anyone lives. US white females showed e_x^0 equal to 72 years in 1950, 79 years in 1980. Elandt-Johnson et al. (1980) apply the expectations comparisons in clinical follow-up studies.

Mortality and its Changes

To classify mortality according to the single parameter of life expectancy captures a good part of the variation in age incidence from one population to another, but not all. For instance a population may have high infant mortality and low mortality in later life while, in another, mortality may be low for infants and high in later life, with the two populations having the same overall expectations. Two dimensions differentiate among patterns much better than one. Coale and Demeny (1983) show four families of model tables. The United Nations (1985) show a Latin American, a Chilean, a South Asian, and a Far Eastern pattern. A particularly effective set of tables is due to William Brass (1971), who regresses the l_x column of a given table on that of a standard table, after both have been transformed by logits, and the regression of the one on the other turns out to be close to a straight line. Given the standard table, Brass's is a two constant system.

As mortality improves along the path that we have seen in advanced countries over the past generation the age specific rates at all ages go down, most being reduced by half in each generation. Because the span of life has changed little, a given per cent fall in age specific rates now has a much smaller effect on life expectancy than an equal percentage fall 50 years ago. In fact a drop of 1 per cent now in all age specific rates causes a rise of only about 0.10 to 0.15 per cent in life expectancy; 50 years ago it caused a rise of 0.30 per cent. This number, the derivative of the life expectancy with respect to the age specific rates, has been called *H:*

$$H = \frac{\int_x^\omega l(x)\ln[l(x)]\,\mathrm{d}x}{\int_x^\omega l(x)\mathrm{d}x}.$$

On the present course it is becoming smaller and smaller, as we proceed to a time when everyone dies at about the same age. Demetrius (1974) has carried this analysis further.

Note that the progress against mortality need not go this route. We can imagine a slowing of the ageing process by which the l_x curve moves out to the right, rather than merely moving up to a horizontal line with a fixed boundary on the right. A slowing of the ageing process by 50 per cent would mean an extension of average life not of $50H$ per cent, or about 7 years, but a full doubling of life expectancy. One of the questions that physicians, pension officials and demographers ask one another is which of the two courses will be taken in the future by mortality improvements, especially at the oldest ages which count more and more for this as mortality under age 70 becomes small.

The life table with one exit–death–can be extended to several exits, representing the several causes of death, and on from these to several increments, taking place not only at age zero, but at arbitrary ages.

Fertility Measures

Children are born to women only at a restricted range of ages, so comparison for births are a somewhat different problem than for deaths. If we divide the number of births B by the whole population P to obtain a crude birth rate then we are subject to the irrelevant variation of the young and old people in the denominator; it is better to divide by the number of women in the childbearing ages. Some further small gain in precision of comparison is obtained by working with age-specific rates, the births ${}_5B_{15}$ to women 15–19 years at last birthday divided by the number of women ${}_5P_{15}$ in the population of that age at mid-period, and similarly for the six other ages under 50. With single years of age, if B_x is average girl births during a year to women aged x, then the rates are $f_x = B_x/P_x$, and these over the childbearing ages may be added to obtain the Gross Reproduction Rate (GRR):

$$\mathrm{GRR} = \sum f_x = \sum (B_x/P_x)$$

including boy and girl births in the numerator B_x gives the total fertility rate (TFR), approximately double the GRR.

For measuring the natural increase of a population survivorship l_x is incorporated in the formula to give the net reproduction ratio (NRR), $R_0 = \Sigma l_x f_x$, where now f_x is again the girl birth rate. R_0 is the number of girl children expected to be born to a girl child on a particular set of rates of birth and death. By virtue of that definition it is the ratio of the number of persons in one generation to the number in the preceding, taken in abstraction from any irregularities in the age distribution, and disregarding the length of time over which one generation is replaced by another.

Estimating the effect of abortion and contraception raises some further issues. Since one abortion of a conception leading to a live birth reduces the number of live births in the population by 1, it might be thought that 1000 abortions would reduce the number of births by 1000, but this is not so. If the probability of a conception that leads to a live birth in a given month is p, and the sterile period following conception is s months, then there will be a birth on the average every $(1/p) + s$ months. If the sterile period following conception when abortion occurs is α, then there will be an abortion on the average every $(1/p) + \alpha$ months. Hence the number of abortions that avoid one birth is

$$\frac{\frac{1}{p} + s}{\frac{1}{p} + \alpha}.$$

This can come out above 2 if no contraception is used, but is only slightly over 1 if the abortion is a backstop to more or less efficient contraception (Potter 1972).

Momentum

With an NRR equal to unity a population will just replace itself over the long run; population in this

condition of bare replacement will ultimately become stationary. If it drops to bare replacement after a history of rapid increase, then because of its young age distribution, with many women in the childbearing ages, it will continue to increase for one or two generations, until it attains a number that may be as much as 70 per cent higher than when its NRR dropped to unity, a phenomenon called population momentum. If the population has been increasing uniformly over a considerable period of time the ratio of the ultimate population to that at the onset of bare replacement is simply expressed as

$$\text{Ratio} = \left(\frac{b}{r}\right)\frac{e_0^0}{\mu}\left(\frac{R_0 - 1}{R_0}\right),$$

where b is the birth rate, r the rate of natural increase, μ the mean age in the stationary population (Keyfitz 1985, p. 156).

This result is exact under the assumptions stated, and is one of numerous inferences from stable population theory.

Pension Cost as a Function of the Rate of Increase

Stable population theory also tells us the relation between certain variables when other circumstances are held constant. A pension of unity to all members of the population over age 65 will cost those aged 20 to 64 at last birthday the annual premium

$$p(r) = \frac{\int_{65}^{\omega} e^{-rx} l(x)\mathrm{d}x}{\int_{20}^{65} e^{-rx} l(x)\mathrm{d}x},$$

and this cost can be approximated as

$$p(r) = p_0 \exp\left[r(m_1 - m_2) - \frac{r^2}{2}\left(\sigma_1^2 - \sigma_2^2\right)\right]$$

where m_1 and m_2 are the mean ages of the 20–64 and the 65 and over respectively, and σ_1^2 and σ_2^2 their variances. Since $m_1 < m_2$ and the term in r^2 is small, the premium is necessarily a decreasing function of the rate of increase of the population (Keyfitz 1985, p. 106).

Kinship

If the population can be assumed to be stable and some assumptions of continuity are made then kin relations become determinate. Knowing the age specific rates of birth and death, and supposing the various demographic events are independent, we can find exact expressions for the probability that a person aged α has a living mother, living grandmother, as well as the expected aunts, cousins etc. (Goodman et al. 1974).

Lotka (1931) gives the probability that a girl aged α has a living mother. His answer is obtained in two steps: (1) with the condition that at the girl's birth the mother was x years old the probability is simple: $l_{x+\alpha}/l_x$; (2) removal of the condition by averaging over all ages of mothers at childbearing gives, on the stable assumption:

$$M_1(\alpha)\int_{\alpha}^{\beta} \frac{l_{x+\alpha}}{l_x} e^{-rx} l(x) f(x) dx.$$

From this it follows that the probability of a living grandmother is

$$M_2(\alpha)\int_{\alpha}^{\beta} M_1(x+\alpha) e^{-rx} l(x) f(x)\mathrm{d}x,$$

and so on. Other expressions are obtainable for sisters, aunts, cousins (Le Bras 1973). Noreen Goldman (1978) has applied the formulas for younger sisters and older sisters, equating the ratio as given in theory to the ratio observed in a sample of a population, and solving for the intrinsic rate. Her method for finding the rate of increase has the advantage of requiring no knowledge of age on the part of respondents.

Notice that the preceding formulas, like others based on stable theory, are essentially comparative statics, and give a result very different in meaning from one based on observed age data.

They answer questions like 'What happens to the premium for old age pensions in the stable condition with the given parameters?' The formula for $M_1(\alpha)$ gives the fraction of girls aged α who have a living mother given the life table and birth rates, and disregarding all else. The observed fraction of girls aged α who have a living mother takes account of all other elements affecting the real population.

Inferring Vital Rates by Indirect Methods

In the absence of complete vital statistics much effort has had to be devoted to inferring vital rates from censuses, and one early method was based on the stable age distribution. In a fast growing population the preponderance of numbers is shifted to the younger ages, and this fact makes it possible to infer the rate of growth from examination of the age distribution. If birth rates and death rates are constant and the population closed, then as we saw the number of persons aged x per current birth is $e^{-rx}l_x$. If the l_x can be taken from a reference or model table, and a census gives c_x persons at age x and c_y persons at age $y > x$, then the equation

$$\frac{c_x}{c_y} = \frac{e^{-rx}l_x}{e^{-ry}{}_y}$$

can be solved to find

$$r = \frac{1}{y-x}\ln\left(\frac{c_x/l_x}{c_y/l_y}\right)$$

(Bourgeois-Pichat 1966).

The matter is not that simple in practice, since growth is irregular, censuses are subject to error, and one does not know what life table to apply. In general any pair of ages combined with a life table gives an estimate, and one can try to use ages that are less vulnerable to reporting error. The theory is readily extended to populations in which mortality is falling (Coale 1963). More recently methods have been developed that do not depend on the assumption of stability (Brass 1975; Preston and Coale 1982; Coale 1984; United Nations 1985).

Periods and Cohorts

Demography moves back and forth between consideration of a population existing at a given moment or period of time, and a cohort that is a group of individuals followed from birth or some other event. Comparison of mortality can be made between periods or between cohorts. The same formulas apply to both, for standardization as well as the life table. In fact, the usual life table is referred to as a synthetic cohort: it treats a set of age-specific rates referring to a particular moment as though they were applicable to individuals and extended over time. Cohorts are in a sense more real than periods, but being only calculations after the last individual member has died, they can never be up-to-date (Ryder 1964).

The cohort – a number of individuals observed from a given starting point – is a demographic unit appropriate to fields other than mortality; one can assemble death and divorce statistics from individual data by following the marriages occurring in a particular year to the time where the couple divorces or one member of the couple dies (Henry 1957a, b; Pressat 1961).

Multi-Dimensional Demography

The above questions and techniques have been largely concerned with counts of people, and in disregard of characteristics other than age and sex. But for many purposes we need to examine marital status, or labour force status, or place of residence within a country. We need to take account of the transitions of individuals, for instance between the states of married and single, between school and labour force, etc. Combinations of sequences are numerous in any of these matters, and in order to bring the number down the Markov assumption is usually introduced, whereby the probability of a person moving into the several states in each period depends only on the last previous state the person was in, and not at all on the path by which he or she arrived at that state.

It fortunately happens that the ordinary life table can be extended to the multi-dimensional case, with matrix analogues for the most common

formulas. If $\mu_{ij}(x)$ is the *rate* at which people aged x are moving from the jth to the ith state, then the *probability* of going from the jth state at the beginning of a period to the ith state at the end of the period, is the ijth element of $\mathbf{P}_x$, where $\mathbf{M}_x$, is the matrix of the μs

$$\mathbf{P}_x = (1 + \mathbf{M}_x/2)^{-1} (1 - \mathbf{M}_x/2)$$

and so on through all the usual life table formulas (Rogers 1975). This way of handling the arithmetic has the convenience of simple formulas, easily implemented on a computer. An equivalent method that dispenses with matrices is due to Robert Schoen (1975) and Leo A. Goodman (1961, 1969).

Mixtures and Heterogeneity

Everything said so far supposes that the several members of the population in any one category have the same probabilities – of dying, of giving birth, or of migrating – an assumption that cannot be correct. The usual demographic models recognize age, sex, and a few other sources of variation among individuals; they make no allowance for statistically unobserved heterogeneity.

Yet we know that some people are in vigorous condition, while others of the same age, sex, etc. are moribund. Among a group of individuals who are not all in the same condition the less vigorous die sooner, leaving the remainder with more favourable mortality than an unselected group would average. This process goes on through life, and the observed death rates, arising as they do from a population selected by differential mortality towards the more robust, are too low to represent an individual who at the start is of average frailty.

If we each had a mark on us indicating our degree of frailty then in estimating our own chances of survival we would use the experience of a group with the same mark as ourselves. We could avoid the unsatisfactory procedure of applying to ourselves the experience of a collection of people among whom average robustness was steadily increasing. Not knowing our condition, we must choose one of two ways of expressing our ignorance and deriving a probability. We can take ourselves as average at the start, and then we must accept that we will have an expectation lower than the published tables show; or else we can take ourselves as the average of the surviving population throughout the whole course, in which case we are supposing that we as individuals are steadily improving in robustness (Vaupel and Yashin 1985).

The recognition of heterogeneity can explain some of the crossovers that are otherwise puzzling, for instance the fact that in the United States blacks show higher mortality than whites at ages up to 70, and beyond that they have lower mortality. Selection by the higher mortality at the younger ages is a way of explaining this; another explanation is defective data.

The curious paradoxes that arise through mixed distributions have been explored by reliability engineers (Mann et al. 1974). In application to demography, the familiar rise in the proportion of divorces with duration of marriage, reaching a peak at five to ten years, could be due to married couples being of two kinds – one group that has a low and constant probability of divorce, not changing with duration of marriage, and another group that has a steadily rising probability with duration of marriage. First the overall rate, following this latter group, rises, but as these divorce and so drop out of the exposed population the rate falls towards that of the lower group. Neither of the component groups has a peak in rates at any time, yet the mix shows such a peak and subsequent fall because those more prone are eliminated from the exposed population.

The point is particularly important in respect of pregnancy. If we follow a group of fertile women through time, and note when they become pregnant we have the same problem of a changing mix, as those that are more fertile drop out, leaving less and less fertile ones behind. That may be a matter of fecundity, the biological ability to have a child, or it may be skill in using birth control, and both of these vary among women (Potter 1972; Potter and Parker 1964). It was Gini (1924) who showed that only in the first month can the rate refer to an unselected group. Goodman (1961)

D

provides methods for the corresponding problem in migration, that had earlier been introduced by Blumen et al. (1955).

The order of magnitude of the effect can be very large in respect of susceptibility to pregnancy, or in respect of divorce; for mortality it cannot be so large because the event in question can only occur once to each member of the population. If a population were divided into three groups, one with an expectation of life of 65 years, one with 73 years, and one with 80 years, then the expected lifetime for the mixed population would be about one year greater than the expected lifetime of the middle group, that we take as the prospect for an individual who is initially of average frailty. About the only general statement that can be made is that expectation as given in published life tables is anything up to one year higher than the initially average person can expect to live.

Forecasting

The activity of demographers that is most often noticed by the public is forecasting: estimating the future population of a country or other area (Brass 1974). The forecasting problem is essentially unsolvable, just as is extrapolating from previous stops to estimate where the wheel will next stop in a casino. There is somewhat more continuity in the demographic than in the casino serials, but to know in advance the major turning points, especially in births, is at least for the present impossible.

While the public may think of demography as principally concerned with the forecasting of population, yet the literature of demography does not give a great deal of attention to this subject, and the best-known demographers have in recent years turned their attention to other problems; explaining the past is providing difficult enough, and until one can say why past events have occurred there is not much prospect of foretelling future ones.

Demographic forecasts are bound to be subject to especially large error for two reasons: they concern the long-term future, and they are self-contained within the narrow set of demographic variables. Forecasting a year ahead would be extremely useful in regard to the unemployment rate or housing construction, not to mention the stock market, while for a year ahead the population is so close to that of today that the forecast is of no interest. Demographic forecasts are typically for 10, 25, and more years into the future.

Since what the population will do depends on many variables outside of demography, it has often been suggested that demographers take into account these non-demographic variables. But that would require knowing future attitudes towards work and the family, and other matters more resistant to forecasting than population itself. Beyond that problem, even if we knew all of these independent variables for the next 25 years, the nature of the functional relation between them and population is beyond present knowledge.

During the present century death rates have been decreasing in most parts of the world, and extrapolations have been moderately successful. The increase in life expectancy has typically been almost three years per decade in developed countries, and has often reached five years per decade elsewhere.

What affects forecasts most is the birth rates assumed, and here is where the biggest failures have been. There was no way to forecast the postwar rise in births shown by developed countries, and equally little understood is the decline of births in the 1960s, and why birth rates continue to be so low. It was during the prosperity of the 1960s that the birth rates started to fall, and during the depressed late 1970s and 1980s that they fall even lower, so we do not know whether births depend directly or inversely on income. A theory that has strong logic on its side, that of Richard Easterlin (1980), by which the small cohort finds itself prosperous and produces a large cohort in its turn has not so far seemed precise enough either in timing or in quantity of the effect to be used by practising forecasters.

Migration is even more difficult for those few countries in which it is substantial. We do not know the amount of immigration into the United

States now, let alone the amount that will occur during the 21st century.

Once the future mortality, fertility and migration are assumed, the forecast is easily made. In the usual projection by age and sex one starts with females, sets up a vertical vector $\mathbf{P}_0$ consisting of the numbers at each age, premultiplies that vector by a matrix whose first row is the age-specific fertility rates for girl children, and whose subdiagonal is the survivorship rates. if $\mathbf{M}$ is the matrix with fertility rates m_{1j} in its first row, and surviviorships $m_{j+1,\ j}$, $j = 1,\ldots,\ n\ -\ 1$, in the sub-diagonal, then the age vector at time 1 is

$$\mathbf{P}_1\ = \mathbf{M}\mathbf{P}_0.$$

and at time t is

$$\mathbf{P}_t = \mathbf{M}^t\ \mathbf{P}_0,$$

if the rates are assumed constant over time (Leslie 1945), If the rates change, the matrix being $\mathbf{M}_1$ in the first period, $\mathbf{M}_2$ in the second period, then

$$\mathbf{P}_t\ = \mathbf{M}_t,\ \ldots,\ \mathbf{M}_2\mathbf{M}_1\mathbf{P}_0.$$

The assumed migrants would be added in each time period.

Experiments have shown that extrapolating birth and death rates does better, though not by much, than supposing that birth and death rates will continue unchanged at their level at the jumping-off point.

Even simpler than projecting with fixed birth rates is using fixed absolute numbers of births into the future. This method, that might be called instant stationarity, also gives results not much inferior to the usual assumption of changing future rates. A rationale for the fixed absolute numbers is provided by the Easterlin hypothesis, by which birth rates are higher for small parental cohorts.

Forecasting Error

Badly needed are probability methods. Some have been proposed (e.g. Pollard 1966) for *ex ante* computation of error, but so far these have had little influence on forecasting practice.

Ex post the problem is simpler. The assessment of earlier projections, leading to an estimate of the intrinsic error of the process, demands first of all a metric that will be comparable between different points of time for a given population, and between large and small populations, growing and declining populations, long- and short-range projections. Such a metric has been found to be the difference between the forecast rate of growth of the population in question and the (subsequently known) realized rate:

$$\text{Error} = \sqrt{\sum\left[\left(\frac{\widehat{p}_t}{\widehat{p}_0}\right)^{1/t} - \left(\frac{p_t}{p_0}\right)^{1/t}\right]^2}$$

where $\widehat{p}_t$ is the forecast population at time t, p_t the realized population, t being the time interval between when the forecast was made and the date to which the projection applies. For some 300 forecasts applying to 15 developed countries, error as so measured turns out to be about 0.003, or 0.3 percentage points.

To interpret this result, consider an estimate for the United States of 268,000,000 for the year 2000, when we are now (1984) at 236,000,000. This is a projected annual rate of increase of 0.8 per cent, so odds are 2 to 1 of the true outcome falling within the range 0.8 ± 0.3 or $0.5 - 1.1$; one can bet 2 to 1 odds that the population in the year 2000 will be in the range $(236)(1.005)^{16}$ to $(236)(1.011)^{16}$, or 256 to 281 millions. This supposes that the present forecast is no better and no worse than the 300 similar forecasts on which this estimate of error has been based (Keyfitz 1981).

Exponential and Logistic Growth

There may have been situations in the past when populations were growing uniformly and it was possible to make some kind of credible prediction by supposing constant increase for the future. By definition of the rate of increase,

$$r = \frac{1}{P_t}\frac{\mathrm{d}P_t}{\mathrm{d}t},$$

so that the population at time t is

$$P_t = P_0 e^{rt}.$$

It is hard to think of cases where such exponential growth persists over more than a very short period.

The patent defect of the exponential that nothing can grow uniformly for very long suggested a further factor in the differential equation to produce the curve known as logistic:

$$r_t = \frac{1}{P_t}\left(1 - \frac{P_t}{A}\right)\frac{\mathrm{d}p_t}{\mathrm{d}t},$$

where A is the asymptotic population at which growth stops. The rate of increase r_t is no longer constant, and the solution of the equation is

$$P_t = \frac{A}{1 + be^{-ct}},$$

where b and c are constants.

The logistic seemed to have merit when births were slowing and total population growth tapering off. It reached the height of its popularity when the Americas could be seen as empty, and as they filled would move towards a population ceiling. Unfortunately the ceiling keeps changing with changing society and technology.

One might take a different line in support of the logistic: not the logic of the model but goodness of fit to the historical series. That does not work either; an inverse tangent, or a cumulative normal fit just as well as a logistic, and an impossible curve, a hyperbola moving off to infinity in a near future, is not much inferior to any of the three in fitting the past (Cohen 1984).

For animal populations the story is different; real niches filling under constant conditions do appear, and in ecological studies the logistic has on occasion provided a useful representation of the process.

Difficult Matters

Some demographic results are perfectly explicable: when Romania suddenly banned abortion, the birth rate, which presumably included a proportion of unwanted children, rose sharply; after the public adapted to the ban the birth rate settled back to where it was. Others remain puzzles even after much study: why does West Germany stay at the lowest recorded fertility of all time, much lower than neighbouring France? The effectiveness of determined population policy in East Germany is partly explained by the large expenditure on it, but not Hungary's extremely low fertility after the war, and the subsequent partial recovery.

Similarly, there is much to explain in poor countries; some countries have seen their fertility fall drastically, while others remain high. Cultural inheritance is apparently a factor. Islamic populations have higher fertility than non-Islamic that are otherwise similar; thus for 1980–85 the UN (1985) estimates Pakistan's TFR (Total Fertility Rate) at 5.84 and Bangladesh's at 6.15 against India's 4.41. What feature of Islam is the cause of the differential remains to be discovered.

A key question in contemporary demography is whether and how quickly the countries whose death rates have fallen can follow through with declines in birth rates that will bring them to zero growth. No one knows for sure whether the fall of deaths – for instance and especially of infant mortality – in and by itself brings about a decline of births; the literature contains proofs that it does and proofs that it does not. Even if we knew for sure that the demographic transition to a stationary condition will take place everywhere, forecasting for the years ahead is impeded by our ignorance of how quickly it will come. And professional opinion on the effectiveness of family planning programmes is by no means unanimous.

See Also

- ▶ Demographic Transition
- ▶ Historical Demography
- ▶ Life Tables
- ▶ Stable Population Theory

Bibliography

Arthur, W.B., and J.W. Vaupel. 1984. Some general relationships in population dynamics. *Population Index* 50(2): 214–226.

Blumen, I., M. Kogan, and P.J. McCarthy. 1955. *The industrial mobility of labor as a probability process*, Cornell studies of industrial and labor relations, vol. VI. Ithaca: Cornell University Press.

Bogue, D.J. 1985. *The population of the United States: Historical trends and future projections*. New York: Free Press.

Bourgeois-Pichat, J. 1966. *The concept of a stable population: Application to the study of populations of countries with incomplete population statistics.* ST/SOA/Series A 139. New York: United Nations.

Bowley, A.L. 1924. Births and population of Great Britain. *Journal of the Royal Economic Society* 34: 188–192.

Brass, W. 1971. On the scale of mortality. In *Biological aspects of demography*, ed. W. Brass, 69–110. London: Taylor and Francis.

Brass, W. 1974. Perspectives in population prediction, illustrated by the statistics of England and Wales. *Journal of the Royal Statistical Society, Series A* 137: 532–583.

Brass, W. 1975. *Methods for estimating fertility and mortality from limited and defective data.* An occasional publication. Chapel Hill: University of North Carolina, International Program of Laboratories for Population Statistics.

Cannan, E. 1895. The probability of cessation of growth of population in England and Wales during the next century. *Economic Journal* 5: 505–515.

Coale, A.J. 1963. Estimates of various demographic measures through the quasi-stable age distribution. In *Emerging techniques in population research* (39th annual conference of the Milbank Memorial Fund, 1962), 175–193. New York: Milbank Memorial Fund.

Coale, A.J. 1966. *Methods of estimating fertility and mortality from censuses of population*. Princeton: Office of Population Research.

Coale, A.J. 1984. Life table construction on the basis of two enumerations of a closed population. *Population Index* 50(2): 193–213.

Coale, A.J., and P. Demeny. 1983. *Regional model life tables and stable populations*, 2nd ed. New York: Academic.

Cohen, J.E. 1984. Demographic doomsday deferred. *Harvard Magazine* 86(3): 50–51.

Demetrius, L. 1974. Demographic parameters and natural selection. *Proceedings of the National Academy of Sciences* 71: 4645–4647.

Easterlin, R.A. 1980. *Birth and fortune: The impact of numbers on personal welfare*. New York: Basic Books.

Elandt-Johnson, R.C., and N.L. Johnson. 1980. *Survival models and data analysis*. New York: Wiley.

Gini, C. 1924. Premières recherches sur la fecondabilité de la femme. *Proceedings of the International Mathematics Congress* 2: 889–892.

Goldman, N. 1978. Estimating the intrinsic rate of increase from the average numbers of younger and older sisters. *Demography* 15: 499–508.

Goodman, L.A. 1961. Statistical methods for the mover-stayer model. *Journal of the American Statistical Association* 56(296): 841–868.

Goodman, L.A. 1969. The analysis of population growth when the birth and death rates depend upon several factors. *Biometrics* 25: 659–681.

Goodman, L.A., N. Keyfitz, and T.W. Pullman. 1974. Family formation and the frequency of various kinship relationships. *Theoretical Population Biology* 5: 1–27.

Graunt, J. 1662. In *Natural and political observations made upon the bills of mortality*, ed. Walter F. Willcox. London/Baltimore: Johns Hopkins University Press, 1939.

Henry, L. 1957a. Fécondité et famille. Mmdèles mathématiques I. *Population* 12: 413–444.

Henry, L. 1957b. Fécondité et famille. Modèles mathématiques II. *Population* 16: 27–48, 261–282.

Kaplan, E.L., and P. Meier. 1958. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53: 457–481.

Keyfitz, N. 1981. The limits of population forecasting. *Population and Development Review* 7(4): 579–593.

Keyfitz, N. 1985. *Applied mathematical demography*, 2nd ed. New York: Springer.

Kitagawa, E.M., and P.M. Hauser. 1973. *Mortality in the United States. A study in socioeconomic epidemiology.* Cambridge, MA: Harvard University Press.

Le Bras, H. 1973. Parents, grandparents, diaeresis bisaieux. *Population* 28: 9–37. Trans. and ed. K. Wachter as *Statistical studies of historical social structure.* New York: Academic, 1978.

Lee, R.D. 1974. The formal dynamics of controlled populations and the echo, the boom and the bust. *Demography* 11: 563–585.

Leslie, P.H. 1945. On the use of matrices in certain population mathematics. *Biometrika* 33: 183–212.

Lotka, A.J. 1931. Orphanhood in relation to demographic factors. *Metron* 7: 37–109.

Lotka, A.J. 1939. *Théorie analytique des associations biologiques.* Part II: *Analyse démographique avec application particulière à l'espèce humaine.* Actualités Scientifiques et Industrielles, No. 780. Paris: Hermann et Cie.

Mann, N.R., R.D. Schafer, and N.D. Singpurwalla. 1974. *Methods for statistical analysis of reliability and life data*. New York: Wiley.

Pollard, J.H. 1966. On the use of the direct matrix product in analysing certain stochastic population models. *Biometrika* 53: 397–415.

Potter, R.G. 1972. Births averted by induced abortion: An application of renewal theory. *Theoretical Population Biology* 3: 69–86.

Potter, R.G., and M.P. Parker. 1964. Predicting the time required to conceive. *Population Studies* 18: 99–116.

Pressat, R. 1961. *L'analyse démographique: méthodes, résultats, applications*. Paris: Presses Universitaires de France, for Institut National d'Etudes Démographiques.

Preston, S.H., and A.J. Coale. 1982. Age structure, growth, attrition, and accession: A new synthesis. *Population Index* 48(2): 217–259.
Rogers, A. 1975. *Introduction to multiregional mathematical demography.* New York: Wiley.
Ryder, N.B. 1964. The process of demographic transition. *Demography* 1(1): 74–82.
Sauvy, A. 1952–1954. *Théorie générale de la population.* Vol. 1: *Economie et population.* Vol. II: *Biologie sociale.* Paris: Presses Universitaires de France.
Schoen, R. 1975. Constructing increment-decrement life tables. *Demography* 12: 313–324.
Scudo, F.M. 1984. The 'golden age' of theoretical ecology: A conceptual appraisal. *Revue Européenne des sciences sociales* 22(67): 11–64.
United Nations. 1985. *World population prospects: Estimates and projections as assessed in 1982.* ST/ESA/SER.A/82. New York: United Nations.
Vaupel, J.W., and I.Y. Yashin. 1985. The deviant dynamics of death in heterogeneous populations. In *Sociological methodology 1985*, ed. N.B. Tuma. San Francisco: Jossey-Bass.
Whelpton, P.K. 1936. An empirical method of calculating future population. *Journal of the American Statistical Association* 31: 457–473.

Demography of the Ancient World

Walter Scheidel

Abstract

This survey of demographic conditions in ancient Greek and Roman history discusses life expectancy and causes of death, reproduction and fertility control, marriage practices and household structure, population size and its change over time, and the relationship between demographic and economic development.

Keywords

Ancient world; Birth rates; Death rates; Demographics; Family; Greece; Life expectancy; Marriage; Population; Rome

JEL Classifications

B11

Ancient demography covers the population history of early civilizations from the third millennium BCE to the seventh century CE.

Due to the uneven distribution of relevant evidence, scholars have focused primarily on Middle Eastern, Greek and Roman populations. This survey deals primarily with the demography of the Greco-Roman world. Demographic conditions in antiquity are generally only dimly perceptible, and attempts to reconstruct them inevitably entail considerable uncertainty and conjecture. Information is provided by tombstone inscriptions, census documents on papyri, skeletal remains and literary accounts.

Ancient birth and death rates were extremely high by modern standards. Mean life expectancy at birth is commonly estimated to have been around 20 to 30 years. The distribution of ages recorded in census returns from Roman Egypt is consistent with model life tables that posit a mean life expectancy at birth of 22 to 25 years. This estimate receives additional support from a variety of other data samples including funerary inscriptions from Roman North Africa, a Roman schedule used to calculate annuities known as 'Ulpian's Life Table', and the age structure of a few cemetery populations. Roman emperors who died of natural causes had a similarly low life expectancy. This suggests that socio-economic standing had little effect on longevity. Mortality regimes were highly localized and determined primarily by the prevalence of particular infectious diseases. In parts of the Roman empire, seasonal variation in mortality can sometimes be reconstructed with the help of dates of death reported in tombstone inscriptions. These seasonality patterns also allow inferences about the underlying disease environments. According to these datasets, seasonal spikes in adult death rates were much stronger than in the more recent past, suggesting that even the most resilient age groups were susceptible to fatal infections. The main causes of death can be inferred from ancient medical texts and literary sources. Gastrointestinal diseases, malaria and tuberculosis were particularly important. Both malaria and leprosy expanded during the Greco-Roman period.

Smallpox epidemics occurred, possibly in Athens in 430 BCE and probably throughout the Roman empire in the 160/180 s CE. Plague spread from 540 to 750 CE in a pandemic that foreshadowed the medieval Black Death.

High levels of mortality required correspondingly high birth rates. The average woman surviving to menopause had to give birth to five or six children to ensure reproduction at replacement level. Birth rates within marriage were higher still: it has been calculated that in Roman Egypt, a woman who had been continuously married between menarche and menopause would on average have given birth eight or nine times. According to census records from the same region, 95 per cent of freeborn children were born to married parents. These documents also allow us to reconstruct the maternal age distribution of childbirths, which implies what is known as a 'natural fertility' regime in which fecundity was a direct function of a woman's age, peaking around age 20 and gradually declining over time. Signs of stopping behavior – that is, the cessation of reproduction in response to family size or composition – are absent from these data.

At the same time, early and near-universal marriage for women and high birth rates went hand in hand with fertility control within marriage. Census returns from Roman Egypt indicate mean birth intervals of 3–4 years. Birth-spacing may have been achieved by prolonged breastfeeding or by other means. Ancient medical texts discuss a variety of putative contraceptives and abortifacients. More drastic intervention in the form of child exposure and infanticide was often socially condoned, although the actual scale of these practices remains unknown. The extent to which parents discriminated against female offspring is particularly controversial. While Greek and Roman sources sometimes refer to femicide, and evidence of male-biased sex ratios has been taken to reflect this custom, we are usually unable to determine whether 'missing' females had been killed or exposed after birth or were merely omitted from written records.

Among Greeks and Romans, (serial) monogamy was the norm. Polygamous unions were largely confined to ruling and elite families in Middle Eastern societies. At the same time, sexual access to slave women facilitated resource polygyny even in formally monogamous settings. In ancient Greek culture, women often appear to have been married off in their mid-teens while men took wives considerably later, around the age of 30. Funerary inscriptions from the western half of the Roman empire point to typical marriage ages of about 20 years for women and 30 years for men. Roman aristocrats generally married at younger ages. For Roman Egypt, the census records reflect mean marriage ages of 17 or 18 years for women and 25 years for men. They also show that whereas almost all women had married by their late 20 s, it was only by age 50 that most men had married at least once. This pattern of moderately early female and late male marriage resembles the so-called 'Mediterranean marriage pattern' which prevailed in the more recent past, suggesting a measure of long-term continuity in that region. Divorce could be initiated by both husbands and wives, and commonly lacked strong stigma. Remarriage was much more common for men than for women, especially after age 30: according to the Roman Egyptian census returns, two-thirds of men but only one-third of women were still married at the age of 50. In the pre-Christian period, celibacy was not normally considered desirable.

Marriages were mostly virilocal or neolocal. Bridal dowries were common but are best attested for elite circles. Slaves could not legally marry but were able to enter informal unions, primarily (but not only) with other slaves. Consanguineous unions were more widespread in the eastern Mediterranean, and especially in the Middle East, than in western Europe. Thus, first-cousin marriage occurred mostly in the East, and occasional half-sibling unions are known from the Greek world. Scholars still debate whether references to married couples of full siblings found in Roman Egyptian census documents reveal genuinely incestuous unions or record the unions of cousins who had legally become siblings through adoption. However, instances of brother–sister and parent–child marriage are credibly attested for

ancient Middle Eastern rulers, and more generally for members of the Zoroastrian community in Mesopotamia and Iran.

The Greek and Latin languages lack specific terms for what we would call the nuclear family. Notions of family and household were more inclusive: next to parents and their offspring, the Greek *oikos* and the Roman *familia* or *domus* routinely encompassed co-resident kin and slaves. At the same time, Roman funerary inscriptions tend to privilege commemorative ties within the nuclear family, showing it to have been the principal locus of familial sentiment and obligation, of inheritance, and probably also of residence. More complex households were common in the eastern Mediterranean and the Middle East. In Roman Egypt, for example, the majority of the rural population belonged to households comprised of extended or multiple families. High death rates offset high fertility, thereby limiting family size, which averaged 4.3 in the same region.

Owing to unpredictable mortality and the desire to preserve male lineages, adoption of relatives appears to have been common. Partible inheritance rather than primogeniture was the norm. Daughters either received dowries as a substitute for an inheritance or inherited alongside their brothers. The social effects of high death rates undermined the formally patriarchal character of ancient households. A significant share of Greeks and Romans must have lost their fathers as minors and were assigned guardians, while many widows were unable to remarry. For these reasons, family units in which women and children were under the control of fathers and husbands were less common and more fragile than modern observers have often imagined.

Population numbers are very poorly known and continue to generate controversy. Statistical documents survive only from parts of Egypt, and literary references to population size are commonly vitiated by rhetorical stylization, ignorance or indifference. Archaeological data help to fill this gap but pose their own problems of interpretation. What we do know is that the Mediterranean regions and its hinterlands underwent significant population growth in the Greco-Roman period. In the Aegean, the collapse of late Bronze Age civilization around 1200 BCE coincided with strong demographic contraction. Population recovered from the early first millennium BCE onward and peaked in the classical period, in the fifth and fourth centuries BCE, when Greece may have been more densely populated than at any other time prior to the 20th century.

During this growth phase, Greek settlers established hundreds of colonies in Sicily, South Italy and the Black Sea littoral. By the fourth century BCE, up to 1000 Greek city-states were inhabited by 7 million people or more. Most of these communities were very small. The conquests of Alexander the Great in the late fourth century BCE triggered Greek emigration to Egypt, Syria, Mesopotamia, Iran and Central Asia. Large-scale state formation under his successors led to the creation of capital cities in excess of 100,000 residents, most notably Alexandria in Egypt. Meanwhile, populations expanded farther west in Italy, where this process drove the conflicts that eventually resulted in Roman regional hegemony, and more generally in western Europe and the Maghreb. A series of Roman census tallies from the last three centuries BCE offers insight into demographic change on the Italian peninsula. Even so, the total size of its population cannot be established with precision: depending on different interpretations of the extant census counts, by the beginning of the Common Era Italy may have been inhabited by no more than 6 million people (including slaves) or by two or even three times as many. These uncertainties interfere with modern assessments of Roman economic performance.

For a variety of reasons, an estimate between the extreme ends of spectrum seems appropriate: with a peak population of perhaps 10 or 12 million people, Roman Italy may well have matched the population densities of the high medieval and early modern periods. The population of the Roman empire as a whole is necessarily even more difficult to ascertain: while a total of 60–80 million seems realistic, a higher figure cannot entirely be ruled out. Maybe 10–20 per cent of these individuals lived in some 2000 cities. The capital city of Rome appears to have grown to a million residents, an urban population unparalleled in Europe prior to

London around 1800. Starting in the late second century CE, epidemics reduced population numbers, although settlement densities remained high into late antiquity. Massive population losses finally accompanied the disintegration of the western half of the Roman empire in the fifth century CE and the onset of recurrent plague pandemics in the 540 s CE.

Despite its overall paucity and numerous shortcomings, demographic information from the Greco-Roman world is of considerable relevance to our understanding of ancient economic history. Centuries of continuous population growth, first in the eastern Mediterranean and later farther west, highlight the scale and persistence of an economic expansion which was driven by the spread of farming, technological innovation and gains from trade. Concurrent urbanization reinforces our impression of dynamic economic development. In the long run, however, ancient economies resembled other premodern economies in their inability to overcome Malthusian pressures through ongoing technological innovation. As population continued to expand, per capita economic growth eventually abated, first in Greece and later in the western Mediterranean. Judging by a variety of archaeological proxies of economic performance, by the time exogenous shocks in the form of plague and invasions began to affect the Roman empire in the second and third centuries CE the economy had already ceased to grow in real terms.

There is no indication that the Greco-Roman economic-demographic expansion significantly improved health or longevity: accretions to the stock of knowledge proved insufficient to mitigate the impact of the main causes of death, and potential gains from infrastructural provisions (such as aqueducts) may well have been offset by the demographic burden of urbanization and rising population densities which increased exposure to infection. In a number of skeletal samples, average body height was smaller in the Roman period than both immediately before and after, which likewise speaks against the notion of improvements in physiological wellbeing. Moreover, widespread skeletal evidence of deficiency diseases points to pervasive morbidity which would have curtailed productivity. High death rates discourage investment in education and impede human capital formation. Correspondingly high fertility depresses female labour participation and the status of women. In this environment, sustainable economic growth, let alone a fertility transition, was not feasible.

D

See Also

- ▶ Economic History
- ▶ Historical Demography
- ▶ Population Dynamics
- ▶ Population Health, Economic Implications of

Bibliography

Bagnall, R.S., and B.W. Frier. 1994. *The demography of Roman Egypt*. Cambridge: Cambridge University Press.

Brunt, P.A. 1971. *Italian Manpower 225 BC–AD 14*. Oxford: Clarendon Press.

Frier, B.W. 1994. Natural fertility and family limitation in Roman marriage. *Classical Philology* 89: 318–333.

Frier, B.W. 2000. Demography. In *The Cambridge Ancient History volume 11*, ed. A.K. Bowman, P. Garnsey, and D. Rathbone. Cambridge: Cambridge University Press.

Hansen, M.H. 2006. *The Shotgun Method: The demography of the Ancient Greek City-State Culture*. Columbia: University of Missouri Press.

Parkin, T.G. 1992. *Demography and Roman Society*. Baltimore: Johns Hopkins University Press.

Pomeroy, S.B. 1997. *Families in Classical and Hellenistic Greece: Representations and realities*. Oxford: Clarendon Press.

Sallares, R. 1991. *The Ecology of the Ancient Greek World*. London: Duckworth.

Saller, R.P. 1994. *Patriarchy, property and death in the Roman family*. Cambridge: Cambridge University Press.

Scheidel, W. 2001a. *Death on the Nile: Disease and the demography of Roman Egypt*. Leiden: Brill.

Scheidel, W. 2001b. Progress and problems in Roman demography. In *Debating Roman demography*, ed. W. Scheidel. Leiden: Brill.

Scheidel, W. 2007. Demography. In *The Cambridge economic history of the Greco- Roman World*, ed. W. Scheidel, I. Morris, and R. Saller. Cambridge: Cambridge University Press.

Scheidel, W. 2008. Roman population size: The logic of the debate. In *People, land and politics: Demographic developments and the transformation of Roman Italy 300 bc-ad 14*, ed. L. De Ligt and S. Northwood. Leiden: Brill.

Denison, Edward (1915–1992)

Barry Bosworth

Keywords

American Economic Association; Capacity utilization; Capital accumulation; Denison, E.; Growth accounting; National income accounting; Production theory; Standardized system of national accounts (SNA); Stone, J. R.N.; Total factor productivity

JEL Classifications

B31

Edward Denison was a major contributor to the development of the US national income accounts and one of the originators of growth accounting. He received a Ph.D. in economics from Brown University in 1941. Denison's early career (1941–56) was spent in the national income division of the US Commerce Department where he worked with Milton Gilbert, George Jaszi, and Charles Schwartz to develop the national accounts of the United States. The United States had published estimates of national income and its components in 1934; and Richard Stone and others developed both expenditure and income-side estimates of GNP for the United Kingdom that were published in 1941. The US expenditure-side estimates were first published in 1942.

Denison participated in a 1944 tripartite meeting with Canada, the United Kingdom, and the United States that worked to establish consensus on a set of concepts and methods for the national accounts. That meeting and subsequent work provided much of the basis for the standardized system of national accounts (SNA) that was adopted and expanded by the United Nations and the OECD. The United States did not initially adopt the SNA; but by 2000 it was following the SNA in all of its important respects.

Denison moved to the Committee on Economic Development (CED) in 1956 where his research focused on identifying the sources of economic growth. In expanding the framework of growth accounting, Denison sought to go beyond a simple partitioning of economic growth into the contributions of the factor inputs and a residual of total factor production. He incorporated changes in the quality of the inputs, such as job skills, economies of scale, and other contributors to the residual, such as research and development. His initial analysis was published by the CED in 1962 as *The Sources of Economic Growth in the United States and the Alternatives Before Us.* A distinctive feature of his approach was the extent to which he anchored it in the basic accounting framework of the national accounts rather than the concepts of neoclassical production theory employed a few years later by Jorgenson and Griliches (1967). This aspect made it easy for other researchers to duplicate his methodology within their own countries.

Denison moved to the Brookings Institution in 1963 and extended his analysis to international comparisons with publication of *Why Growth Rates Differ* (1967). Two important later contributions were *How Japan's Economy Grew So Fast* (with W.K. Chung, 1976) and *Accounting for Slower Economic Growth: The United States in the 1970s* (1979). In *Accounting for Slower Growth,* he explored a wide range of popular explanations for the productivity slowdown, including higher energy prices, government regulation, and reduced R&D expenditures, and argued that their effects were too small to account for the magnitude and persistence of the slowdown. He received the Distinguished Fellow Award of the American Economic Association in 1981.

Denison's exchanges with Jorgenson and Griliches (1967, 1972a), while centred around differences in their approaches to measuring the contributions to growth, served to highlight an ongoing debate about the relative importance of capital accumulation and total factor productivity gains. Denison's approach, by minimizing several aspects of the measurement of capital's contribution, tended to support the conventional wisdom of the time that TFP accounted for a substantial portion of growth. Jorgenson and Griliches were attempting to argue that careful measurement of the factor

inputs could drastically shrink the residual contribution of TFP. Denison won out on the issue of the relative importance of TFP by pointing to some problems with Jorgenson–Griliches adjustment for variations in capacity utilization; but the longer-term value of the debate was in showing that their approaches were quite similar. In subsequent years, the Jorgenson–Griliches approach, with its anchor in production theory, has dominated the conceptual discussion. However, many of the empirical studies continue to follow Denison's careful use of national income accounts data.

See Also

- ▶ Economic Growth
- ▶ Growth Accounting
- ▶ Total Factor Productivity

Selected Works

1962. *The sources of economic growth in the United States and the alternatives before us.* New York: Committee for Economic Development.
1967. *Why growth rates differ?* Washington, DC: Brookings Institution.
1969. Some major issues in productivity analysis: An examination of estimates by Jorgenson and Griliches. *Survey of Current Business* 49(5, Part 2): 1–27.
1972. Final comments. *The measurement of productivity.* Washington, DC: Brookings Institution.
1976. (With W. K. Chung.) *How Japan's economy grew so fast: The sources of postwar expansion*. Washington, DC: Brookings Institution.
1979. *Accounting for slower economic growth: The United States in the 1970s.* Washington, DC: Brookings Institution.

Bibliography

Jorgenson, D., and Z. Griliches. 1967. The explanation of productivity change. *Review of Economic Studies* 34: 249–83.
Jorgenson, D. and Z. Griliches. 1972a. Issues in growth accounting: A reply to Edward F. Denison. *The measurement of productivity.* Washington, DC: Brookings Institution.
Jorgenson, D. and Z. Griliches. 1972b. Final reply. *The measurement of productivity.* Washington, DC: Brookings Institution.

Dependency

D

José Gabriel Palma

Abstract

The focus of all 'dependency' analyses is the development of peripheral capitalism (or lack of it). One approach, begun by Baran, Sweezy and Frank, attempted to construct a theory of the practical impossibility of capitalist development in the periphery. A second emerged from the Structuralist School, especially Furtado, Pinto and Sunkel, and tried to reformulate the classical ECLAC analysis from the perspective of the obstacles to 'national' development. A third, initiated by Cardoso and Faletto, concentrated on studying 'concrete situations of dependency' – how the specific dynamic of different peripheral societies emerges from the interaction between their internal and external structures.

Keywords

Baran, P.; Capitalism; Capitalist development; Dependency; Economic development; Exploitation; Frankfurt School; Furtado, C.; Harrod–Domar theory; Imperialism; Industrialization; Lenin, V. I.; Marx, K. H.; Marx's analysis of capitalist production; Monopoly capital; Multinational corporations; Periphery; Socialism; Structuralism; Surplus; Underdevelopment

JEL Classifications

O1

Dependency theories emerged in Latin America in the early 1960s as a challenge to traditional

Marxist and structuralist thinking regarding whether capitalist development in the periphery was both still *viable* (given the transformations of the world economy after the Second World War), and still *necessary* (as an unavoidable transition step towards socialism).

There can be little doubt that the Cuban Revolution was a turning point in Marxist analysis of capitalist development in the periphery. The events in Cuba gave rise to a new approach, of which most of the 'dependency analyses' form part. This argued that capitalism had totally lost its historical 'progressive' role in the periphery (if it ever had one); that is, it was both no longer capable of developing the productive forces of backward societies, and (thus) no longer able to bring them closer towards socialism. Consequently, this approach also argued against the politics of the popular fronts in the periphery and in favour of an immediate transition towards socialism.

Following traditional Marxist analysis, the pre-dependency, pre-Cuban Revolution approach saw capitalism as still historically progressive in the periphery; however, it argued that its key historical task – the 'bourgeois-democratic' revolution – was being inhibited by a new alliance between imperialist forces and the traditional oligarchies. The bourgeois-democratic revolution was the revolt of the emerging capitalist forces of production against the old pre-capitalist order. This revolution would be based on an alliance between the rising bourgeoisie and other progressive forces of society; the principal battle line would be between the new capitalist elites and the traditional oligarchies – between industry and land, capitalism versus pre-capitalist forms of monopoly and privilege. Because it would be the result of the pressure of a rising class whose path was being blocked in political, economic and social terms, this revolution would bring to the periphery (as it had done in the centre) not only political emancipation but economic progress as well.

One of the main analytical challenges facing the pre-dependency Marxist analysis was to explain why the 'bourgeois–democratic' revolution in the periphery was not really happening as expected (a phenomenon that was seriously hindering the process of capitalist development there). Since Lenin, this analysis had identified imperialism as the unmistakable main obstacle facing this revolution. The traditional oligarchies could not be the reason for this as on their own, they were not expected to prove any match for the new emerging capitalist classes. Therefore, the principal target in this struggle was unmistakable: North American imperialism. The allied camp for this fight, by the same reasoning, was also clear: everyone, except those (pre-capitalist) internal groups allied with imperialism. Thus, the anti-imperialist struggle was at the same time a struggle for domestic capitalist development and industrialization. The state and the 'national' bourgeoisie appeared as the potential leading agents for the development of the new capitalist economy, which in turn was viewed as a necessary stage towards socialism.

The Cuban Revolution questioned the very essence of this approach, insisting that the domestic bourgeoisies in the periphery no longer existed as a progressive social force but had become 'lumpen', 'rent seekers', incapable of rational accumulation and rational political activity, dilapidated by their consumerism, and blind to their 'real' long-term interests. It is within this framework, and with the explicit motive of developing theoretically and documenting historically this new approach that dependency analysis appeared on the scene. At the same time, both inside and outside the Economic Commission for Latin America (ECLAC), two other major Dependency Schools began to develop (see structuralism).

The general focus of all 'dependency' analyses is the development of peripheral capitalism (or, rather, the lack of it). More specifically, these studies attempted to analyse the obstacles to capitalist development in the periphery from the point of view of the new interplay between 'internal' and 'external' structures that had emerged after the Second World War. However, this interplay was analysed in several different ways.

With the necessary degree of simplification that every classification of intellectual tendencies entails, I distinguish between three major approaches – not mutually exclusive from the point of view of intellectual history – in 'dependency' analysis. First is the approach begun by Paul Baran, Paul Sweezy and Andre Gunder

Frank; its essential characteristic is that it attempted to construct a comprehensive theory of the practical impossibility of capitalist development in the periphery. In these theories the 'dependent' character of peripheral economies is the crux on which the whole analysis of underdevelopment turns; that is, dependency is seen as causally linked to permanent capitalist underdevelopment.

The second approach is associated with the ECLAC Structuralist School, especially Celso Furtado, Aníbal Pinto and Osvaldo Sunkel. These writers sought to reformulate the classical ECLAC analysis of Latin American development from the perspective of a critique of the obstacles to 'national' development. This attempt at reformulation was not just process of adding new elements (mainly political and social) that were lacking in the original Prebisch–ECLAC analysis (see Prebisch, Raúl), but a thoroughgoing attempt to proceed beyond that analysis, adopting an increasingly different perspective.

Finally, the third approach, started by Fernando Henrique Cardoso and Enso Faletto, attempted to distance itself from the first by deliberately avoiding the formulation of a mechanico-formal theory of dependency and underdevelopment – specifically, by trying to avoid a mechanico-formal theory of the inevitability of underdevelopment in the capitalist periphery based on its dependent character. In turn, it concentrated on the study of what have been called 'concrete situations of dependency'; that is to say, the precise forms in which the different economies and polities of the periphery have been articulated with those of the advanced nations at different times, and how their specific dynamics have thus been generated.

The First Approach: Dependency as a Formal Theory of the Inevitability of Capitalist Underdevelopment: On Cutting a Knot That Could Not Be Unravelled

There is no doubt that the 'father' of this approach was Baran. His principal contribution (1957) took up the approach of the Sixth Congress of the COMINTERN regarding the supposedly irresolvable nature of the contradictions between the economic and political needs of imperialism and those of the processes of political transformation, economic development and industrialization of the periphery.

To defend its interests, international monopoly capital would not only form alliances with pre-capitalist domestic oligarchies intended to block progressive capitalist transformations in the periphery, but its activities would also have the effect of distorting the process of capitalist development in these countries. As a result, international monopoly capital would have easy access to peripheral resources and finance, and the traditional élites in the periphery would be able to maintain their monopoly on power and their traditional (mostly predatory and rent-seeking) modes of surplus extraction. Within this context the possibilities for any form of dynamic economic growth in dependent countries were extremely limited or non-existent; the surplus they were able to generate (mainly from primary commodity export activities) was largely appropriated by foreign capital, or otherwise squandered by traditional elites. Therefore, long-term economic stagnation and underdevelopment was inevitable. The only way out was political. At a very premature stage, capitalism had become a fetter on the development of the productive forces in the periphery and, consequently, its historical role had already come to an early end.

Baran developed his ideas influenced both by the Frankfurt School's general pessimism regarding the nature of capitalist development (see Jay 1996) and by Sweezy's (1946) proposition that the rise of monopolies imparts to capitalism a tendency towards stagnation and decay (see monopoly capitalism). He also followed the main growth paradigm of his time, the Harrod–Domar model, which held that the size of the investable surplus was the crucial determinant of growth (together with the efficiency with which it was used: the incremental capital–output ratio).

Starting out with Baran's analysis, Frank (1967) attempted to prove the thesis that the only

political and economic solution to capitalist underdevelopment was a radical transformation of an immediately socialist character. For our purposes we may identify three levels of analysis in Frank's model of the 'development of underdevelopment'. In the first (arguing against 'dualistic' analyses), he attempted to demonstrate that the periphery had been incorporated and fully integrated into the world capitalist economy since the very early stages of colonial rule. In the second, he tried to show that such incorporation into the world capitalist economy had transformed the countries in question immediately into capitalist economies. Finally, in the third level, Frank attempted to prove that the integration of these supposedly capitalist economies into the world capitalist system was achieved through an interminable metropolis–satellite chain, through which the surplus generated at each stage was successfully siphoned off towards the centre. Therefore, for Frank the choice was clear: continue to endlessly underdevelop within capitalism, or socialist revolution.

In my opinion, the real value of Frank's analysis is his critique of the supposedly dual structure of peripheral societies. Frank argues convincingly that the different sectors of the economies in question are and have been, since very early in their colonial history, well integrated to the world economy. Moreover, he has correctly emphasised that this integration has not automatically brought about capitalistic economic development, such as 'optimistic' models (derived from Adam Smith) would have predicted, in which increased international trade and the division of labour would inevitably bring about economic growth and prosperity. Nevertheless, Frank's error lies in his attempt to explain this phenomenon by using the same economic deterministic framework of the model he purports to transcend. In fact, he merely turns it upside-down: integration into the world economy cannot possibly bring about capitalism development in the periphery because the development of the industrialised centre necessarily requires the underdevelopment of the periphery. Frank's error is characteristic of the whole tradition of which he is part, including Baran (1957), Sweezy (1946), Amin (1970) and Wallerstein (1974, 1980) among the better known. In their analysis, there is always a priority of external over internal structures; in order to do this, they have to separate almost metaphysically the two sides of the opposition (the internal and the external), losing in the process the notion of movement through the dynamic of the contradictions between these two structures. The analysis which emerges is one typified by 'antecedent causation and inert consequences'.

It is not surprising that this type of analysis leads Frank to develop a circular concept of capitalism. Although it is evident that capitalism is a system where production for profits via exchange predominates, the opposite is not necessarily true: the existence of production for profits in the market is not necessarily an indication of capitalist relationship of production. For Frank, this is a sufficient condition for the existence of capitalist forms of surplus extraction (and for the periphery to have been 'capitalist' since the beginning of colonial rule).

Although Frank did not go very far in his analysis of the world capitalist system as a whole, of its origins and its development, Amin (1970) and Wallerstein (1974, 1980) tackled this tremendous challenge. The central concerns of Frank's theory of the 'development of underdevelopment' are also addressed by dos Santos (1970), Marini, Caputo, Pizarro, Hinkelammert, and continued later on by many non-Latin American social scientists. The most thoroughgoing critiques of these theories of underdevelopment have come from Brenner (1977), Cardoso (1972), Kay (1989), Laclau (1971), Lall (1975), Palma (1978), and Warren (1980).

I would argue that the theories of dependency examined here are mistaken not only because they do not 'fit the facts', but also – and equally important – because their mechanico-formal nature renders them both static and ahistorical. Their analytical focus has not been directed to the understanding of how new forms of capitalist development in the periphery have been marked by a series of specific economic, political, and social contradictions, instead only to assert the claim that capitalism had lost, or never had, a historically progressive role in the periphery.

Now, if the argument is that the progressiveness of capitalism has manifested itself in the periphery differently from in advanced capitalist countries, or in diverse ways in the different branches of the peripheral economies, or that it has generated inequality at regional levels and in the distribution of income, and has been accompanied by such phenomena as unemployment, and has benefited the elite almost exclusively, or again that it has taken on a cyclical nature, then this argument does no more than affirm that the development of capitalism in the periphery has been characterized by its contradictory and exploitative nature. The specificity of capitalist development in the Third World stems precisely from the particular ways in which these contradictions have been manifested, the different ways in which many of these countries have faced and temporarily overcome them, the ways in which this process has created further contradictions, and so on. It is through this process that the specific dynamic of capitalist development in different peripheral countries has been generated.

Reading their political analysis, one is left with the impression that the whole question of what course the revolution should take in the periphery revolves solely around the problem of whether or not capitalist development is viable. In other words, their conclusion seems to be that, if one accepts that capitalist development is feasible on its own terms, one is automatically bound to adopt the political strategy of waiting for and/or facilitating such development until its full productive powers have been exhausted, and only then to seek to move towards socialism. As it is precisely this option that these writers wish to reject, they have been obliged to make in their work a forced march back towards a pure ideological position to deny any possibility of capitalist development in the periphery.

The Second Approach: Dependency as a Reformulation of the ECLAC Analysis of Latin American Development

Towards the end of the 1960s the analysis of ECLAC regarding Latin American development suffered a gradual decline due to several key factors (see Furtado, Celso). The apparently gloomy panorama of capitalist development in Latin America in the 1960s led to substantial ideological changes in many influential ECLAC thinkers, and it strengthened the convictions of the Marxist 'dependency' writers reviewed earlier. The former were faced with the problem of trying to explain the apparent failure of their structuralist policies, particularly concerning import-substituting industrialization (see structuralism). The latter felt vindicated in their view of the unfeasibility of any form of 'dependent capitalist development'.

Finally, by making a basically ethical distinction between 'economic growth' and 'economic development', most of the research done within the perspective of this second approach followed two separate lines, one concerned with the obstacles to economic growth (and in particular to manufacturing), the other with the apparently perverse character taken by capitalist development. The fragility of this formulation lies in its inability to distinguish between a socialist critique of capitalism and the analysis of the actual obstacles to capitalist development in the periphery.

The Third Approach: Dependency as a Methodology for the Analysis of 'Concrete Situations of Development'

In my critique of the dependency studies reviewed so far, I have described the fundamental elements of what I understand to be the third of the three approaches within the dependency school. This approach is primarily associated with the work of Cardoso and Faletto, dating from the completion of their 1967 book.

Briefly, this third approach to the analysis of dependency can be summarized as follows.

1. In common with the two other approaches to 'dependency' discussed already, this third approach sees the Latin American economies as an integral part of the world capitalist system, in the context of increasing internationalization of the system as a whole. It also argues that the central dynamic of that system lies

outside the peripheral economies and that, therefore, the options which are open to them are limited (but not determined) by the development of the system at the centre. In this way the 'particular' is in some way conditioned by the 'general'. Therefore, a basic element for the analysis of these societies is given by the understanding of the general determinants of the world capitalist system, which is itself rapidly changing. However, the theory of imperialism, which was originally developed to provide an understanding of the dynamics of that system, has had enormous difficulty in keeping up with the significant and decisive changes in the capitalist system since the death of Lenin. During this period, capitalism underwent substantial changes, and the theory failed to keep up with them properly.

One widely recognized characteristic of the third approach to dependency has been its effort to incorporate these transformations. For example, this approach was quick to grasp that the rise of the multinational corporations after the Second World War progressively transformed centre–periphery relationships, as well as relationships between the countries of the centre. As foreign capital became increasingly directed towards manufacturing industry in the periphery, the struggle for industrialization, which was previously seen as an anti-imperialist struggle, in some cases increasingly become the goal of foreign capital itself. Thus dependency and industrialization ceased to be necessarily contradictory processes, and a path of 'dependent development' for important parts of the periphery became possible.

2. The third approach has not only accepted but has also tried to enrich the analysis of how developing societies are structured through unequal and antagonistic patterns of social organization, showing the social asymmetries, the exploitative character of social organization and its relationship with the socio-economic base. This approach has also given considerable importance to the particular aspects of each economy like the effect of the diversity of natural resources, geographic location and so on, thus also extending the analysis of the 'internal determinants' of the development of peripheral economies.
3. However, while these improvements are important, the most significant feature of this approach is that it attempts to go beyond the analysis these internal and external elements, and insists that from the premises so far outlined one arrives at only a partial, abstract and indeterminate characterization of the historical process in the periphery, which can only be overcome by understanding how the 'general' and the 'specific' determinants interact in particular and concrete situations. It is only by understanding the specificity of 'movement' in the peripheral societies as a dialectical unity of both these internal and external factors that one can explain the particularity of social, political and economic processes in these societies.

Only in this way can one explain how, for example, the same process of mercantile expansion could simultaneously produce systems of slave labour, systems based on other forms of exploitation of indigenous populations, and incipient forms of wage labour. What is important is not simply to show that mercantile expansion was the basis of the transformation of most of the periphery, and even less to deduce mechanically that that process made these countries immediately capitalist. Rather, this approach emphasizes the specificity of history and seeks to avoid vague, abstract concepts by demonstrating how, throughout the history of backward nations, different sectors of local classes allied or clashed with foreign interests, organized different forms of the state, sustained distinct ideologies or tried to implement various policies or defined alternative strategies to cope with a constantly changing imperialist challenge.

The study of the dynamic of dependent societies as a dialectical unity of internal and external factors implies that the conditioning effect of each on the development of these societies can be separated only by undertaking a static (and metaphysical) analysis. Equally, if the internal dynamic of

the dependent society is a particular aspect of the general dynamic of the capitalist system, it does not imply that the latter produces concrete effects in the former, but only that it finds concrete expression in that internal dynamic.

The system of 'external domination' reappears as an internal phenomenon through the social practices of local groups and classes, who share the interests and values of external forces. Other internal groups and forces oppose this domination, and in the concrete development of these contradictions the specific dynamic of the society is generated. It is not a case of seeing one part of the world capitalist system as 'developing' and another as 'underdeveloping', or of seeing imperialism and dependency as two sides of the same coin, with the underdeveloped or dependent world reduced to a passive role determined by the other.

There are, of course, elements within the capitalist system that affect all developing economies, but it is precisely *the diversity within this unity* that characterizes historical processes. Thus the analytical focus should be oriented towards the elaboration of concepts capable of explaining how the general trends in capitalist expansion are transformed into specific relationships between individuals, classes and states, how these specific relations in turn react upon the general trends of the capitalist system, how internal and external processes of political domination reflect one another, both in their compatibilities and their contradictions, how the economies and polities of peripheral countries are articulated with those of the centre, and how their specific dynamics are thus generated.

However, as is obvious, this third approach to the analysis of peripheral capitalism is not unique to 'dependency' studies and as such, in time, has superseded them.

See Also

- ▶ Baran, Paul Alexander (1910–1964)
- ▶ Engels, Friedrich (1820–1895)
- ▶ Furtado, Celso (1920–2004)
- ▶ Lenin, Vladimir Ilyich [Ulyanov] (1870–1924)
- ▶ Marx, Karl Heinrich (1818–1883)
- ▶ Marx's Analysis of Capitalist Production
- ▶ Monopoly Capitalism
- ▶ Structuralism
- ▶ Sweezy, Paul Marlor (1910–2004)

Acknowledgment *I am extremely grateful to Fiona Tregenna for many constructive comments.*

D

Bibliography

Amin, S. 1970. *L'accumulation à l'échelle mondiale: critique de la théorie du sous-développement.* Paris/New York: Anthropos/Monthly Review Press, 1975.

Baran, P. 1957. *The political economy of growth*. New York: Monthly Review Press.

Brenner, R. 1977. The origins of capitalist development: A critique of neo–Smithian Marxism. *New Left Review* 104: 25–93.

Cardoso, F.H. 1972. Dependency and development in Latin America. *New Left Review* 74: 83–95.

Cardoso, F.H. and Faletto, E. 1967. *Dependencia y Desarrollo en América Latina.* Mexico/Berkeley: Siglo XXI/University of California Press, 1977.

Dos Santos, T. 1970. The structure of dependence. *American Economic Review* 60: 231–236.

Frank, A.G. 1967. *Capitalism and underdevelopment in Latin America: Historical studies of Chile and Brazil*. New York: Monthly Review Press.

Jay, M. 1996. *The dialectical imagination: A history of the Frankfurt School and the Institute for Social Research 1923–1950*. Berkeley: University of California Press.

Kay, C. 1989. *Latin American theories of development and underdevelopment*. London: Routledge.

Laclau, E. 1971. Feudalism and capitalism in Latin America. *New Left Review* 67: 19–38.

Lall, S. 1975. Is dependence a useful concept in analysing underdevelopment? *World Development* 11: 799–810.

Owen, R., and B. Sutcliffe, eds. 1972. *Studies in the theory of imperialism*. London: Longman.

Palma, J.G. 1978. Dependency: A formal theory of underdevelopment or a methodology for the analysis of concrete situations of underdevelopment? *World Development* 6: 881–924.

Sweezy, P.M. 1946. *The theory of capitalist development*. London: D. Dobson.

Wallerstein, I. 1974. *The modern world system: Capitalist agriculture and the origins of the European world–economy in the sixteenth century.* New York: Academic.

Wallerstein, I. 1980. *The modern world system II: Mercantilism and the consolidation of the European world-economy, 1600–1750*. New York: Academic.

Warren, B. 1980. In *Imperialism: Pioneer of capitalism*, ed. J. Sender. London: Verso.

Depletion

Arnold C. Harberger

The concept of depletion is that counterpart of depreciation which is normally applied in extractive industries. The need for a different concept may fairly be questioned; as in many facets of economics, the explanation is more historical than analytical. Traditionally, land has been regarded as a non-depreciable asset, yet mineral rights have most often (though by no means always) been viewed as attached to the land. The special term *depletion* thus applies to the special circumstances where land loses value (actually or potentially) through a process of extraction of some non-reproducible element in the soil or subsoil.

The foregoing definition should make it clear that it is in principle not the land that is depleted, but the mineral deposits contained therein. It would accordingly not be appropriate in principle to allow that the full amount paid for a property should be deductible in concept of depletion. In practice, the full cost is frequently allowed, but typically only in circumstances where the mineral rights constitute the lion's share of the market value of the holding.

When the corporation income tax was introduced in the United States in 1909, the legislation embodied the concept of cost depletion. There was a transition provision (grandfather clause), however, which applied to mineral deposits that were already being exploited at the time of the law's enactment. Here, the allowable basis on which depletion could be claimed was the fair market value of the mineral holding when the law took effect. Ironically, this tiny and apparently innocuous clause contained the seeds of the half-century of political controversy and of the billions of dollars of economic waste that resulted from 'percentage depletion'.

The process began with the clamour for oil to fuel the World War I effort of Britain, France, and their allies, as the United States from 1916 onward abandoned any pretence of neutrality and moved step by step toward the status of a direct belligerent. Those involved in the oil industry were quick to point out that the wells discovered as a result of their own exploration activities were being taxed more heavily than 'old wells' that had already been functioning when the income tax law took effect. The oil interests succeeded by 1918 in obtaining parity of treatment; that is, newly discovered wells could now claim depletion allowances based not on their cost, but on their 'discovery value' – their market value at the time of their discovery.

Economists might well wonder how replacing cost depletion by discovery depletion could be so bad, since economic forces normally work to bring about a close relationship between the value produced by an economic activity and its cost. This normal tendency was slightly marred by the tax-paying entity being allowed to opt each year for the larger of discovery or cost depletion, but this was only a minor flaw in the legislation. The major flaw lay in the difficulty of reconciling legal and economic concepts of cost in face of the aleatory nature of petroleum exploration.

Petroleum exploration (and the search for most other minerals) has always been a rather risky business. Firms have typically drilled something like ten exploratory wells for each successful find. The economists' concept would say that the economic cost of each firm's successes was that of all the wells it drilled, and that the cost of all discoveries to a whole economy was the full cost of all exploration undertaken, whether successful or not. One way of reflecting the economists' concept would be to require that a firm capitalize all of its exploration costs (whether successful or not), and then write off the sum total of these costs over the economic life of the successful finds. This can prove difficult, for much exploration is undertaken by consortia, formed *ad hoc* for a particular series of attempts. The alternative actually followed under cost depletion was to permit the deduction of the specific costs of each successful well against the revenues therefrom, while the costs of dry holes simply became losses, to be written off against income from any source. As

long as income was present against which to write off dry hole costs, they were in effect, under this treatment, allowed to be expensed. This gave oil exploration – even under cost depletion – a certain tax advantage over other types of investment, whose total costs were required to be capitalized initially, then written off gradually over the asset's useful life (yielding, of course, a present value of total write-off equal to only a fraction of the costs when incurred).

This relatively modest tax advantage of oil exploration was greatly magnified when discovery depletion was introduced. Assume that the value of the successful wells is precisely equal to the total economic costs incurred in finding them, and that the latter include 80 per cent of dry hole costs. Under discovery depletion the total write-off would turn out to be 180 per cent of exploration costs. The successful wells would obtain over their economic life a write-off equal to their discovery value (100 per cent of total exploration costs), while the 80 per cent of exploration costs incurred on what turned out to be dry holes would end up as losses written off against other income from any source.

The possibility of writing off more than 100 per cent of costs is the nub of the debate that raged for decades concerning percentage depletion. For percentage depletion, at least as it was first introduced and as it prevailed for many, many years was simply discovery depletion in another guise. It was introduced in 1926 precisely for the purpose of breaking a log-jam of litigation over the appropriate value to set on newly discovered wells. The precise percentage ($27\frac{1}{2}$ per cent of the gross value of oil or gas at the wellhead) was chosen so as to approximate the relationship between actual depletion allowances (based on legally sanctioned discovery values) and the production values to which they applied.

The tragedy of the depletion story did not really begin to unfold until much later. At the time discovery depletion was introduced, the US corporation income tax rate was 12 per cent; over the decade of the 1920s it oscillated between 10 and 13.5 per cent. At such tax rates even the double deduction of the cost of capital assets would have relatively little effect. If the tax consequence of the extra deduction is considered to be like a subsidy, its effects would be similar to those of a subsidy which caused the equilibrium rate of return to be about 9 per cent in investments in oil exploration if it would normally be 10 per cent in other activities – hardly a gross distortion.

The problems came later as tax rates first crept, then zoomed upward to finance a growing government and finally a major war effort. The corporation income tax rate was 52 per cent for many years following World War II; gradual reductions brought it to 46 per cent in 1981. Unfortunately, the same tax provisions have very different incentive effects at a 52 per cent rate than at one of 12 per cent or so.

A simple way of representing an equilibrium situation of a firm with respect to investment in a particular type of capital asset is to equate, at the after-tax rate of discount (r) the costs of the asset to the firm (net of tax offsets, if any) with the present value of the net-of-tax income stream produced by the asset. Let the present value (at the discount rate r) of the gross-of-tax income stream produced by the asset be Y. Let the asset's cost be C_a; this is also the amount which will be written off over time in the form of depreciation allowances. The present value of those allowances will be dC_a; d is a positive fraction, which is smaller the larger is r, the longer is the period over which depreciation is spread, and the later within that period the allowances are concentrated. The equating of costs with benefits yields

$$C_a = Y - t(Y - dC_a), \tag{1}$$

or

$$C_a = Y(1 - t)/(1 - \mathrm{d}t), \tag{2}$$

where t is the applicable tax rate.

The effect of discovery depletion was to permit the writing-off of the full value of the successful wells, while at the same time allowing the writing-off of dry-hole costs. If β is the fraction of total exploration costs represented by dry holes, the equilibrium investment under discovery depletion would be characterized by

$$C_d(1-\beta t) = Y(1-t)/(1-\mathrm{d}t). \qquad (3)$$

Here C_d is the cost which would characterize an oil exploration investment producing a gross-of-tax income stream whose present value (discounted at the rate r) is Y, where the fruits of the investment are subject to discovery depletion, and the costs of dry holes are written-off against other income that is taxable at the normal rate, t.

It is easily seen, comparing (3) with (2) that

$$C_d\ /C_a = 1/(1-\beta t). \qquad (4)$$

This means that under discovery depletion, if (as was approximately the case) dry-hole costs amount to some 80 per cent of exploration costs, a marginal investment producing a given income stream would have costs equal to 1/(1 – 0.096) times those of an ordinary corporate investment producing a similar income stream when the tax rate was 12 per cent; this same equilibrium at the margin would with a 52 per cent tax rate be generated by an investment whose costs were 1/(1 – 0.416) times those corresponding to an ordinary investment producing the same income stream. The mere upward drift of tax rates, then, magnifies the distortion from a fraction of just a little over 10 per cent to a fraction of more than 2/3.

Percentage depletion differed from discovery depletion in specifying that the depletion allowance should be a specified fraction ($27\frac{1}{2}$ per cent for oil and gas) of the gross value of the product as it emerged from the wells (or mi^2). Gross value of product differs from the gross income attributable to capital by the amount of labour and materials costs involved in the extractive process. These costs are not great for oil and gas; for a typical well, the depletion allowances averaged about 35 per cent of the cash flow attributable to capital when the statutory depletion rate was $27\frac{1}{2}$ per cent of a broader income concept. Denoting this observed fraction (depletion allowance/cash flow attributable to capital) by p, and by C_p the costs which would yield the after-tax rate r under percentage depletion treatment, we have

$$C_p(1-\beta t) = Y - t(Y - pY), \qquad (5)$$

or

$$C_p = Y(1-t+tp)/(1-\beta t). \qquad (6)$$

Investments under percentage depletion (with expensing of dry-hole costs) and under ordinary income taxation can be compared by looking at the relative amounts of cost that it would be barely worthwhile to incur in order to produce similar income streams under the two tax treatments. This is given by equation (7)

$$C_p\ /C_a = \begin{array}{l}[(1-t+tp)(1-dt)] \\ \times/[(1-\beta t)(1-t)].\end{array} \qquad (7)$$

For $p = 0.35$, $d = 0.65$ (an approximate figure for oil and gas wells under cost depletion when $r = 0.1$), and $\beta = 0.8$, this ratio comes to 1.07 for $t = 0.12$ and to 1.62 for $t = 0.52$. Ironically, the great bulk of the economic inefficiency produced by percentage depletion came long after it was enacted, as a result of tax rate changes that were debated, passed and signed into law without even the most cursory consideration being given to their consequences in magnifying the anti-economic incentive effects of percentage depletion.

Various factors have led to the progressive reduction of these anti-economic incentives. Beginning with the Tax Reform Act of 1969, the statutory rate of percentage depletion was reduced. Indeed, after the 1974 oil crisis, percentage depletion was eliminated for integrated firms, and after the crisis of 1979 an (ostensibly transitory) windfall profits tax was enacted. In addition, the percentage depletion rate for independent producers and royalty owners, and for secondary and tertiary production, was reduced in stages to 15 per cent as of 1984. Finally, the corporation income tax rate itself was reduced, reaching the level of 46 per cent in 1979, and accelerated depreciation was authorized for a wide range of assets of ordinary (non-extractive) businesses, consequently raising the value of d.

Some notion of the prospective level of the relative incentive to oil exploration, as of this writing, can be obtained by assuming $d = 0.8$ (compared with 0.65), $p = 0.2$ (compared with 0.35), and $t = 0.46$ (compared with 0.52). This

calculation yields a value for C_p/C_a of 1.17 (without taking any account of the transitory windfall profits tax which, so long as it remains in effect, further reduces the incentive in question).

In a real sense, the saga of percentage depletion, so far as the oil and gas industry of the United States is concerned, may be said to have passed into history. Once a source a major economic distortion and of bitter political battles, it has become (for oil and gas) a comparatively innocuous piece of special-interest incentive legislation.

But that is not the whole story. Percentage depletion only began with oil and gas in 1928. Coal, metals, and sulphur were added in 1932; fluorspar, rock asphalt, and ball and sagger clay came into the list in 1945; other additions were made in 1944, 1947, and 1951 (when even sand, gravel, slate, and stone came to be included); and finally, in 1954, percentage depletion treatment was extended to 'all other minerals' (with a few stated exceptions).

The incentive works in a significantly different way for most other minerals than for oil and gas, because of the great importance of a continuing process of exploration and discovery in the latter case. The analysis presented above for oil and gas treats the exploration activity as the fundamental act of investment, and the production process as simply the reaping of its fruits. For minerals like coal, sulphur, clay, gypsum, and many of the metals, reserves are well known and exploration is an insignificant element in the economic picture. In these cases, assuming reserves to be large and the true loss of property value through depletion to be small, the effect of percentage depletion allowances is substantially equivalent to that of a subsidy at a rate equal to the depletion rate times the applicable rate of tax. Thus, for many of the products subject to it, percentage depletion works as a subsidy to extraction rather than exploration.

See Also

- ▶ Exhaustible Resources
- ▶ Natural Resources
- ▶ Neutral Taxation
- ▶ Taxation of Capital

Bibliography

Agria, S.R. 1969. Special tax treatment of mineral industries. In *The taxation of income from capital*, ed. A.C. Harberger and M.J. Bailey. Washington, DC: The Brookings Institution.

Gravelle, J.C. 1985. Effective federal tax rates on income from new investments in oil and gas extraction. *The Energy Journal* 6(Special Tax Issue): 145–153.

Harberger, A.C. 1955. The taxation of mineral industries. In *Federal tax policy for economic growth and stability*, ed. U.S. Congress and Joint Economic Committee, 439–449. Washington, DC: GPO.

Harberger, A.C. 1961. The tax treatment of oil exploration. In *Proceedings of the second energy institute*, 256–269. Washington, DC: The American University.

Robinson, M.S. 1983. *Essays on the taxation of oil and natural gas*. Doctoral Dissertation, Stanford University.

Steiner, P.O. 1959. In *Tax revision compendium*, vol. II, ed. US House of Representatives, Committee on Ways and Means, 949–966. Washington, DC: GPO.

Wright, D. 1976. *The taxation of petroleum production*. Doctoral Dissertation, Havard University.

D

Deposit Insurance

Stephen G. Cecchetti

Abstract

The purpose of deposit insurance is to ensure financial stability, as well as protect the interests of small investors. But with government guarantees in hand, bankers take excessive risks, driving up the chances of failure. Evidence suggests that these schemes increase rather than decrease the probability of financial crises. There is a good chance that deposit insurance does more harm than good. This article surveys the rationale for and history of deposit insurance, and discusses its consequences and possible alternatives.

Keywords

Assets and liabilities; Asymmetric information; Bagehot, W.; Banking crises; Banking industry; Deposit insurance; Excessive risk taking; Federal Reserve System; Financial intermediaries; Financial market contagion;

Great Depression; Lender of last resort; Moral hazard; Non-bank financing mechanisms; Risk

JEL Classifications
E5

People living in countries where bank deposits are insured would never question the wisdom of an explicit insurance scheme. The idea that their savings are protected by a government-backed guarantee is something they simply take for granted. Only some crazy economist would ask whether deposit insurance makes sense. Well, does it? Surprisingly, the evidence is that it may not. Deposit insurance, which is supposed to stabilize the financial system, may do more harm than good.

This article examines the nature of deposit insurance by answering the following series of questions: (*a*) What do financial intermediaries do that warrants government intervention? (*b*) What is the history of deposit insurance? (*c*) Does deposit insurance do what it is designed to do? And (*d*), are there any alternatives?

Financial Intermediaries, Banks, and Bank Runs

The term 'financial intermediaries' encompasses a large set of institutions that include depository institutions as well as insurance companies, securities firms and pension funds. The first of these – what we all call 'banks' – are both the most commonly known to individuals and provide the broadest array of services. They pool savings, accepting resources from a large number of small savers in order to provide large loans to borrowers; provide access to the payments system, so that individuals can make and receive payments; provide liquidity, allowing depositors to transform their financial assets into money quickly and easily at low cost; and diversify risk, giving even the smallest saver a mechanism for diversification.

To appreciate the importance of financial intermediaries, consider what it would be like without them. If banks didn't exist, all finance would be direct, with borrowers obtaining funds straight from the lenders. Such a system would be costly and ultimately ineffective. It would be so difficult and expensive for borrowers and lenders to find each other, and then to come to agreement over the terms of a loan, that it is unlikely there would be any transactions at all. And without a financial system to transfer funds from savers to investors, there would be no economic development. The world would be a very different place.

Because of the services they provide, banks face a risk that other financial institutions (and industrial firms) do not. They are vulnerable to runs. Here's why. Banks issue liquid liabilities in the form of short-term demand deposits, and hold illiquid long-term assets, structured as securities and loans. The bank promises all its depositors that, if they want the entire balance of their checking account, they just have to come and ask. If a bank has insufficient funds to meet requests for withdrawal on demand, it will fail.

Banks not only guarantee their depositors immediate cash on demand; they promise to satisfy depositors' withdrawal requests on a first-come, first-served basis – what is called a 'sequential service constraint'. This commitment has important implications. Suppose depositors begin to lose confidence in a bank's ability to meet their withdrawal requests. True or not, reports that a bank has become *insolvent* can spread fear that it will run out of cash and close its doors. Mindful of the bank's first-come, first-served policy, panicked depositors rush to convert their account balances into cash before other customers arrive. Such a *bank run* can cause a bank to fail. Importantly, if people believe that a bank is in trouble, that belief alone can make it so.

While banking system panics and financial crises can result from false rumours, they can also come about for more concrete reasons. Widespread downturns in economic activity drive down the value of loans and securities, so bank capital (the difference between assets and liabilities) falls. If things get bad enough, banks become insolvent and fail. A big economic downturn can put the entire financial system at risk. Gorton (1988) reports that significant contractions are associated with all seven of the severe financial

panics in the United States that occurred between 1871 and 1914.

In a market-based economy, the opportunity to succeed is also an opportunity to fail. It would be natural to dismiss bank failures as analogous to the closing of an unpopular restaurant. But, while individual banks should be, and are, allowed to fail, the fact that banks are dependent on one another (in a way that restaurants are not) means that when one bank fails it puts others at risk.

Banks are linked both on their balance sheets and in their customers' minds. In recent years in the United States, inter-bank loans make up roughly four per cent of bank assets – an amount that represents almost half of bank capital. If one bank fails, it could put the system at risk. Information asymmetries are the reason that a depositor run on a single bank can turn into a bank panic that threatens the entire financial system. Most of us are not in a position to assess the quality of a bank's balance sheet. So, when rumours spread that a certain bank is in trouble, depositors everywhere begin to worry about their own banks' financial condition. Concern about even one bank can create a panic that causes profitable banks to fail, leading to a complete collapse of a country's banking system. Bank failure is contagious.

All of this leads to the following conclusions. Not only are individual banks fragile and vulnerable to runs, but the entire banking system is prone to panics. Contagion creates an externality that provides the economic justification for government intervention in the system.

Deposit Insurance and the Government Safety Net

Government officials intervene in the financial system both to protect small investors and to ensure financial stability. They do it with two related tools: the lender of last resort, where a central bank that can issue liabilities without limit provides loans to banks that are illiquid but not insolvent; and deposit insurance.

History reveals that the presence of a lender of last resort significantly reduces, but does not eliminate, bank panics. The series of three bank panics in the United States during the Great Depression of the 1930s, described in Friedman and Schwartz (1963), is one example of a failure of this sort. The Federal Reserve System was in place and had the capacity to operate as a lender, but did not.

The first national deposit insurance scheme was enacted by the US Congress in 1935 as a direct response to the bank panics in the 1930s. White (1995) sets out the history, noting that the debate was contentious, and that the stated purpose of deposit insurance was to stabilize the banking system. As surprising at it may seem from a modern perspective, investor protection per se was not the point.

When one thinks about deposit insurance, it is important to keep in mind that no private fund can be large enough to withstand a system-wide panic. Only the fiscal authority (possibly combined with the central bank) has the necessary resources.

For decades the US system was nearly unique. In 1974 only 12 countries had explicit national deposit insurance systems. Explicit deposit insurance is a phenomenon of the last quarter of the 20th century, when it became a part of the generally accepted best-practice advice international organizations gave to developing countries. Demirgüç-Kunt and Kane (2002) report that by 1999 the number of countries with deposit insurance had risen to 71 (with the insurable limits ranging up to more than eight times a country's per capita GDP). Prior to this, most systems were implicit, whereby depositors would exert their substantial political influence to force fiscal authorities to supply unlimited deposit guarantees in the event of a bank failure. This is all somewhat surprising, given the obvious political appeal of any system that has no immediate budgetary outlay associated with it. What politician wouldn't want to make an apparently costless promise to protect the bank deposits of his or her constituents?

Does Deposit Insurance Work?

In their classic theoretical treatment of deposit insurance, Diamond and Dybvig (1983) show that, if

self-fulfilling depositor runs result from information asymmetries, then government-supplied insurance can improve social welfare. But at what cost?

Insurance changes people's behaviour. Protected depositors have no incentive to monitor their bankers' behaviour. Knowing this, a bank's managers take on more risk than they would otherwise, since they get the benefit of risky bets that pay off while the government assumes the costs of the ones that don't. In protecting depositors, then, the government creates moral hazard. This is not just a theory. In 1980, the deposit insurance limit in the United States was raised to $100,000, four times its earlier level. Over the following ten years, several thousand depository institutions (banks and savings and loans) failed. That was more than four times the number that failed in the first 46 years of explicit deposit insurance. While a vast majority of the institutions that failed in the 1980s were small, the cost of reimbursing depositors exceeded 3 percent of one year's GDP. The bill was ultimately paid by US taxpayers.

The problem of excessive risk taking did not stop with the resolution of the 1980s crisis. Today, the US banking system's assets are worth between 10 and 12 times their equity. In the 1920s, this same leverage ratio was closer to four. Industrial firms typically have leverage that is half that lower number. In other words, deposit insurance has driven up leverage in banking. And with the increase in leverage comes an equal increase in risk (as measured by the standard deviation of returns).

So, in an attempt to solve one problem, deposit insurance created another. And to combat bankers' excessive risk taking, governments were forced to set up regulatory and supervisory structures. Among other things, there are now constraints on the assets banks can hold, rules governing the minimum levels of capital that banks must maintain, and requirements that banks make public information about their balance sheets. Supervisors have to enforce the detailed web of regulations.

Does this complex mechanism actually work to stabilize the financial system? The evidence is not encouraging. Demirgüç-Kunt and Kane (2002) summarize international research and conclude that explicit deposit insurance actually makes financial crises more likely. When countries have either implemented a new scheme or expanded an existing one, the probability of crises has increased.

To make matters worse, the creation of deposit insurance retards the evolution of non-bank financing mechanisms. Cecchetti and Krause (2005) find that countries with more extensive deposit insurance schemes tend to have both smaller financial markets and a fewer publicly traded firms per capita. To put it bluntly, deposit insurance is bad for financial development, and may be bad for real economic growth.

Are There Alternatives?

So, if deposit insurance schemes do more harm than good, what should we do to stabilize the financial system? The natural response of an economist is to use the price system. Measure how risky a bank's balance sheet is, and set its deposit insurance premiums accordingly. Beginning in 1991, the US Federal Deposit Insurance Corporation did implement a risk-based premium structure. But this is extremely difficult to do well. Banks can always find ways to evade detailed rules, exploiting the system to reduce the prices they pay. In the end, this is not a solution.

There are three other options. We could implement changes that further restrict the assets held by banks, eliminating their asset transformation function. We could increase our reliance on the central bank's lender-of-last-resort function. Or it may be possible to design a scheme to ensure that large depositors will impose discipline on the risk taking of bank managers.

Proposals for narrow banking are in the first category. A narrow bank is an institution that holds only a very limited set of very low-risk, highly liquid assets, such as short-term government securities. Since insolvency is impossible for such an institution, liability holders would not have to worry about the quality of the narrow bank's assets, and there would be no fear of a run. Deposit insurance would be unnecessary.

Second, it may be possible to address the potential for systemic bank panics by improving the effectiveness of the lender of last resort. In

1873, Walter Bagehot suggested that, in order to prevent the failure of solvent but illiquid financial institutions, the central bank should lend freely on good collateral at a penalty rate. By lending freely, he meant providing liquidity on demand to any bank that asked. Good collateral would ensure that the borrowing bank was in fact solvent, and a high interest rate would penalize the bank for failing to manage its assets sufficiently cautiously. While such a system could work to stem financial contagion, it has a critical flaw. For it to work, central bank officials who approve the loan applications must be able to distinguish an illiquid from an insolvent institution. But during times of crisis computing the market value of a bank's asset is almost impossible, since there are no operating financial markets and no prices for financial instruments. Because a bank will go to the central bank for a direct loan only after having exhausted all opportunities to sell its assets and borrow from other banks without collateral, its illiquidity and its need to seek a loan from the government draw its solvency into question. Officials anxious to keep the crisis from deepening are likely to be generous in evaluating the bank's assets, and to grant a loan even if they suspect the bank might be insolvent. And, knowing this, bank managers will tend to take too many risks.

Finally, we could require that banks issue subordinated debt. These are unsecured bonds, with the lender being paid only after all other bondholders are paid. Someone who buys a bank's subordinate debt has a very strong incentive to monitor the risk-taking behaviour of the bank. The price of these publicly traded bonds then provides the market's evaluation of the quality of the bank's balance sheet and serves to discipline its management.

By eliminating the accountability of bank managers to their depositors, deposit insurance encourages risky behaviour. So, while financial stability is clearly in the public interest, deposit insurance may not be.

See Also

- ▶ Bagehot, Walter (1826–1877)
- ▶ Financial Intermediation
- ▶ Financial Structure and Economic Development
- ▶ Moral Hazard
- ▶ Risk

Bibliography

Bagehot, W. 1873. *Lombard street: A description of the money market*. London: Henry S. Kin & Co.

Cecchetti, S. 2006. *Money, banking, and financial markets*. Boston: McGraw-Hill Irwin.

Cecchetti, S., and S. Krause. 2005. Deposit insurance and external finance. *Economic Inquiry* 43: 531–541.

Demirgüç-Kunt, A., and E. Kane. 2002. Deposit insurance around the globe: Where does it work? *Journal of Economic Perspectives* 16(2): 175–195.

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

Friedman, M., and A. Schwartz. 1963. *A monetary history of the United States: 1867–1960*. Princeton: Princeton University Press.

Gorton, G. 1988. Banking panics and business cycles. *Oxford Economic Papers* 40: 751–788.

White, E. 1995. *Deposit insurance*. Policy research working paper No. 1541. Washington, DC: World Bank.

Depreciation

Gautam Mathur

Abstract

Depreciation estimates the decline in the value of capital over time. It is highly important to capital accounting, since the rate of dividend is calculated as the ratio of the surplus to the current value of assets. The causes of the depreciation of equipment are twofold: its productivity may fall with age, and over time its expected remaining earning life is shorter. Depreciation is taken to be the difference between gross and net investment but total new employment is given by gross investment: the physical counterpart of replacement (of equipment or manpower) is needed give the full picture.

Keywords
Champernowne, D. G.; Depreciation; Kahn, R. F.; National income statistics; One-hoss shay; Robinson, J. V.; Sraffa, P.; Torrens, R.; Von Neumann, J.

JEL Classifications
G0

Depreciation estimates the decline in the value of capital as a result of ageing, its maximum value being near its age of manufacture and its minimum value when it is dismantled and sold as scrap. It is of great importance to capital accounting, for the rate of dividend is calculated as the ratio of the surplus to the current value of assets. The reduction in value of equipment comes about from two causes – firstly that its productivity may fall with age; and secondly that, as time advances, the expected remaining earning life of the plant is shorter. Hence, the capitalized value of the present value of expected future stream of quasi-rents from an old piece of equipment is smaller for any given rate of interest than for a younger machine.

'One-Hoss Shay' Assumption

The influence of declining productivity over time may be eliminated by assuming a 'one-hoss shay' type of equipment, which keeps its efficiency constant over its service life and falls to pieces at the end. However, the product of a process is not only its current output but the stock of equipment which remains at the end of the production period – as stressed by von Neumann (1933) and by Sraffa (who in 1960 referred to Robert Torrens as having insisted in the years 1818 and 1821 on its being considered as a part of output).

On account of the shorter remaining service life of equipment at the end of a period (and the consequent smaller number of expected items of quasi-rent in its stream of earnings), there is lesser value of capital remaining at the end of a production period – a so-called 'year'. This reduction in value of a stock output affects adversely the productivity in value terms (even with 'one-hoss shay' equipment) and it measures the depreciation. There is, therefore, an aggravated tendency of the value of capital embodied to fall as the plant is older.

Shape of Decline in Valuation Curve

In a straight line approximation, depreciation is taken as constant in absolute amount per year. In a formula using the exponential concept depreciation is at a constant rate; hence, the fall in value is more when machines are younger and higher priced, than when they are older – as in radioactive decay, that is, it indicates a curve convex to the origin. But depreciation is at higher rates for older capital in service – not as would be given by an exponentially falling value of equipment at a constant rate with respect to time. When there is a rising rate of reduction of value, it makes the decline more than exponential as the machines are older, and yields a steeply falling value towards the end of the service life, that is, it yields a curve with respect to time which is concave to the origin. The straight line approximation of value of capital (with respect to its age) which is used in some calculations is thus wide of the mark; and even the exponentially falling value according to a constant rate of reduction does not make the value of old machines decline sufficiently markedly.

In a Sraffa or von Neumann valuation of capital (of different ages taken as different commodities) this decline is well brought out automatically, for differences between value of the commodity called 'equipment t years old' and the one called '$t + 1$ years old', increases as t becomes larger.

This aspect of the Sraffa system (1960) was not known to Joan Robinson or to Professor Richard Kahn and D.G. Champernowne in 1954 when the text of *Accumulation of Capital* (1956) was being finalized – especially its Mathematical Appendix (to a part of which the latter two had contributed as authors, their names appearing in the original printed text). It is all the more remarkable that it was discovered that, in the measurement of value

of ageing equipment, one could strike upon another useful device – of balanced age composition of capital.

Balanced Age Composition of Capital

In demographic studies as part of the subject of manpower, it is well known that, for a population of human beings growing at g per cent, there are higher numbers of children of age t in comparison to those a year older (of age $t + 1$) by the factor $(1 + g)$. The same principle can be applied to a population of plants, and we can derive a universe of plants ordered according to their ages in this particular manner. One can try to ascertain what the number of plants in a cohort of each age is and the value of capital embodied in each cohort.

The value of plant at the centre of gravity of the age-composition pyramid may then be used as the standard unit of measurement of the value of a plant of any particular age. The result would be in agreement with the well-known, but rather mystifying, Kahn–Champernowne formula of the reciprocal value of a new plant in terms of value of the plant of average age. This reciprocal will be called a K-C unit in honour of those two authors who worked out the said formula.

Kahn–Champernowne Units of Measurement

In a generalized version of this concept, as the set of pieces of capital of constant physical productivity and of balanced age composition growing exponentially at a steady rate, keep the composition in terms of relative sizes of cohorts constant; hence the value of the average plant does not change. This is the justification of the K–C units.

In terms of a balanced age composition of equipment (with T years expected service life since its manufacture), a piece of equipment t 'years' old is replaced at the end of t years by a piece which was $t - 1$ years old in the beginning of the year. Except for this replacement by equipment which is now of the same age as the piece it substitutes, there is no depreciation visible in the physical system or its statistical depiction.

Redundancy of Gross and Net Concepts

It is to be remembered that Joan Robinson had correctly realized that depreciation was not a physical phenomenon but a notional or value one. The implication of depreciation not being a physical phenomenon in terms of effect upon the concepts of gross and net investment had to wait until the von Neumann model was integrated (in 1960) with the Robinsonian golden-age system. In traditional analysis the system is depicted as z machines (newly produced and added) in a factory, and at the same time another z machine rendered inoperative (by completion of their natural life). But the net investment is not an act of accretion–depreciation in physical terms; for the machines added through current investment are new ones and the depletion is of old machines – and it makes no sense if value measurement were not resorted to for calculating the excess of accretion over depreciation.

The balanced age composition is a device by which one can realize that in a von Neumann system as a growing economy m machines of age t years exist and $m(1 + g)$ of age $t - 1$ years are automatically substituted a year later by $m(1 + g)$ machines – also now of t years age. The stock as well as each age cohort grows at rate G, and depreciation of value by ageing is exactly counterbalanced by that much capital of erstwhile younger age and erstwhile higher value (but now of the same age and the same value as the m plants at the beginning of the year) replacing it. In addition one has mg times more machine of age t. The total stock grows at a given rate of growth, and depreciation is also compensated for exactly, for $m(1 + g)$ is equal to m for replacement, and mg for accumulation for each age cohort.

In Sraffa–von Neumann analysis (as a simplified purposive model combining the two general constituents of those two models and integrating the resultant with the Robinsonian golden-age system), this fact was noticed in 1961, and it was

discovered that in a state of steady growth and balanced age composition depreciation of a stock of inputs in terms of writing down of value of equipment (due to ageing) is a dispensable concept (Mathur 1965). Each age cohort is replenished exactly by an age cohort from within the system in physical numbers and value, and there is nothing to be written down of any piece of equipment by a chartered accountant at the end of the year. The pieces of equipment of each age are higher by the rate g and valuation is required for finding out cumulative accumulation of equipment in each age cohort. In a body of equipment of balanced age composition as the value of capital of different ages differs by the amount of depreciation, the concept of depreciation is required for measuring aggregate accumulation due to ageing, not for decumulation due to ageing as was required in the traditional concept.

It is because of the total absence of writing down of value of stock of any age that it was realized that there is no concept of gross or net necessary in such a reckoning (Mathur 1965), and depreciation is important not as the difference between gross and net investment, but as the difference of value of an older machine in relation to a younger one for purposes of measuring accumulation (of positive-age equipment from within the firm and of new equipment from the manufacturers). It is only when the age composition is grossly unbalanced – as for newly established firms – that it may be necessary to use depreciation in the traditional sense of writing down value of stocks. But in that case measurement of depreciation or of amount to be written off is itself a procedure not entirely free from logical doubts.

Depreciation and Maintenance

In manpower-employment terms, total new employment is given by gross investment and not by net investment, because the amount spent on activities of maintaining capital intact (repairing, renovating) also creates employment, and not only the building of new capital. Hence, in national income statistics, it is gross investment which creates manpower employment and not net investment by itself. The difference between gross and net is taken to be depreciation, but in manpower terms it does not so follow – for employment created for maintenance of a machine (like a sealed unit) might be very low, and yet the reduction of its value year by year very high due to ageing. When viewing manpower statistics, the activity of operatives of a particular type ought to be supplemented by statistics of valuation (Mathur 1983). While figures in terms of counting heads are important for a physical count, greater economic significance would be acquired if the productivity of each type of human equipment were determined and its true value calculated in K–C units with respect to a balanced age composition and age structure (Mathur 1964). But depreciation in value terms alone without the physical counterpart of replacement (of equipment or manpower) also tells us an incomplete story, and only valuation and quantification (in physical terms) together give the full picture.

See Also

► Amortization

Bibliography

Champernowne, D.G., and R.F. Kahn. 1956. The value of invested capital. Mathematical Note appended at the end of Robinson (1956).

Mathur, G. 1964. The valuation of human capital for manpower planning. *Applied Economic Papers (Hyderabad)* 4 (2): 14–35.

Mathur, G. 1965. *Planning for steady growth*. Oxford/New York: Basil Blackwell/Augustus Kelly.

Mathur, G. 1983. Web of inequity. Presidential Address to the Silver Jubilee Session of the Indian Labour Economic Association, Lucknow, 1982. *Indian Journal of Labour Economics*, Sec. VIII.

von Neumann, J. 1933. A model of general economic equilibrium. *Review of Economic Studies* 13 (1945): 1–9.

Robinson, J. 1956. *The accumulation of capital*. London: Macmillan.

Sraffa, P. 1960. *Production of commodities by means of commodities*. Cambridge: Cambridge University Press.

Depressions

Sidney Pollard

Modern economies have a tendency to grow, but their growth is irregular. The periodic variations of most indicators and even occasional reversals of direction inevitably prompt the question whether these changes are of a random nature, or indicate broader sweeps, swings or cycles. The trade cycle, or *Juglar* cycle (named after Clément Juglar) of an average duration of about seven years appeared frequently enough in the course of the nineteenth century to be generally recognized as a cyclical phenomenon. Other possible recurrent movements were the inventory *Kitchin* cycles (Joseph Kitchin) of 2–3 years, and the longer *Kuznets* (Simon Kuznets) swings of 20–25 years, indicating alternate phases of European and American long-term investment and of transatlantic migration.

By general consent, the trough phases of none of these are normally termed depressions. That term is reserved for longer periods of more serious adversity on an international scale, in particular the Great Depressions of *c*1873–96 and of the 1930s. By analogy, the distressed years following the Napoleonic wars and the years since the downturn of 1973 have also been included in that category. In view of the fact that these seem to have occurred at fairly regular intervals, attempts have not been lacking to explain them as part of an alternating movement also.

The first major theory of long swings was that of N.D. Kondratieff, who published his findings in 1922–8 (Kondratieff 1935), following some earlier Marxists, notably J. van Gelderen in 1913. Kondratieff's cycles are essentially price movements, including swings of other indicators expressed in money, such as wages and foreign trade. His depression periods of 1810/17–1844/51, and 1870/5–1890/6, were periods of falling prices. In addition to prices, Kondratieff also considered variations in some key industrial products such as coal and iron, and noted the particularly sharp deterioration of agriculture in the depressions. According to him, upswings were preceded by technological innovations and rises in gold production, while wars and revolutions also take place then; in the depressions their effects fade. Although essentially presented as a set of empirical findings, Kondratieff's swings may be interpreted as investment cycles of long duration.

Kondratieff's views were rejected in his own country, and found little echo in the west at first. They were introduced to a wider audience by being incorporated by J.A. Schumpeter (1939) in a grand concept of a cyclical development of modern capitalism. In this scheme, each Kondratieff consisted of six Juglars, the depression phase containing three Juglars in which the peaks tended to be weaker, the troughs more pronounced than in the upswing. Schumpeter's periodization was similar to Kondratieff's, except that the first turning point was set in 1842–3 rather than 1844–51. His depressions were periods which lacked the stimulus of what to him was the major driving force of industrial capitalistic society, the appearance of bunched major innovations, such as railways, steel or electricity. Each depression was introduced by a major crisis, including a high level of unemployment, following an exceptional investment boom or wartime expansion. Schumpeter also did not fail to notice the serious agricultural price fall and the pervasive agrarian distress that was part of each depression period.

The notion that bunched innovations are associated with recurrent depressions, that their introduction leads to loss of jobs, and that the depression can be reversed, in the end, only by a renewed massive introduction of a new technology, has also reappeared in the depression of the 1970s, in such theories as those propagated by Gerhard Mensch (1975), Ernest Mandel (1975) and Robert Boyer (1979). Mensch noted that the previous depressions were started by panics, in 1825, 1873 and 1929 respectively, among those who had the responsibility for economic policy, but he held that their basic cause was the diminishing returns on technical innovations, as

major breakthroughs are completed, leaving room for only minor adjustments of ever lesser significance, while at the same time on the demand side there is a diminishing rate of growth of marginal utility, as markets become saturated with the new commodity or new productive technique. A renewed upward movement would therefore require a new set of major innovations.

Boyer's theory is a complex one. According to him, expansion periods are periods of 'extensive' accumulation (which may roughly be translated as investment), when capital of a known kind is spread into new areas. By contrast, depressions are periods in which opportunities for investment of the current type have become exhausted, and new techniques are being developed in what he terms 'intensive' accumulation. Since this has to be undertaken ahead of demand, it tends to be unprofitable and the general profit level therefore tends to fall until new structural parameters evolve, including new wage labour relations, competitive systems, capital relations, monetary and currency systems, and a redefined role of individual economies within the international division of labour. These then allow a renewed period of 'extensive' accumulation. Depressions are thus hinge periods of structural change between growth periods.

A Keynesian version of the Kondratieff was developed by W.W. Rostow in the 1940s (Rostow 1978). His depression period extended over 1815–48, 1873–96 and 1920–36, and they were thus more in line with the reference cycles, made up of numerous indicators which have been developed since then by other scholars. The main characteristics of these depressions were falling prices, especially agricultural prices, falling interest rates and low profits, while incomes shifted in favour of wage earners, or at least in favour of those who remained in employment. Their cause lay in declining employment opportunities for capital, or, expressed in different terms, a failure of investment to mop up all available savings. The preceding prosperity periods had been marked, in each case, by investment in long gestation projects such as railways, as well as by major wars with their positive employment effects. In the depressions, the earlier investments were bearing fruit in lower costs which contributed to the fall in prices. The effects of the major gold discoveries at the onset of each upward phase, in *c*1850 and in the 1890s, were not forgotten, and the possible contribution of their diminishing yields to the deflationary tenor of the age, measured against the growing demand for gold owing to expanding world transactions and the extension of the gold standard, especially after 1873, was noted. However, in view of the growth of paper credit, the relationship between the amount of gold mined annually and the world price level was clearly a complex one, while the coincidence of the gold discoveries just at the point when gold was at its most valuable, i.e. when prices were lowest, could not be ignored.

Not surprisingly, the same descriptive emphases as well as the same building bricks of explanations tend to recur in most of these cyclical theories. None of them, however, entirely succeeds in clarifying the question whether the depressions were part of an immanent, endogenous rhythm of world capitalist development, or whether they were in each case the result of a fortuitous conjunction of circumstances. The bunching of innovations, for example, which occurs repeatedly as part of the explanatory model, sometimes in addition linked explicitly to a succeeding phase in which investment opportunities are exhausted, might derive either from a 'law' of the workings of market economies, or they might have a wholly exogenous explanation. Such a 'law' might be based on mass psychology, the successive waves of optimism or pessimism among entrepreneurs; it might be based on economic functions, according to which innovations gain by being accompanied by others; or it might conceivably have some other inherent mechanism as its source, according to which each phase necessarily bears the seeds of the next phase within itself. On the other hand, the bunching might have exogenous causes, such as development breakthroughs in science and technology which make several related processes possible at once, the opening up of overseas mineral supplies, overseas markets or the peopling of free land in empty continents. Most theories on offer balanced carefully between these two possibilities without coming down on either side. Similar ambiguities

might be found in the other explanations offered, such as changes in relative costs and prices, the 'terms of trade', on an international scale. In any case, a mere three cycles at most were too few to provide certainty that endogenous swings or cycles were at work.

The alternative of looking at each of the depression periods as a unique historical event has therefore found many adherents. The first to be studied in any depth was the 'Great Depression' of 1873–96 (to use the terminal dates which today would command most support). It was particularly marked in Britain, where in consequence it has received the greatest amount of attention, and this, in turn, was associated with the fact that it coincided with a turning point, or 'climacteric', in British economic fortunes in international comparison. This was the period in which foreign manufacturing competition became a serious threat to Britain for the first time, and in which, therefore, a mood of general pessimism could spread more easily. Further, as modern research shows, the secular growth of the British economy did indeed experience a marked slow-down in those decades, showing that the instinct of contemporaries was sound. The continuation of the slow-down, bringing growth practically to a complete halt in the following upswing phase of 1896–1914 does, however, throw some doubt on the decline of the long-term growth rate as a characteristic of the 'depression' which allegedly came to an end in *c*1896. By contrast, both the USA and Germany, the other leading industrial nations, were then in a phase of rapid secular growth so that they were much better able to overcome the effects of a depression which by no means passed them by entirely.

The depression was ushered in by the collapse of what was perhaps the most expansive boom of the century (1871–3), particularly in the capital goods industries. It had been marked by much speculation and was followed by numerous bankruptcies. Overseas borrowing countries defaulted on their debts, home investments, made at a time when costs peaked, proved unprofitable, and the French indemnity to Germany, paid in part through London, further upset financial markets. In the period as a whole it was not only financiers who had a thin time but indeed all profit earners, though their plight may have been exaggerated, and interest rates were low. These low interest rates, part of the 'Gibson paradox', have frequently been used as a proof against the proposition that the depression was brought about by deflationary conditions which, in turn, derived from a shortage of gold, for in that case, interest rates should have been high.

In real terms, however, the 'depression' was far less clear-cut. Serious unemployment existed only in the late 1870s and in 1884–7, and even then it was mild by the standards of the 20th century. GNP failed to grow in three years only, and in each case the decline was marginal, never exceeding 0.5 per cent of the preceding year. It was, in fact, part of the persistent complaint of industrialists, before the Royal Commission on the Depression of Trade reporting in 1886 and elsewhere, that competition forced them to sell large quantities for which the unit profit margin had become exceedingly small. Real wages rose satisfactorily.

Even in agriculture, the sector which complained most loudly, the sharp decline in profits and in income was limited to the grain farmers. Here the combination of much reduced transport costs by rail and steamship, and much reduced production costs in the fertile lands above all of North America, but also of Russia, Australia and India, had led to a particularly rapid collapse in prices, to which home growers had not had time to adjust. This may be considered to be part of a long-term, and economically desirable process of an international division of labour in which Britain turned increasingly to producing manufactures and importing food. By contrast, producers of meat and dairy products, market gardeners and horse breeders did not fare too badly, benefiting rather than suffering from the cheap grain imports.

Thus it was essentially profit earners, and financiers for whom the period was one of depression. The standard of living of the rest of the population went up satisfactorily. But it is precisely profit earners and financiers who control the press and shape 'informed' economic opinion. Perhaps they have misled us. The 'Great Depression', in the view of some, was a myth (Saul 1969).

The depression of the 1930s has possibly an even stronger claim to be considered a unique event, with its own particular explanation for which no parallels can be found in other periods. Marxists saw it as heralding the frequently predicted final end of capitalism, while other ascribed its severity to the direct and indirect consequences of the war. Its terminal dates are not entirely unquestioned. Most indices show a more drastic fall in 1920–21 than in 1929–30. Moreover, the inflation in Germany and some other countries in the early 1920s was, in some ways, more devastating than anything that happened in the 1930s. At the other extreme, the trade cycle boom of 1937 was extremely weak and carried many signs of turning into a further severe depression, which was avoided only by the preparations for war. In more normal circumstances, the whole of the three Juglars, 1920–39 or even 1920–45, might have been taken as the depression, in parallel with the dating of 1873–96. However, such was the extent of their economic adversity, that it is only the four to six years following the New York stock exchange crash of autumn 1929 which are commonly referred to as the 'Great Depression' (van der Wee 1972).

The share collapse at the beginning was soon followed by other financial disasters. The failure of the Austrian Creditanstalt, by far the largest bank in the country, caused waves which also brought down some German banks. Britain, unable to stand the international drain on her gold reserve, went off the gold standard maintained precariously at high cost up to then, and was followed by many other countries. As the crisis deepened, American banks and finance houses went under in large numbers. Profits and interest rates fell. The British bank rate at two per cent ushered in an unprecedented period of 'cheap money'.

More significant for this depression than financial crashes was the collapse in the real world of production and trade, and above all the unprecedented unemployment. Unemployment of registered labour affected 15–30 per cent of the population at risk in many countries, while many more workers had removed themselves from the register, having lost all hope of finding work, and others worked short time only. Even after the onset of the recovery in 1933–4, a large hard core of intractable unemployment remained almost everywhere. Similarly unprecedented was the decline in output and in real national income over several years. The American GDP, for example, fell for four consecutive years, from an index of 163.0 in 1929 to 115.0 in 1933 (1913 = 100), or by almost 30 per cent; in Germany, the drop in the same years was from 121.6 to 102.0, or by 16 per cent; and even in the United Kingdom there was a six per cent drop over the two years 1929–31.

For several of the contributory causes of the depression and for its severity, the war could be held responsible. Thus it was in the war years, above all, that overseas countries had been encouraged to increase their food supplies to Europe where agricultural output suffered in the fighting, and when, afterwards, Europe returned to its previous output levels, growing world food surpluses began to burden world markets in the 1920s. Similar exceptional wartime demands had brought about an over-capacity in some industrial sectors also, such as shipbuilding, leading to exceptionally heavy unemployment there, but it was in primary products that the price fall was particularly severe in the depression, bankrupting overseas countries and their firms, though benefiting those Europeans who still had an income.

Among other consequences of the war were international war debts, the reparations imposed on Germany, and the creation of new, small, non-viable states in Europe. Debts and reparations bedevilled political relations, helped to radicalize the German electorate and inhibited the German chancellor Brüning from conducting a more vigorous reflationary policy in 1930. The hostility engendered then contributed to the protectionist and beggar-my-neighbour reactions of most governments to the depression so that trade was reduced, real costs rose, and the world became divided into several currency blocs, of the pound sterling, the dollar, the franc and the mark, severely restricting trade and payments between them. Britain, meanwhile, had become too weak to keep the world's exchanges in balance, as she had done in the 19th century, and the USA, which had emerged from the war as the leading economic power, was unwilling to accept that role, so that it went by default (Kindleberger 1973).

Lastly, the war had strengthened the trade unions everywhere, and had led to greatly extended state welfare provisions, including more extensive unemployment benefits. This led to a rigidity of wages downwards, which, according to some, prevented an early adjustment to the depression and helped to account for its exceptional severity. Such views, widely held at the time, were strongly opposed by Keynes, whose comprehensive theory appeared in print only in 1936, well after the worst of the depression was over. According to him, cutting wages would have done little or nothing to improve employment; what was wanted was the creation of new purchasing power, which in the final analysis would have to come from the Government.

Keynes's view on the depression was dominant in the following decades, though by no means accepted by all. In the 1960s, an alternative theory, developed by Milton Friedman and others in the USA, began to gain wider support. In their view, the money supply, which to the Keynesians played only a subordinate role via the rate of interest, occupied the centre of the stage. As it happened, as far as the cure for a deep depression occurring in a deflationary phase was concerned this view did not differ too widely from Keynes's. According to Friedman and his associates, the depression had been greatly aggravated, at any rate in the USA, by repeated reductions in the quantity of money made available to the system, whereas the economy could have been revived by a controlled monetary expansion.

It is in their relative assessments of cause and cure of the depression of the 1970s, which is marked by a concurrent monetary inflation, that Keynesians and Monetarists are totally at loggerheads. The depression itself was triggered by the floating of the dollar in 1973 and the drastic oil price rises engineered by the OPEC countries in 1973–4 and again in 1979, but these themselves were symptoms and consequences of a creeping inflation which had accompanied the remarkable world boom of 1945–73. Depressed conditions at a time of inflation, 'stagflation', could not really occur according to the Keynesians, and the fact that it did, helped to discredit them, at least for a time. The underlying inflation after the war was, no doubt, at least to some extent due to Keynesian-type employment policies on the part of most governments. It is the monetarist view that only severe cuts in the expansion of the money supply, bringing with them at least temporarily drastic increases in unemployment, offer a way out.

Inflation apart, the current depression bears some marked similarities to the experience of earlier ones. Among them there is the technological explanation of unemployment, this time looked for in the electronics field; there is the rise of economic nationalism, fostered by subsidies, hidden tariffs, and industries maintained as status symbols, especially by weaker economies; there is the rigidity explanation, blaming trade unions and social insurance schemes for the unwillingness to change occupations or reduce wages; there is, despite international agreements to the contrary, the use of currency manipulation to gain trading advantages; there is also a huge international indebtedness, which the debtors are unwilling or unable to pay, and the creditors unwilling or unable to forego. In duration, if not in severity, the current depression is likely to match its predecessors.

See Also

- ▶ Kondratieff Cycles
- ▶ Long Swings in Economic Growth
- ▶ Trade Cycle

References

Boyer, R. 1979. La crise actuelle: une mise au point en perspective historique. *Critiques de l'économie politique,* April–September, 5–113.

Brunner, K. (ed). 1981. *The Great Depression revisited*. Boston/The Hauge: Nijhoff.

Friedman, M., and A. Schwartz. 1963. *A monetary history of the United States*. Princeton: Princeton University Press.

Kindleberger, C.P. 1973. *The world in depression, 1929–1939*. Berkeley: University of California Press.

Kondratieff, N.D. 1935. The long waves in economic life. *The Review of Economic Statistics* 17(6): 105–115.

Mandel, E. 1975. *Late capitalism*. London: New Left Books.

Mensch, G. 1975. *Das technologische Patt. Innovationen überwinden die Depression*. Frankfurt: Fisher.

Rostow, W.W. 1978. *The world economy: History and prospect*. Austin/London: Texas University Press/ Macmillan.
Saul, S.B. 1969. *The myth of the Great Depression*. London: Macmillan.
Schumpeter, J.A. 1939. *Business cycles*. 2 vols. New York: McGraw-Hill.
Temin, P. 1976. *Did monetary forces cause the Great Depression?* New York: Norton.
van der Wee, H. (ed). 1972. *The Great Depression revisited. Essays on the economics of the thirties*. The Hague: Nijhoff.

Derived Demand

John K. Whitaker

Keywords

Cost functions; Derived demand; Elasticity of substitution; Hicks, J. R.; Marshall, A.; Robinson, J. V.

JEL Classifications

D1

The idea that the demand for intermediate goods is *derived* from the demand for the final goods they help produce is obvious and appealing. It was implied by Cournot (1838, pp. 99–116) and explicitly stated by Gossen (1854, pp. 31, 113) and Menger (1871, pp. 63–7). That the British classical school failed to make use of such a perspective – Mill's famous proposition that 'demand for commodities is not demand for labour' (1848, Book I, ch. 5) came close to denying it – was doubtless due to the strong emphasis placed on prior accumulation of capital as a prerequisite for production. But it was Alfred Marshall in his *Principles of Economics* (1890, pp. 381–93, 852–6) who introduced the term 'derived demand' and developed the concepts of the derived demand curve for an input and the elasticity of derived demand.

Marshall focused on a case in which a commodity is produced by the cooperation of several inputs, which are thus jointly demanded for the purpose, the demand for each being derived from the demand for the product. His formal analysis proceeded on the assumption that the inputs were all combined in fixed proportions (which might vary with the scale of output) although he suggested that the variable-proportions case would be similar.

A derived demand curve can be constructed for a selected input on the assumptions that production conditions, the demand curve for output, and the supply curves for all other inputs remain fixed, and that the competitive markets for output and all other inputs are always in equilibrium. The resulting derived demand curve can most easily be interpreted as the outcome of a hypothetical experiment. Make the selected input available, perfectly elastically, at an arbitrary price, y, per unit. Now ascertain, under the above conditions about the markets for output and other inputs, what quantity, x, of the selected input would be demanded. All other markets must be in equilibrium, and each seller or buyer must be optimally adjusted to the assumed terms of availability of the selected input. Repeating this experiment for different values of y would generate the inverse of the relationship between x and y, $y = f(x)$, whose graphical representation is Marshall's derived demand curve for the selected input. Bringing this demand curve into conjunction with the actual supply curve of the selected input will determine the actual equilibrium price and quantity for this input and thereby implicitly determine the actual equilibrium prices and quantities of output and all other inputs. But the point of obtaining the derived demand curve is not to permit such a two-stage determination of the actual equilibrium. It is rather to permit a simplified analysis of the effect of *changes* in the supply conditions of the selected input when supply conditions of other inputs, as well as technology and the demand conditions for output, remain unaltered.

Marshall invoked a simple example in which the final product, a knife, is obtained by joining costlessly a unit each of the two inputs, blades and handles. The derived demand curve for handles is then given by the rule that y, the derived demand price for x handles, is the demand price for x knives less the supply price for x blades.

Marshall analysed the conditions producing a low elasticity of derived demand for an input, a condition which would encourage supply restriction. The first condition, the lack of a good substitute, is already implied by the fixity of production coefficients. The second is that the demand for the final output be inelastic. The third, aptly described by Henderson (1922, p. 59) as 'the importance of being unimportant', is that expenditure on the input in question be only a small fraction of total production cost. The final condition is that cooperating inputs be in inelastic supply. These last three conditions ensure that a large rise in the price of the input will not raise product price much, that a rise in the product price will not reduce sales much, and that a reduction in sales and production will lower the cost of cooperating inputs substantially.

The next major contribution was that of Hicks (1932, pp. 241–6) who formally relaxed the assumption of fixed production coefficients. He analysed the consequences of input substitutability for a two input case with constant returns to scale in production, making use of his newly invented concept of the elasticity of substitution. His principal finding was that, to get a low elasticity of derived demand, 'It is "important to be unimportant" only when the consumer can substitute more easily than the entrepreneur', that is, only when the elasticity of demand for the product exceeds the elasticity of input substitution (1932, p. 246). This finding, which is not easily explained intuitively, has been the subject of intermittent controversy, aptly summarized and resolved in Maurice (1975). The extension of Hicks's analysis to the many-input case has been accomplished by Diewert (1971), using an elegant dual approach based on the cost function concept. However, modem theoretical work is more prone to work explicitly and symmetrically with complete systems of input demand equations for firm and industry.

More or less contemporaneously with Hicks, Joan Robinson (1933, chs. 23, 24) was studying the derived demand curve for an input in cases where the final product is sold by a monopolist, who might also acquire cooperating inputs monopsonistically. The question of when areas under a derived demand curve can be given a welfare interpretation, analogous to consumer surplus for a final demand curve, has been broached by Wisecarver (1974).

The concept of derived demand finds its main application in discussions of labour-market questions, and Marshall's tools still play a significant part in the teaching and writing in that area.

See Also

- ▶ Acceleration Principle
- ▶ Marshall, Alfred (1842–1924)

Bibliography

Cournot, A.A. 1838. *Mathematical principles of the theory of wealth.* Trans., New York: Macmillan, 1897.

Diewert, W.E. 1971. A note on the elasticity of derived demand in the n-factor case. *Economica* 38: 192–198.

Gossen, H.H. 1854. *The laws of human relations.* Trans., Cambridge, MA: MIT Press, 1983.

Henderson, D.H. 1922. *Supply and demand*. London: Nisbet.

Hicks, J.R. 1932. *The theory of wages*. London: Macmillan.

Marshall, A. 1890. *Principles of economics*, vol. 1, 8th ed. London: Macmillan, 1920.

Maurice, S.C. 1975. On the importance of being unimportant: An analysis of the paradox in Marshall's third rule of derived demand. *Economica* 42: 385–393.

Menger, C. 1871. *Principles of economics.* Trans., New York: New York University Press, 1981.

Mill, J.S. 1848. *Principles of political economy*. London: Parker.

Robinson, J. 1933. *The economics of imperfect competition*. London: Macmillan.

Wisecarver, D. 1974. The social costs of input-market distortions. *American Economic Review* 64: 359–372.

Design and Impact of Physician Payment Incentives: Theory and Evidence

Douglas A. Conrad

Abstract

This article delineates a broad conceptual framework for predicting physician behavioral response to financial incentives. The framework views incentives within a two-tiered

hierarchy: (1) private health plan or public program payments to the provider organization, e.g., the medical group practice, or integrated delivery system (external incentives); (2) method of compensation to the individual within the provider organization (internal incentives.

Within that framework several dimensions of physician behavior and outcomes are examined – total per capita payments for the physician's panel of patients, quality of services, and physician productivity, in order to provide an integrated view of physician response to financial incentives. Next, a set of propositions are derived broadly from the conceptual framework. Those propositions are evaluated broadly by comparison to the extant empirical evidence. The article concludes by discussing the implications of the empirical evidence for theory and practice pertaining to physicians' response to payment incentives.

In order to fundamentally move payment models from the dominant fee-for-service structure, public policymakers and private health plans must create a "burning platform", in which neither providers nor health insurers perceive that continued reliance on FFS is an option. To catalyze this revolution in incentives, purchasers (public programs, employers, and other sponsors of health plans) must insist on disruptive change toward value-based, rather than volume-based, payment. Payment incentive design will influence care delivery – precisely because different delivery models (e.g., small independent practices, independent practice associations, multispecialty medical groups, and integrated delivery systems) have distinct capabilities for assuming and managing population health risk. Those differences highlight the importance of adjusting payment levels for differences in population health risk, but should not be used as a reason to delay implementation of value-based payment reform.

Keywords

Incentives; Payment reform; Pay-for-performance programs; Value-based payment

JEL Classification

I1; I11; I13

Introduction

Fundamental change in physician payment arrangements is a critical element in discussions of healthcare reform, particularly but not exclusively, in the USA, UK and Europe (Davis and Guterman 2007; Chernew 2010; Arrow et al. 2009; Reinhardt 2011; Campbell et al. 2009; Sutton et al. 2010; Conrad 2009; Christianson and D. Conrad 2011; Kirschner et al. 2012). This article offers a conceptual framework and a set of testable propositions regarding the impact of physician financial incentives, surveys the relevant empirical literature, considers the implications for policy and practice, and concludes with observations on future directions for physician payment reform.

Increasingly, policy reforms have moved away from volume-based payment models and towards value-based models (Miller 2007). Volume-based payment, based on fee-for-service (FFS), tends to encourage over-use of services, especially those with generous profit margins (Ginsburg et al. 2007; Ginsburg. 2012; Hadley et al. 2009–2010), thereby contributing to higher total healthcare costs. An alternative payment model, fixed payment per person per period (capitation), offers strong incentives for cost minimisation. However, that method introduces its own potential inefficiencies, e.g. implicitly discouraging delivery of higher cost services irrespective of potential health benefit, given zero marginal revenue per service (Conrad and Christianson 2004). Moreover, capitation payment exposes providers to population health (actuarial) risk that only integrated delivery systems are sufficiently large in scale and scope of practice to manage effectively (Guterman et al. 2009).

Recently, global capitation payment models – in which the provider organisation receives a fixed payment per person per month for the full continuum of services – have added population risk adjustment in an attempt to

mitigate the actuarial risk borne by providers. Risk adjustment allows clinicians to focus on quality and cost performance rather than selection of favourable health risks (Song et al. 2012). Payment to physicians based on 'value' – translated as patient health benefit from treatment relative to its cost – poses difficult implementation challenges of its own, as witnessed in the first 15 years of experience with pay-for-performance (P4P) programmes (Houle et al. 2012; van Herck et al. 2010; Mehrotra et al. 2010; Conrad and Perry 2009; Rosenthal and Frank 2006). The predominant exemplars of value-based payment are pay-for-performance (P4P), in which provider organisations are paid based on clinical quality or efficiency measures; bundled payment per episode of care with adjustments for quality (Hussey et al. 2011); shared savings models, which divide savings in total healthcare costs between provider and payer; and risk-adjusted global capitation.

Blended payment models (combining capitation with FFS) have emerged as a means of balancing the comparative advantages and disadvantages of FFS and capitation payment, recognising that neither 'corner solution' is likely to achieve the optimum level of quantity and quality of health service, at which marginal health benefit equals marginal cost of production. These blended models include:

- Partial capitation, which combines FFS payment for a subset of services with capitation for services such as care coordination and evaluation and management that are less amenable to piece-rate production (Robinson 2001; Newhouse 1998).
- Mixed models that blend elements of a capitation payment per member per month (pmpm), pay-for-performance (P4P) incentives and FFS (Kantarevic et al. 2011).
- Bundled payment, which pays the accountable provider organisation a fixed amount for the bundle of services required for treatment of an episode of care (Burns and Bailit 2012; de Brantes et al. 2011; Hussey et al. 2011; Chernew 2010).
- Shared savings arrangements, which pay fee-for-service to provider organisations, but periodically share savings if total payments are less than a predetermined total healthcare cost (budget) target. This payment model generally includes a threshold for quality performance or pay-for-performance incentives and has been applied in the Medicare Physician Group Practice demonstration (Colla et al. 2012), the Medicare Shared Savings Program (Berwick 2011) and private health plans (Larson et al. 2012).

P4P programmes, which emerged in health services in the late 1990s and early 2000s, can be viewed as an **adjunct** incentive mechanism, intended to strengthen quality incentives at the margin of general payment by rewarding or penalising providers on the basis of their performance (Chaix-Couturier et al. 2000; Rosenthal et al. 2005, 2007). These programmes mirror the performance-based compensation arrangements in general industry. P4P rewards providers for following evidence-based practice protocols, thus reducing possible stinting on care under capitation, while also discouraging over-use of expensive services through incentives for generic drug prescribing, appropriate use of antibiotics and asthma controller medications, for example (Conrad 2009).

Objective of This Article

This article delineates a broad conceptual framework for predicting physician behavioural response to financial incentives. The framework views incentives within a two-tiered hierarchy: (1) private health plan or public programme payments to the provider organisation, such as the medical group practice or integrated delivery system (external incentives); (2) method of compensation to the individual within the provider organisation (internal incentives). This multitiered structure of physician financial incentives was explicated in the context of health maintenance organisations (HMOs) in the 1990s (Hillman et al. 1992), but applies generally in today's healthcare environment (Robinson et al. 2009; Rosenthal et al. 2002). The framework

draws primarily on the principles of agency theory (McGuire 2000; Prendergast 1999; Ellis and McGuire 1990), behavioural economics (Kahneman and Tversky 1979; McNeil et al. 1982; Mehrotra et al. 2010) and cognitive psychology (Deci et al. 1999; Deci and Ryan 1985).

Within that framework several dimensions of physician behaviour are examined – total per capita payments for the physician's panel of patients, quality of services and physician productivity – in order to provide an integrated view of physician response to financial incentives. Next, a set of propositions derived broadly from the conceptual framework are presented. Those propositions are then evaluated broadly by comparison to the extant empirical evidence. Finally, the article concludes by discussing the implications of the empirical evidence for theory and practice pertaining to physicians' response to payment incentives.

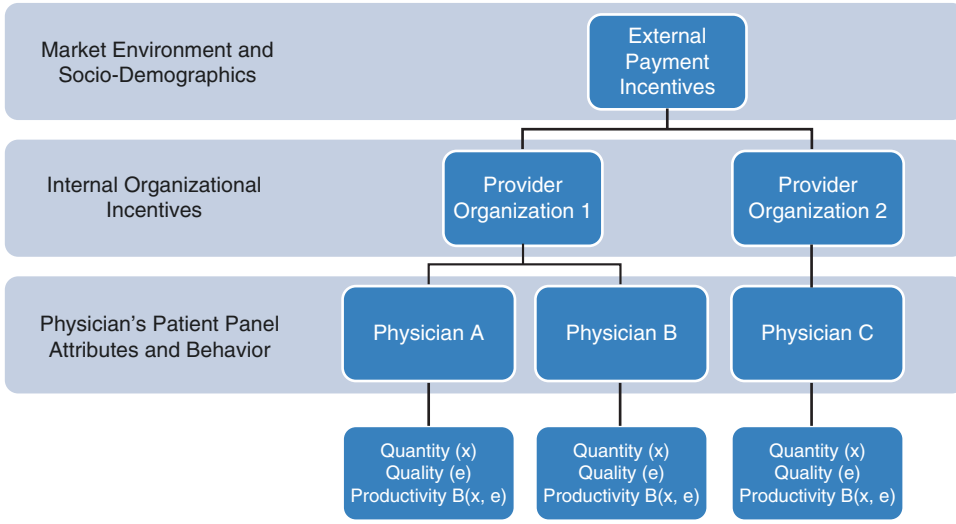
Conceptual Framework

Figure 1 represents the conceptual framework. The hierarchy of payment incentives begins at the top with public and private payers (public insurance programmes and private health plans), whose incentives are designed in response to the structure of their market environment (e.g. competition with other payers, characteristics of product and labour markets, regulation) and sociodemographic characteristics of the population (e.g. income, education, mobility). In turn, those external incentives influence the internal incentives and structure of provider organisations (e.g. individual physician compensation methods, integration of primary care and specialty physicians, care management capabilities, size).

At the third tier in this hierarchy of incentives and structure, the individual physician's behaviour is shaped by the attributes and behaviour of his or her patient panel, as well as by relationships with other physicians both within the provider organisation and in other organisations. By displaying multiple provider organisations and multiple physicians, the figure allows for both inter-organisation and inter-physician competition.

McGuire's algebraic and narrative exposition (2000) of physician agency adds depth to this pictorial representation. Consider the following physician net income function:

$$\pi = n[B(x, L_e) - p_d x] \quad [R + (p_s - c)x],$$



Design and Impact of Physician Payment Incentives: Theory and Evidence, Fig. 1 Physician behavioural response to payment incentives

where:

π = total net income (profit) per patient for the physician;

$n[B(x, L_e) - p_d x]$ = number of patients in the physician's panel, which depends on L_e, the likelihood that the physician will supply effort, e (a measure of non-observable, non-contractible 'quality'), as well as the observable quantity of services per patient (x), for which the patient pays p_d, the out-of-pocket price net of the portion paid by the private health plan or public programme);

$[B(x, L_e) - p_d x]$ = 'net benefit' expected by the patient, given the perceived likelihood of different levels of effort and observed quantity of services supplied by the physician;

R = prospective fixed revenue per patient (i.e. the capitation payment per period);

p_s = supply price to the physician;

c = marginal cost per unit of service (assumed equal to average cost without loss of generality).

The equation illustrates the salient features of the agency problem facing physicians and their patients. To the extent that p_d (the insurance-subsidised price facing the patient) is less than c, over-use of services will result (the moral hazard problem). Similarly, asymmetric information exists between physician and patient, which potentially leads to an undersupply of effort (quality), which is largely unobservable to the patient and thus non-contractible. By setting the fixed capitation (R) relatively high and lowering the rewards to FFS by reducing p_s, the health plan can mitigate one form of over-use. However, to avoid stinting on care by the physician-agent, the plan must provide patients with verifiable information on quality of physician effort (e), while also introducing demand-side incentives for patients to use such quality information in their selection of physician. Furthermore, there are high-value, discrete services such as preventive services which are amenable to FFS payment, so the plan will ensure that their net income margin, $(p_s - c)$, is positive.

Tiered provider networks, in which p_d is set lower for patients selecting physicians with higher quality of effort, are a health plan mechanism for potentially inducing providers to compete on quality (Sinaiko 2011; Baker et al. 2007; Harris 2002). Similarly, value-based insurance design (VBID) is another health plan policy instrument that rewards patients for cost-effective treatment choices by lowering p_d (Fendrick and Chernew 2009). Analogously, P4P programmes both reduce the opacity of quality information to patients and increase the elasticity of demand for physicians offering a higher level of net benefit.

Agency theory, as illustrated above, assumes net income-maximising behaviour by physicians and net benefit-maximising responses by patients. For present purposes, the 'principal' being directly served by the physician-agent is the patient, but one must also acknowledge other implicit agency roles of the physician – attempting to advance the population health for society and improving patient health and reducing healthcare costs per insured person for health insurers. While these assumptions are a reasonable starting point for predicting incentive effects, I have integrated elements of cognitive psychology and the closely related disciplines of behavioural and organisational economics to refine the conceptual framework. Cognitive psychology and its body of experimental and observational evidence (Deci et al. 1999; Deci and Ryan 1985) are particularly informative regarding the roles of and relationship between extrinsic and intrinsic motivation, as is the literature on economic behaviour and organisations (Congleton 1991; Frey 1997; Jack 2005). Behavioural economics, grounded in prospect theory (Kahneman and Tversky 1979; Kahneman et al. 1986) highlights the importance of how decisions are framed and the use of heuristics rather than optimisation in predicting behavioural response – thus potentially enriching predictive power.

Propositions from the Conceptual Framework

Several propositions flow from this integrated framework:

1. Incremental, continuous and more frequent rewards for improved clinical quality performance will produce greater quality gains than all-or-none rewards for exceeding a fixed

quality threshold (Avery and Schultz 2007; Conrad and Perry 2009; Mehrotra et al. 2010). Figure 2 illustrates this point. As a corollary, quality targets based on achievable benchmarks (Kiefe et al. 2001) will lead to greater improvement than fixed standards.

2. Downside risk – in the form of withholds or retrospective penalties – will generate greater improvement per dollar than 'upside potential', or bonuses (Kahneman and Tversky 1979). This prediction derives from what behavioural economists refer to as 'loss aversion', but is also consistent with traditional concepts of diminishing marginal utility of income. However, the use of penalties or withholds could produce negative responses if perceived as unfair by participants (Mehrotra et al. 2010; Kahneman et al. 1986). This reasoning implies the need to optimise the balance between risk and reward in the structure of incentives.
3. Involvement of stakeholders in incentive design is expected to enhance the strength of performance effects – first, by enhancing the credibility and perceived fairness of the quality target, and second, by enhancing communication and awareness of the incentive programme (Conrad 2009; Mehrotra et al. 2010; Kirschner et al. 2012).
4. Performance incentives based on the subject's own performance (whether at the level of the organisation, team or individual) will lead to greater improvement than those tied to performance relative to one's peers, or 'contests' (Mehrotra et al. 2010; Conrad 2009; Conrad and Perry 2009; Frolich et al. 2007).
5. Dollar-for-dollar, incentives aimed at the individual subject or team are likely to generate stronger behavioural responses than those at the organisational level, but this greater incentive strength must be weighed against the direct utility of organisational rewards in supporting enhanced quality infrastructure (Conrad and Perry 2009; Young and Conrad 2007).
6. Because process quality indicators are more controllable by the provider than patient health outcomes, performance incentives geared to care processes are more likely to achieve significant improvement in clinical quality than outcome-based incentives of equivalent size (Conrad 2009; Conrad and Perry 2009; Frolich et al. 2007).
7. In principle, increasing the size of the financial incentive will induce greater quality improvement, other things equal (Van Herck et al. 2010; Conrad and Christianson 2004; Rosenthal and Frank 2006). However, offsetting factors include the potential for extrinsic reward to 'crowd out' intrinsic motivation (Jack 2005; Deci et al. 1999; Frey 1997; Congleton 1991; Deci and Ryan 1985). Furthermore, the marginal utility of each new incentive dollar declines as income increases. There may be a 'sweet spot' in scaling incentive size (see Fig. 2): large enough to offset any diminution in inherent motivation, but not so large as to lead to 'teaching to the test' and other unintended consequences (Mehrotra et al. 2010; Van Herck et al. 2010). The concern about extrinsic rewards crowding out intrinsic motivation is reinforced by the empirical literature on the effects of physician ownership of testing facilities on utilisation rates (Swedlow et al. 1992; Mitchell 2008), as well as the effects on pharmaceutical use of financial incentives for prescribers (Sturm et al. 2007).
8. Quality-based financial incentives that are consistent across multiple health plans and purchasers are more likely to achieve their intended effects on quality than those implemented by a single payer because multiple and potentially conflicting signals are confusing to providers (Van Herck et al. 2010).
9. Physician productivity, whether measured by patient visits or relative value-weighted services per hour worked, will be positively related to the structure of individual physician compensation. The more individualised the incentive and the greater the weight on measured production, the greater will be individual physician productivity, other things being equal (Conrad et al. 2002; Gaynor and Gertler 1995). As a corollary, high-powered, individual production incentives will be more

prevalent as the number of the physicians in the group practice increases.

10. To the extent that internal individual physician compensation incentives are consistent with external incentives used by health plans to pay provider organisations, physician practices will be more likely to succeed financially and to organise patient care more effectively (Hillman et al. 1992). Thus, one would expect FFS payment by plans to induce provider organisations to place a heavier weight on individual productivity-based compensation; conversely, if plans are predominantly paying capitation, one would expect greater prevalence of salary-based compensation.
11. A partial capitation arrangement (high level of *R*, or fixed prospective revenue), coupled with substantial performance incentives based on work effort, clinical quality and patient experience metrics, is expected to out-perform narrowly conceived incentive structures that emphasise one performance dimension. The Alternative Quality Contract of Blue Cross Blue Shield of Massachusetts (Song et al. 2012) illustrates many of these characteristics:
    - Long-term (5-year) risk-adjusted capitation contracts between the health plan and medical group, with gradual tightening of the pre-determined capitation budget constraint over time
    - Lower rates of increase in capitation levels for groups with higher costs per capita at baseline
    - P4P bonuses for quality and patient experience as a balance to the contract's high-powered cost containment incentives.

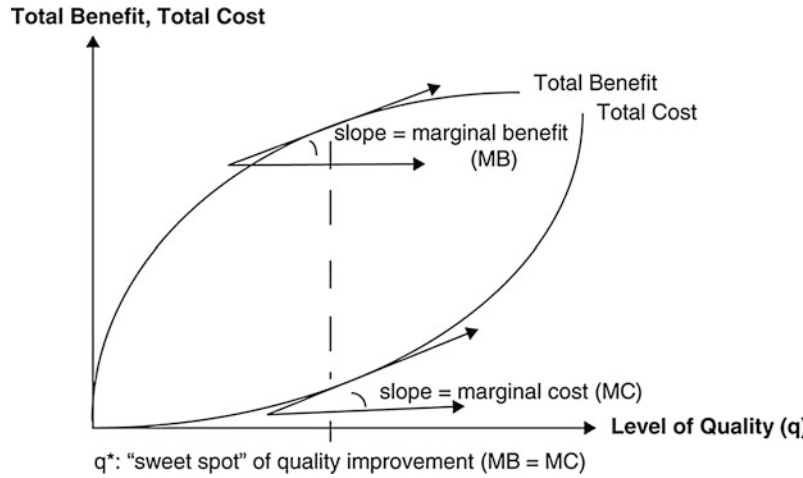
Evidence Summary

The extant empirical literature offers general, but somewhat mixed, support for the preceding theoretical propositions.

For example, the evidence is ambiguous on the comparative efficacy of absolute, continuous rewards versus fixed all-or-none thresholds and on the significance of frequency of incentive payment. Van Herck et al. (2010) report that both fixed threshold targets and continuous improvement scales have led to positive quality gains in the Quality Outcomes Framework of the UK, while findings in other studies are inconclusive. In contrast, the findings are clear that larger quality incentive effects are observed at lower levels of baseline performance (Levin-Scherz et al. 2006; Young et al. 2007; Sutton et al. 2010). Van Herck et al. (2010) summarise the average effect size of quality incentive programmes as approximately 5% improvement, with considerable variation across measures and programme designs and lack of a discernible dose–response relationship (Frolich et al. 2007). The one specific study of the effect of frequency of incentive payment (Chung et al. 2010b) finds no significant differences between quarterly payment ($1250 per 3 months) and annual payment ($5000 per 12 months).

The paucity of empirical evidence in healthcare on the comparative incentive power of rewards versus penalties in P4P precludes firm conclusions (Conrad and Perry 2009). However, in their research Kahneman and Tversky (1979) did find that individuals were more responsive to incentives framed as a loss rather than a gain. For example, an early experiment with physicians asked to choose between surgery or radiation therapy for a patient with cancer found that the choice of surgery was significantly more likely when the surgical risk was characterised as probability of living versus probability of dying (McNeil et al. 1982).

The systematic reviews by Van Herck et al. (2010) and Petersen et al. (2006) indicate that communication and participant awareness are significant contributors to incentive programme success. They point to several studies illustrating substantial positive P4P effects of direct communication with provider stakeholders; effect sizes ranged from 5% to 20% (Gilmore et al. 2007; Beaulieu and Horrigan 2005; Chung et al. 2003; Larsen et al. 2003; Amundson et al. 2003). Kirschner et al. (2012) recently fashioned a P4P programme for primary care in the Netherlands that directly involved target users in programme design, which resulted in incentive weights for



Design and Impact of Physician Payment Incentives: Theory and Evidence, Fig. 2 Benefit–cost tradeoffs in quality improvement

clinical care that were twice those for the other two domains: practice management and patient experience. The average target performance bonus was 5% to 10% of general practice income, and penalties were not included in the programme, which has drawn a growing number of voluntary participants.

The empirical evidence generally validates the superiority of absolute performance incentives over those based on relative performance metrics. The one study of relative performance incentives (Young et al. 2007) found no significant effect on quality. Rosenthal and Dudley (2007) reported that 70% of P4P programmes based their payments on absolute standards and 25% incented improvement. Based on their systematic review, Van Herck et al. (2010) recommended a blended incentive based on absolute achievement and improvement and argued that such a design would induce improved performance among both high and low performers.

The review by Frolich et al. (2007) suggests that individual incentives may be more powerful than group incentives: three of four randomised controlled trials (RCTs) of individual incentives showed significantly positive quality effects, whereas only one of three RCTs of group-level incentives led to significant quality improvement. There is only one study directly comparing the quality impact of individual or team versus organisation-level incentives (Chung et al. 2010a). That study found that eight of nine reported and incentivised quality measures showed greater improvement after the quality incentive switched from group-based to physician-specific. However, since the improvement in the clinic switching to individual-specific incentives was not consistently different from that of two comparison clinics not experiencing the change, the authors could not rule out other factors as the cause of increased improvement in quality.

In a recent systematic review of P4P incentives targeting individual practitioners, Houle et al. (2012) concluded that uncontrolled studies (before–after designs and cohort comparisons) tended to find positive effects of P4P on quality; however, better designed controlled studies did not validate those positive findings.

Van Herck et al. (2010) concluded that process-based P4P programmes yield greater quality gains than those using outcome targets. They reported that intermediate outcome measure improvement was generally in between the rates reported for programmes based process and outcome measures, respectively. The increasing use of outcome incentives in P4P (Rosenthal et al. 2007) offers evidence that health plans and healthcare purchasers are blending process and outcome measures as a means of mitigating any provider tendency to 'game' the incentive mechanism by focusing improvement on measured indicators rather than other, equally important, unmeasured care processes (McGlynn 2007; Eggleston 2005;

Holmstrom and Milgrom 1991). In their study of intended and unintended consequences of the Quality and Outcomes Framework incentive programme in the UK, Sutton et al. (2010) found that incentivised quality indicators in the targeted general practices improved substantially more (by 14.6%) than in the untargeted practices.

There is no conclusive evidence of a consistent dose–response relationship between incentive size and the magnitude of quality improvement (Frolich et al. 2007; Conrad and Perry 2009; Van Herck et al. 2010) – most likely due to the wide variation in study designs and quality indicators observed. However, there is emerging evidence that 'size matters'. For example, Sutton et al. (2010), in the aforementioned study, concluded that provider response was greater for incentivised process measures (e.g. recording blood pressure, cholesterol) that offered greater rewards but also more stringent criteria. Another study (de Brantes and D'Andrea 2009) demonstrated a strong and statistically significant effect of incentive size and the rate of quality improvement programme adoption by physicians for both the Physician Office Link and Diabetes Care Link programmes. Notably, the latter study did not present data regarding actual quality improvement and thus does not speak directly to performance, but rather 'intent-to-perform'.

Signs of 'teaching to the test' or other forms of effort diversion are few in the extant studies (Van Herck et al. 2010), but are reported in some reviews (Petersen et al. 2006; Christianson et al. 2008). On the other hand, Sutton et al. (2010) observed positive spillovers in the UK incentive programme. They estimated an improvement of 10.9% on non-incentivised quality indicators within the targeted group of patients. Another systematic review (Gillam et al. 2012) of the Quality and Outcomes Framework in the UK did sound a cautionary note: '. . . overall, patients reported seeing their usual physician less often and gave lower satisfaction ratings for continuity of care' (p. 464). The authors observed that some health professionals recognised a tendency toward 'protocol-driven care' that might detract from patientled consultations and closer attention to patients' concerns.

While head-to-head comparisons between P4P programmes involving a single payer versus multiple payers are largely absent, the general pattern of findings across studies has led researchers to conclude that fragmentation of incentives among payers does account for some dilution in programme effect (Van Herck et al. 2010; de Brantes and D'Andrea 2009; Pearson et al. 2008; Campbell et al. 2007; Rosenthal and Frank 2006; Pourat et al. 2005).

The findings on physician productivity incentives are virtually unanimous. Gaynor and Gertler (1995) confirmed their hypotheses that: (a) physician group size would be positively related to the use of individual productivity incentives, and (b) use of individual productivity incentives would systematically increase visit productivity. In a separate national study, Conrad et al. (2002) found that compensation incentive effects on productivity followed a predictable gradient. Controlling for other factors, physician productivity tracked the following pattern (from highest to lowest): individual productivity incentive, equal share of group net revenue, fixed salary, and per member per month (quasi-capitation) compensation. A study by Kantarevic et al. (2011) reported that an enhanced FFS model deployed in primary care practices in Ontario, Canada (including increased fees for targeted services, premium payment for extended hours, bonuses for chronic disease management, and incentives for patient enrollment) led to significant increases in services delivered, patient visits, and distinct patients seen (i.e. panel size), compared to the standard FFS payment arrangements.

The hypothesis that individual physician compensation methods within provider organisations will be aligned with the form of external plan payment incentives is confirmed. Rosenthal et al. (2002) observed this pattern in California in their study of medical groups and IPAs, as did Robinson et al. (2009) in their more recent study. Gaynor and Gertler (1995) also found a significant positive relationship between FFS-based health plan payments and high-powered, individual production-based physician compensation in their national study of physician partnerships. The Geisinger Health System has structured its

internal physician compensation as a fixed 78.5–80% component to reflect expected work effort, with the remaining 20–21.5% variable component tied to other goals of quality, efficiency and integration of care (Lee et al. 2012). The stated rationale for the roughly 80% expected productivity component is the predominance of FFS payment in Geisinger's market.

The first 2 years' experience of the BCBSMA Alternative Quality Contract in Massachusetts is instructive. The savings in total health care costs per capita, compared to the comparison group of non-participating practices, were 1.9% in year 1 and 3.3% in year 2. Those savings were primarily achieved by shifting referrals to facilities with lower prices for selected procedures, imaging and tests, as well as reduced utilisation among some groups. Quality of care also improved relative to comparison group practices, and chronic care management, adult preventive care and pediatric care improved more in year 2 than in year 1. However, the cost savings achieved in the first 2 years by the participating medical groups have not been large enough to cover the plan's bonus payments and infrastructure costs.

A systematic review of the evidence regarding financial incentive effects on the quality of patient care delivered by primary care physicians (Scott et al. 2011) offered a useful critique of the prevailing empirical evidence, in particular, lack of attention to potential estimation bias due to physician self-selection into different financial incentive arrangements and the paucity of appropriately designed studies. The authors remark that incentive effects were positive, but modest, and only for some primary outcomes.

Implications for Policy and Practice

This analysis has drawn on the best available theory and empirical evidence to derive a series of propositions that policymakers and practitioners might use in formulating incentive designs tailored to physicians. In general, those propositions have been validated in broad terms by the empirical literature, but there is a wide band of uncertainty surrounding the estimates of financial incentive effects, and the effect sizes are modest. That said, certain central tendencies do emerge from this article's synthesis of theory and evidence:

- To engage physicians and their organisations, incentives must be clear, predictable and right-sized (sufficient to offer a competitive return on practice improvement, but not so large as to crowd out providers' intrinsic motivation).
- Mixed payment arrangements (blended payment, partial capitation) are likely to dominate single-dimension incentives (fee-for-service or global capitation) and should be designed to fit the current care management capacity of the provider organisation *and* to stimulate care delivery redesign.
- Physician and organisational stakeholders should participate actively in the negotiation of incentive contracts, and contract terms must be transparent to all major parties.
- The focus on process indicators in the physician reporting and incentive structure – rather than patient health outcomes – seems to match the state of the art in health plan-based quality measurement, reporting and assessment systems over the past two decades.
- In order to sustain innovation in physician financial incentives, payers and providers will need to discover means of reducing negotiation costs and of right-sizing the payment incentives, so that both parties – payer and provider – will anticipate a competitive return on their investment.

In keeping with the modest success of first-generation physician incentive programmes in the USA, multi-payer collaboration and efforts to couple quality-based financial incentives with programmes geared to both practice redesign and shared savings between payers and providers are evolving rapidly (Damberg et al. 2009; Davis and Guterman 2007). A full-court press on quality and efficiency, based on common and broadly defined clinical and economic metrics among multiple payers and providers, seems to be the logical next step in payment reform and health delivery system transformation.

My review of the theory, evidence, and history of physician financial incentives leads me to the following conclusion. In order to fundamentally move payment models from the dominant fee-for-service structure, public policymakers and private health plans must create a 'burning platform', in which neither providers nor health insurers perceive that continued reliance on FFS is an option. To catalyse this revolution in incentives, purchasers (public programmes, employers and other sponsors of health plans) must insist on fundamental change toward value-based, rather than volume-based, payment of physician organisations. Payment incentive design will influence care delivery – precisely because different delivery models (e.g. small independent practices, independent practice associations, multi-specialty medical groups and integrated delivery systems) have distinct capabilities for assuming and managing population health risk. Those differences highlight the importance of adjusting payment levels for differences in population health risk, but should not be used as a reason to delay implementation of value-based payment reform.

See Also

- ▶ Health Insurance, Economics of
- ▶ Health Economics

Bibliography

Amundson, G., L.I. Solberg, M. Reed, E.M. Martini, and R. Carlson. 2003. Paying for quality: Compliance with tobacco cessation guidelines. *Joint Commission Journal on Quality and Safety* 29: 59–65.

Arrow, K., A. Auerbach, J. Bertko, M.S. Brownlee, et al. 2009. Toward a 21st century health care system: Recommendations for health care reform. *Annals of Internal Medicine* 150: 493–495.

Avery, G., and J. Schultz. 2007. Regulation, financial incentives, and the production of quality. *American Journal of Medical Quality* 22(4): 265–273.

Baker, L., K. Bundorf, A. Royalty, C. Galvin, and K. McDonald. 2007. Consumer-oriented strategies for improving health benefit design: An overview. *Agency for Healthcare Research and Quality Report*, No. 07–0067. Rockville.

Beaulieu, N.D., and D.R. Horrigan. 2005. Putting smart money to work for quality improvement. *Health Services Research* 40: 1318–1334.

Berwick, D.M. 2011. Launching accountable care organizations: The proposed rule for the medicare shared savings program. *New England Journal of Medicine* 364(16), e32.

Burns, M.E. and M.H. Bailit. 2012. Bundled payment across the U.S. today: status of implementation and operational findings. *HCI3 Improving Incentives Issue Brief,* April, 1–16. Available at http://www.hci3.org/content/hci3-improving-incentives-issue-brief-bundled-payment-across-us-today Accessed 20 Nov 2012.

Campbell, S., D. Reeves, E. Kontopantelis, E. Middleton, B. Sibbald, and M. Roland. 2007. Quality of primary care in England with the introduction of pay for performance. *New England Journal of Medicine* 357(2): 181–190.

Campbell, S.M., D. Reeves, E. Kontopantelis, B. Sibbald, and M. Roland. 2009. Effects of pay for performance on the quality of primary care in England. *New England Journal of Medicine* 361: 368–378.

Chaix-Couturier, C., I. Durand-Zaleski, D. Jolly, and P. Durieux. 2000. Effects of financial incentives on medical practice: Results from a systematic review of the literature and methodological issues. *International Journal for Quality in Health Care* 12(2): 133–142.

Chernew, M. 2010. Editorial: Bundled payment systems: Can they be more successful this time? *Health Services Research* 45(5, Part I): 1141–1147.

Christianson, J.B. and D. Conrad. 2011. Provider payment and incentives. In *The oxford handbook of health economics,* eds. S. Glied and P.C. Smith, pp. 624–648. Oxford and New York: Oxford University Press.

Christianson, J.B., S. Leatherman, and K. Sutherland. 2008. Lessons from evaluations of purchaser pay-for-performance programs: A review of the evidence. *Medical Care Research and Review* 65(6 Suppl.): 5S–35S.

Chung, R.S., H.O. Chernicoff, K.A. Nakao, R.C. Nickel, and A.P. Legorreta. 2003. A quality-driven physician compensation model: Four-year follow-up study. *Journal for Healthcare Quality* 25: 31–37.

Chung, S., L.P. Palaniappan, L.M. Trujillo, H.R. Rubin, and H.S. Luft. 2010a. Effect of physician-specific pay-for-performance incentives in a large group practice. *American Journal of Managed Care* 16: E35–E42.

Chung, S., L. Palaniappan, E. Wong, H. Rubin, and H. Luft. 2010b. Does the frequency of pay-for-performance payment matter? Experience from a randomized trial. *Health Services Research* 45: 553–564.

Colla, C.H., D.E. Wennberg, E. Meara, J.S. Skinner, D. Gottlieb, V.A. Lewis, C.M. Snyder, and E.S. Fisher. 2012. Spending differences associated with the medicare physician group practice demonstration. *Journal of the American Medical Association* 308(10): 1015–1023.

Congleton, R.D. 1991. The economic role of a work ethic. *Journal of Economic Behavior and Organization* 15: 365–385.

Conrad, D.A. 2009. Incentives for health care improvement. In *Performance measurement for health system*

improvement: Experience, challenges, and prospects, eds. P.C. Smith, E. Mossialos, I. Papinicolas and S. Leatherman, pp. 582–612. The cambridge health economics, policy, and management series. London: Cambridge University Press.

Conrad, D.A. and J.B. Christianson. 2004. Penetrating the 'black box': Financial incentives for enhancing the quality of physician services. *Medical Care Research and Review,* 61(3, Special Supplement), 37S–68S.

Conrad, D.A. and L. Perry. 2009. Quality-based financial incentives in health care: Can we improve quality by paying for it? *Annual Review of Public Health,* 30, 357–371.

Conrad, D.A., A. Sales, S.Y. Liang, A. Chaudhuri, C. Maynard, L. Pieper, L. Weinstein, D. Gans, and N. Piland. 2002. The impact of financial incentives on physician productivity in medical groups. *Health Services Research* 37(4): 885–906.

Damberg, C.L., K. Raube, S.S. Teleki, and E. dela Cruz. 2009. Taking stock of pay-for-performance: A candid assessment from the front lines. *Health Affairs* 28(2): 517–525.

Davis, K., and S. Guterman. 2007. Rewarding excellence and efficiency in Medicare payments. *The Milbank Quarterly* 85(3): 449–468.

de Brantes, F.S., and B.G. D'Andrea. 2009. Physicians respond to pay-for-performance incentives: Larger incentives yield greater participation. *American Journal of Managed Care* 15(5): 305–310.

de Brantes, F., A. Rastogi, and C.M. Soerensen. 2011. Episode of care analysis reveals sources of variations in costs. *American Journal of Managed Care* 17(10): e383–e392.

Deci, E.L., and R.M. Ryan. 1985. *Intrinsic motivation and self-determination in human behavior*. New York: Plenum.

Deci, E.L., R. Koestner, and R.M. Ryan. 1999. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin* 125(6): 627–668.

Eggleston, K. 2005. Multitasking and mixed systems for provider payment. *Journal of Health Economics* 24: 211–223.

Ellis, R.P., and T.G. McGuire. 1990. Optimal payment systems for health services. *Journal of Health Economics* 9: 375–396.

Fendrick, A.M., and M.E. Chernew. 2009. Value-based insurance design: Maintaining a focus on health in an era of cost containment. *American Journal of Managed Care* 15(6): 338–343.

Frey, B.S. 1997. On the relationship between intrinsic and extrinsic work motivation. *International Journal of Industrial Organization* 15: 427–439.

Frolich, A., J.A. Talavera, P. Broadhead, and R.A. Dudley. 2007. A behavioral model of clinician responses to incentives to improve quality. *Health Policy* 80: 179–193.

Gaynor, M., and P. Gertler. 1995. Moral hazard and risk-spreading in partnerships. *RAND Journal of Economics* 26(4): 591–613.

Gillam, S.J., A.N. Siriwardena, and N. Steel. 2012. Pay-for-performance in the United Kingdom: Impact of the quality and outcomes framework – a systematic review. *Annals of Family Medicine* 10(5): 461–468.

Gilmore, A.S., Y.X. Zhao, N. Kang, K.L. Ryskina, A.P. Legorreta, D.A. Taira, and R.S. Chung. 2007. Patient outcomes and evidence-based medicine in a preferred provider organization setting: A six-year evaluation of a physician pay-for-performance program. *Health Services Research* 42: 2140–2159.

Ginsburg, P.B. 2012. Fee-for-service will remain a feature of major payment reforms, requiring more changes in Medicare physician payment. *Health Affairs* 31(9): 1977–1983.

Ginsburg, P., H.H. Pham, K. McKenzie., and A. Milstein. 2007. Distorted payment system undermines business case for health quality and efficiency gains. *Issue Brief* No. 112. Center for Studying Health System Change.

Guterman, S., K. Davis, S. Schoenbaum, and A. Shih. 2009. Using Medicare policy to transform the health system: A framework for improving performance. *Health Affairs* 28(2): w238–w250.

Hadley, J., J. Reschovsky, C. Corey, and S. Zuckerman. 2009–2010. Medicare fees and the volume of physicians' services. *Inquiry,* 46(4): 372–390.

Harris, K.M. 2002. Can high quality overcome consumer resistance to restricted provider access? Evidence from a health plan choice experiment. *Health Services Research* 37(3): 551–571.

Hillman, A.L., W.P. Welch, and M.V. Pauly. 1992. Contractual arrangements between HMOs and primary care physicians: Three-tiered HMOs and risk pools. *Medical Care* 30(2): 136–148.

Holmstrom, B., and P. Milgrom. 1991. Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, and Organization* 7(Special Issue): 24–52.

Houle, S.K.D., F.A. McAlister, C.A. Jackevicius, A.W. Chuck, and R.T. Tsuyuki. 2012. Does performance-based remuneration for individual health care practitioners affect patient care? A systematic review. *Annals of Internal Medicine* 157(12): 889–899.

Hussey, P.S., M.S. Ridgely, and M.B. Rosenthal. 2011. The PROMETHEUS bundled payment experiment: Slow start shows problems in implementing new payment models. *Health Affairs* 30(11): 2116–2124.

Jack, W. 2005. Purchasing health care services from providers with unknown altruism. *Journal of Health Economics* 24: 73–93.

Kahneman, D., and A. Tversky. 1979. Prospect theory: An analysis of decision under risk. *Econometrica* 47: 263–292.

Kahneman, D., J.L. Knetsch, and R.H. Thaler. 1986. Fairness as a constraint in profit seeking: Entitlements in the market. *American Economic Review* 76(4): 728–741.

Kantarevic, J., B. Kralj, and D. Weinkauf. 2011. Enhanced fee-for-service model and physician productivity: Evidence from family health groups in Ontario. *Journal of Health Economics* 30: 99–111.

Kiefe, C.I., J.J. Allison, O.D. Williams, S.D. Person, M.T. Weaver, and N.W. Weissman. 2001. Improving quality improvement using achievable benchmarks for physician feedback: A randomized controlled trial. *Journal of the American Medical Association* 285(22): 2871–2879.

Kirschner, K., J. Braspenning, J.E. Annelies Jacobs, and R. Grol. 2012. Design choices made by target users for a pay-for-performance program in primary care: An action research approach. *BMC Family Practice* 13: 25. Available at: http://www.biomedcentral.com/1471-2296/13/25 (accessed 20 February 2013).

Larsen, D.L., W. Cannon, and S. Towner. 2003. Longitudinal assessment of a diabetes care management system in an integrated health network. *Journal of Managed Care Pharmacy* 9: 552–558.

Larson, B.K., A.D. Van Citters, S.A. Kreindler, K.L. Carluzzo, J.N. Gbemudu, F.M. Wu, E.C. Nelson, S.M. Shortell, and E.S. Fisher. 2012. Insights from transformations underway at four Brookings-Dartmouth accountable care organization pilot sites. *Health Affairs* 31(11): 2395–2406.

Lee, T.H., A. Bothe, and G.D. Steele. 2012. How Geisinger structures its physician compensation to support improvements in quality, efficiency, and volume. *Health Affairs* 31(9): 2068–2073.

Levin-Scherz, J, N. DeVita, and J. Timbie. 2006. Impact of pay-for-performance contracts and network registry on diabetes and asthma HEDIS® measures in an integrated delivery network. *Medical Care Research and Review,* 63(1, Special Supplement), 14S–28S.

McGlynn, E.A. 2007. Intended and unintended consequences: What should we really worry about? *Medical Care* 45(1): 3–5.

McGuire, T.G. 2000. Physician agency. In *Handbook of health economics*, Vol. 1A, eds. A.J. Culyer and J.P. Newhouse, pp. 461–536. Amsterdam: Elsevier.

McNeil, B.J., S.J. Pauker, H.C. Sox Jr., and A. Tversky. 1982. On the elicitation of preferences for alternative therapies. *New England Journal of Medicine* 306(21): 1259–1262.

Mehrotra, A., M.E.S. Sorbero, and C.L. Damberg. 2010. Using the lessons of behavioral economics to design more effective pay-for-performance programs. *American Journal of Managed Care* 16(7): 497–503.

Miller, H.D. 2007. *Creating payment systems to accelerate value-driven health care: Issues and options for policy reform*. Washington, DC: The Commonwealth Fund.

Mitchell, J.M. 2008. Utilization trends for advanced imaging procedures: Evidence from individuals with private insurance coverage in California. *Medical Care* 46(5): 460–466.

Newhouse, J.P. 1998. Risk adjustment: Where are we now? *Inquiry* 35(2): 122–131.

Pearson, S.D., E.C. Schneider, K.P. Kleinman, K.L. Coltin, and J.A. Singer. 2008. The impact of pay-for-performance on health care quality in Massachusetts: 2001–2003. *Health Affairs* 27(4): 1167–1176.

Petersen, L.A., L.D. Woodard, T. Urech, C. Daw, and S. Sookanan. 2006. Does pay-for-performance improve the quality of health care? *Annals of Internal Medicine* 145(4): 265–272.

Pourat, N., T. Rice, M. Tai-Seale, G. Bolan, and J. Nihalani. 2005. Association between physician compensation methods and delivery of guideline-concordant STD care: Is there a link? *American Journal of Managed Care* 11(7): 426–432.

Prendergast, C.R. 1999. The provision of incentives in firms. *Journal of Economic Literature* 37: 7–63.

Reinhardt, U.E. 2011. The many different prices paid to providers and the flawed theory of cost shifting: Is it time for a more rational all-payer system? *Health Affairs* 30(11): 2125–2133.

Robinson, J.C. 2001. Theory and practice in the design of physician payment incentives. *Milbank Quarterly* 79(2): 149–177.

Robinson, J.C., S.M. Shortell, D.R. Rittenhouse, S. Fernandes-Taylor, R.R. Gillies, and L.P. Casalino. 2009. Quality-based payment for medical groups and individual physicians. *Inquiry* 46: 172–181.

Rosenthal, M.B., and R.A. Dudley. 2007. Pay-for-performance: Will the latest payment trend mprove care? *Journal of the American Medical Association* 297(7): 740–744.

Rosenthal, M.B., and R.G. Frank. 2006. What is the empirical basis for paying for quality in health care? *Medical Care Research and Review* 63(2): 135–157.

Rosenthal, M.B., R.G. Frank, J.L. Buchanan, and A.M. Epstein. 2002. Transmission of financial incentives to physicians by intermediary organizations in California. *Health Affairs* 21(4): 197–205.

Rosenthal, M.B., R.G. Frank, Z. Li, and A.M. Epstein. 2005. Early experience with pay-for-performance: From concept to practice. *Journal of the American Medical Association* 294(14): 1788–1793.

Rosenthal, M.B., B.E. Landon, K. Howitt, H.R. Song, and A.M. Epstein. 2007. Climbing up the pay-for-performance learning curve: Where are the early adopters now? *Health Affairs* 26(6): 1674–1682.

Scott, A., P. Sivey, D. Ait Ouakrim, L. Willenberg, L. Naccarella, J. Furler, and D. Young. 2011. The effect of financial incentives on the quality of health care provided by primary care physicians (Review). *Cochrane Database of Systematic Reviews,* Issue 9. Article No.: CD008451. doi: 10.1002/14651858.CD008451.pub2.

Sinaiko, A.D. 2011. How do quality information and cost affect patient choice of provider in a tiered network setting? Results from a survey. *Health Services Research* 46(2): 437–456.

Song, Z., D.G. Safran, B.E. Landon, M.B. Landrum, Y. He, R.E. Mechanic, M.P. Day, and M.E. Chernew. 2012. The 'Alternative Quality Contract', based on a global budget, lowered medical spending and improved quality. *Health Affairs* 31(8): 1885–1894.

Sturm, H., A. Austvoll-Dahlgren, M. Aaserud, A.D. Oxman, C. Ramsey, A. Vernby, and J.P. Kosters. 2007. Pharmaceutical policies: Effects of financial incentives for prescribers. *Cochrane Database of Systematic Reviews* 18(3): CD006731.

Sutton, M., R. Elder, B. Guthrie, and G. Watt. 2010. Record rewards: The effects of targeted quality incentives on the recording of risk factors by primary care providers. *Health Economics* 1: 1–13.

Swedlow, A., G. Johnson, N. Smithline, and A. Milstein. 1992. Increased costs and rates of use in the California workers' compensation system as a result of self-referral by physicians. *New England Journal of Medicine* 327(21): 1502–1506.

Van Herck, P., D. De Smedt, L. Annemans, R. Remmen, M.B. Rosenthal, and W. Sermeus. 2010. Systematic review: Effects, design choices, and context of pay-for-performance in health care. *BMC Health Services Research,* 10: 247.

De-Skilling

A. L. Friedman

The proposition that there is a long run tendency for workers to become de-skilled as part of the basic operation of capitalist economies can be found in Marx (1867). The change in capitalist stages from Cooperation to Manufacture was distinguished by the division of labour under individual capitalists. The effect of this on workers is that they are ordered to specialize in a narrow range of tasks. The worker is transformed from an all-round craftsman into what Marx calls a detail worker. His detail dexterity becomes over-exercised, and he is thereby turned into a 'crippled monstrosity'. The de-skilling process continues with the next stage of capitalism, Modern Industry. Under Manufacture, the traditional skills of workers are still required collectively even if individual workers may lose the ability to perform all the tasks required in a single trade. With Modern Industry the heart of the labour process becomes the machine. Workers become appendages of the machines. Their tasks concern feeding, minding and maintaining machines rather than parts of a skilled labour process.

This process of de-skilling is viewed by Marx as fundamental to the logic of capitalist development. Under competitive pressure capitalists must reduce costs. Labour costs can be reduced considerably if the skill required of workers is removed. This widens the supply of suitable workers, thereby swelling the reserve army of labour, and keeps wages down to subsistence levels. It means that expensive apprentice schemes do not have to be supported. It also means that workers, both individually and as a collectivity, do not possess the secrets that were the basis of the guilds' power to limit labour through long apprenticeships, even after most individual workers no longer used their acquired skills, This facilities the employment of women and children and reduces further the level of wages because now workers do not have to be paid a family wage.

That the division of labour could have a de-skilling effect on workers was well known before Marx. Smith (1776), the champion of the division of labour for its effects on productivity, noted that specialization could produce boredom and a loss of traditional skills. However, this was not emphasized by Smith, and he believed that the harmful effects could be effectively countered by public education.

One of the main benefits to employers of the division of labour was that tasks which were divided up into segments each requiring different skill levels, in principle could then be given out to workers with just those skills required to do the job. This would allow employers to pay for only those skills actually used in the labour process. This effect of the division of labour, known now as the Babbage Principle (Babbage 1835), was reiterated by Marx.

There are two counter-arguments to the de-skilling hypothesis which have become very influential. The first is associated with Durkheim (1893). The effect of the division of labour is specialization. However, for Durkheim specialization means stimulating a diversity of skills rather than a loss of general skill. Diversity means that individuals are able to act in accordance with their own individualized preferences rather than preferences which are homogenized by social pressures. This allows an organic solidarity to be achieved in society as opposed to the mechanical solidarity of undifferentiated individuals which Durkheim primarily associates with primitive societies. A similar attitude towards the division of labour underpins the human capital view of skill (Becker

1964). Specialization requires specialist skills. The opportunities are thereby created which allow individuals to choose to invest in acquiring those skills by foregoing immediate earnings.

The second argument against de-skilling is one which accepts Marx's judgement for the nineteenth century and early twentieth century, but which claims that he is now outdated. According to this view, skill requirements depend on overall types of technologies used in production (Woodward 1958; Blauner 1964). The shift from unit and small-batch production techniques to large-batch and mass-production techniques involved a loss of craft skills and a rise in worker alienation. However, it is argued that more recently technology has been changing once more, this time towards process production technologies. These require new skills associated with scientific and technical training and result in work which is also less alienating.

Both groups arguing against the de-skilling thesis have pointed to the universal rise in years of formal education and the relative growth in white-collar jobs as evidence against the de-skilling hypothesis.

The de-skilling hypothesis was revived by Braverman (1974). Braverman's argument was essentially that capitalism had not changed. At base the trend to de-skilling had continued. Although certain changes had occurred since Marx's time, they amounted to a change in the methods of de-skilling rather than in the fundamental direction of de-skilling. For Braverman the process of de-skilling involves four processes. First, the shop floor loses the right to design and plan work; that is, the separation of planning from doing or conception from execution, in an overall sense. Second, work is fragmented into meaningless segments. Third, tasks are redistributed among unskilled and semi-skilled labour according to the Babbage Principle, and conceptual activities are concentrated on as few workers as possible. Fourth, work is monitored closely.

According to Braverman, the essential problem for management is to control the variability and uncertainty associated with the transformation of worker's capacity to work (labour power) into actual work. Control over this transformation is the key to profitability. This has not changed since Marx's time. What has changed is the application of scientific principles to this task. The 20th century is distinguished by the application of Frederick Taylor's system of scientific management to this task.

Braverman accuses the Durkheimian school of concentrating only on what Marx would have called the social division of labour, the separation of trades or broad occupations, and ignoring what Marx called the manufacturing division of labour, the dividing up of work tasks within individual firms. He also argues against the straightforward determination of social processes by technology. Skill levels depend on the uses made of technology and on the social organization through which technology is used. This depends on the purposes of the agents in control of that social organization. That is, it is the fundamental needs of capitalists, not the technology itself, which should be the starting point of the analysis of skills. In particular, Braverman notes that the rise in white-collar work has not primarily occurred by a rise in the proportion of highly skilled technicians. Most white-collar occupational growth has been due to a rise in the proportion of low-level clerical staff. They have been de-skilled in precisely the same way as manual workers. Concerning rising levels of formal education, Braverman cites the considerable evidence that educational achievement required for recruitment exceeds the actual requirements for carrying out work in most job categories.

The de-skilling hypothesis has received considerable critical attention by theorists who have been inspired by Braverman. For Edwards (1979), what has grown during the 20th century is capitalist control over the labour process. This has occurred by substituting personal and direct ways of controlling labour with structural methods, such as by building controls into the overall pattern of machine design for whole factories or by bureaucratic procedures governing worker behaviour which apply to all workers. This change involves de-skilling some workers and re-skilling others at the same time.

Friedman (1977) associates de-skilling with one of two general types of management strategy: the Direct Control strategy. This strategy focuses

on reducing the discretion individual workers can exercise. The alternative strategy, Responsible Autonomy, focuses on achieving high flexibility from workers by allowing them high discretion and by encouraging their loyalty to to the firm. According to Friedman, there is a perpetual tension between these strategies. During times of severe product market competition based on price (rather than quality), and times of excess labour supply, managers will be pushed toward Direct Control strategies and de-skilling. However, in the opposite environmental conditions, or when pushed by strong worker resistance, they will be encouraged to move towards Responsible Autonomy strategies.

Littler (1982) also notes that the problem of high-cost skilled workers can be avoided rather than confronted, by firms moving to new sites or when technological changes lead to new firms making products which had been made using skilled workers. The overall effect may be to change the composition of skills required, but individual skilled jobs are not directly redefined, they simply disappear.

Braverman has been widely criticized for ignoring the effects of worker resistance on managerial behaviour. According to Edwards, the Taylorist programme was a failure due to strong worker resistance. Friedman criticizes Braverman for presuming that the basic problem of management is to deal with worker variability and uncertainty. Managers often take actions to counter specific moves by workers, rather than merely to reduce the harmful effects that unanticipated worker resistance may cause. Also, by reducing the opportunities for workers to disrupt production, managers lose the advantages to be gained by harnessing worker initiative and loyalty. These circumstances allow the Responsible Autonomy strategy, and its attendant re-skilling possibilities, to be profitable.

Currently, while most of the post-Braverman theorists view de-skilling as an important symptom of capitalist development, it is viewed as a process which directly affects particular groups of workers in only limited episodes of capitalist development. Often the process occurs simply by capital moving around highly skilled groups. Also, the redivision of labour does not always run in the de-skilling direction. Sometimes capitalist development involves strategies which allow re-skilling.

See Also

- ▶ Division of Labour
- ▶ Durkheim, Emile (1858–1917)

Bibliography

Babbage, C. 1835. On the economy of machinery and manufacture. Excerpted in *Design of jobs*, ed. L.E. Davis and J.C. Taylor. Harmondsworth: Penguin, 1972.

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

Blauner, R. 1964. *Alienation and freedom: The factory worker and his industry*. Chicago: University of Chicago Press.

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

Durkheim, E. 1893. *The division of labour in society*. New York: Free Press, 1964.

Edwards, R. 1979. *Contested terrain*. London: Heinemann.

Friedman, A.L. 1977. *Industry and labour*. London: Macmillan.

Littler, C.R. 1982. *The development of the labour process in capitalist societies*. London: Heinemann.

Marx, K. 1867. *Capital*, vol. 1. London: Lawrence & Wishart, 1970.

Smith, A. 1776. In *An inquiry into the nature and causes of the wealth of nations*, ed. E. Cannan. London: Methuen, 1961.

Woodward, J. 1958. *Management and technology*. London: HMSO.

Destutt De Tracy, Antoine Louis Claude (1754–1836)

R.F. Hébert

Keywords

Capitalization theory of taxation; Destutt de Tracy, A. L. C.; Ideology; Labour theory of value; Physiocracy; Productive and unproductive labour

JEL Classifications

B31

French philosopher and economist, Tracy was born into a noble family of the *ancien régime* at Paris on 20 July 1754 and died in the same city on 10 March 1836. His life spanned the most tumultuous period of French history, from the twilight of the Old Regime to the dawn of capitalism, romanticism and socialism. One of the last *philosophes*, Tracy began as an eighteenth-century classical metaphysician, preoccupied with the sensationalist doctrine of Locke and Condillac, and ended up, in the words of Auguste Comte, as the philosopher 'who had come closest to the positive state'. In the interim he knelt at the feet of Voltaire; served alongside Lafayette in the Royal Cavalry, and as deputy to the French Estates General and the Constituent Assembly; was imprisoned during the Reign of Terror; released after Thermidor (escaping the guillotine by a mere 2 days); subsequently helped to establish his country's first successful national programme of public education; led the opposition to Napoleon from his seat in the French Senate; regained his title under the Bourbon Restoration; counted among his associates the likes of Mirabeau, Condorcet, Cabanis, DuPont de Nemours, Jefferson, Franklin, Lavoisier, Ricardo and Mill; and retained his early sympathies for liberty throughout.

Long before it took on its pejorative sense at the hands of Marx, Tracy coined the term 'ideology' (by which he meant the science of ideas) to describe his philosophy, which embraced and intertwined psychological, moral, economic and social phenomena, but which gave primacy to economics because he thought that the purpose of society was to satisfy man's material needs and multiply his enjoyments. Tracy rejected the Physiocratic notion of value, substituting a labour theory that Ricardo subsequently endorsed in his *Principles*. Like Say, he denied Smith's distinction between productive and unproductive labour. But unlike Smith or Say, he reduced all wealth, including land, to labour. On numerous other topics (that is, wages, profits, rents, exchange, price variations, international trade) he was far less thorough and rigorous than either Smith or Say, but his exposition of the capitalization theory of taxation was superior to the rest. In the final analysis, his *Traité* was not properly a treatise on political economy so much as a part of a general study of the human will. Yet the resulting lack of depth did not impair his remarkable ability to allure great minds. Ricardo found him 'a very agreeable old gentleman', and Jefferson was influenced to the point of including 'ideology' among the ten projected departments in his plan for the University of Virginia.

Along with Say, Destutt de Tracy was one of the earliest members of the French liberal school. Patrician, philosopher and patriot, caught in the grips of major social and economic upheaval, he denounced the interests of his own class (the *rentiérs*) and became the spokesman of a nascent capitalism in which he had neither role nor vested interest.

Selected Works

1804–18. *Eléments d'idéologie*, 5 vols. Paris: Courcier.

1811. *A commentary and review of Montesquieu's spirit of laws*. Trans, Philadelphia: William Duane.

1817. *A treatise on political economy,* ed. T. Jefferson. New York: Augustus M. Kelley, 1970.

Bibliography

Allix, E. 1912. Destutt de Tracy, économiste. *Revue d'Economie Politique* 26: 424–451.

Kennedy, E. 1978. *Destutt de Tracy and the Origins of 'Ideology'*. Philadelphia: American Philosophical Society.

Picavet, F.J. 1918. In *Les Idéologues*, ed. chs 5 and 6. Paris: F. Alcan.

Determinacy and Indeterminacy of Equilibria

Chris Shannon

Abstract

This article discusses work on the determinacy and indeterminacy of equilibria in models of competitive markets. Determinacy typically refers to situations in which equilibria are finite

in number, and local comparative statics can be precisely described. The article describes basic results on generic determinacy for exchange economies and the general underlying principles, together with various applications and extensions including incomplete financial markets and markets with infinitely many commodities.

Keywords

Ambiguity aversion; Arrow–Debreu model; Comparative statics; Concavity; Continuous-time trading; Continuum of equilibria; Determinacy and indeterminacy of equilibria; Edgeworth box economy; Excess demand functions; Existence of equilibrium; Implicit function theorems; Incomplete markets; Infinite horizons; Infinite-dimensional economies; Lipschitz continuous functions; Loss aversion; Multiple equilibria; Reference dependence; Restricted participation; Sard's th; Transversality th; Uniqueness of equilibrium; Walras's Law

JEL Classifications

D5

Introduction

The Arrow–Debreu model of competitive markets is one of the cornerstones of economics. Part of the explanatory power of this model stems from its flexibility in capturing price-taking behaviour in many different markets, and from the predictive power arising from the great generality under which equilibrium can be shown to exist. This predictive power is significantly enhanced when equilibria are determinate, meaning that equilibria are locally unique and local comparative statics can be precisely described. Instead, when equilibria are indeterminate, even arbitrarily precise local bounds on variables might not suffice to give a unique equilibrium prediction, the model might exhibit infinitely many equilibria, and each might be infinitely sensitive to arbitrarily small changes in parameters.

Simple exchange economies cast in an Edgeworth box with two agents and two goods illustrate the possibility of indeterminacy in equilibrium. One easy example arises when agents view the goods as perfect substitutes. In this case, every profile of initial endowments leads to a continuum of equilibria. Another example comes from the opposite extreme, in which each agent views the goods as perfect complements. Every profile of initial endowments dividing equal social endowments of the two goods leads to a continuum of equilibria. These examples may seem degenerate, since they involve individual demand behaviour either extremely responsive to prices, or extremely unresponsive to prices. Similar examples can be constructed using preferences that are less extreme, however, and that can be chosen to satisfy a number of regularity conditions including strict concavity, strict monotonicity, and smoothness. Problems from standard graduate texts illustrate this possibility. In fact, indeterminacy is unavoidable, at least for some endowment profiles, in almost any model that may exhibit multiple equilibria for some choices of endowments. The conditions leading to unique equilibria or unambiguous global comparative statics are well-known to be very restrictive, suggesting that equilibrium indeterminacy may be a widespread phenomenon.

In a deeper sense, however, these examples of indeterminacy remain knife-edge. Under fairly mild conditions on primitives, if an initial endowment profile leads to indeterminacy in equilibrium, arbitrarily small perturbations in endowment profiles must restore the determinacy of equilibrium. More powerfully, the set of endowment profiles for which equilibria are determinate is generic, that is, an open set of full Lebesgue measure. Explaining this remarkable result – originally postulated and established by Debreu (1970) – and its many extensions and generalizations is the focus of this article. Section "Determinacy in Finite Exchange Economies" lays out the basic question of determinacy of equilibrium in finite exchange economies, and sketches the results. Section "Determinacy and Indeterminacy: A New Approach to Many Problems" describes the general underlying principles,

together with various applications and extensions. Section "Determinacy in Infinite-Dimensional Economies" concludes by examining recent work on determinacy in markets with infinitely many commodities.

Determinacy in Finite Exchange Economies

Imagine a family of exchange economies, each with a fixed set of L commodities and a fixed set of m agents, $i = 1, \ldots, m$, with given preferences $\{\succcurlyeq_i\}_{i=1,\ldots,m}$, indexed by varying individual endowments $(e_1, \cdots, e_m) \in \mathbf{R}^{mL}_{++}$. Denote the social endowment $\overline{e} := \sum_i e_i$ and a particular profile of individual endowments by $e := (e_1, \cdots, e_m) \in \mathbf{R}^{mL}_{++}$. An economy $E(e)$ then refers to the exchange economy with preferences $\{\succcurlyeq_i\}_{i=1,\ldots,m}$ and endowment profile e. For simplicity this article focuses on exchange economies. Mas-Colell (1985) is a comprehensive reference that includes discussion of extensions allowing for production.

The crucial departure in Debreu (1970) is to view each economy as a member of this parameterized family, and to ask whether perhaps almost no economies exhibit indeterminacy or pathological comparative statics when indexed this way. To formalize this, Debreu (1970) summarizes an agent's choice behaviour by a C^1 demand function $x_i : \mathbf{R}^L_{++} \times \mathbf{R}^L_{++} \to \mathbf{R}^L_+$ satisfying basic properties such as homogeneity of degree 0 in prices, Walras's Law, and boundary conditions as prices converge to zero. This leads to the familiar characterization of equilibria as zeros of excess demand:

$$0 = z(p, e) := \sum_i x_i(p, e_i) - \overline{e}.$$

Two simplifying normalizations are then commonly adopted. Demand functions derived from optimal choices of price-taking agents are homogeneous of degree zero in prices, so normalize by setting $p_1 \equiv 1$. Normalized prices thus can be taken to range over $\mathbf{R}^{L-1}_{++}$. Next, Walras's Law ensures that excess demand functions are not independent across markets, as $p \cdot z(p, e) = 0$ for each $p \in \mathbf{R}^{L-1}_{++}$. This renders one market clearing equation redundant, and leads to the characterization of equilibria by normalized price vectors $p \in \mathbf{R}^{L-1}_{++}$ such that

$$z_{-L}(p, e) = 0$$

where, adopting common conventions, the subscript $-L$ refers to all goods except L, so $z_{-L}(p, e) = (z_1(p, e), \ldots, z_{L-1}(p, e))$. Using these normalizations, the equilibrium correspondence can be defined by

$$E(e) := \{(x, p) \in \mathbf{R}^{mL}_+ \times \mathbf{R}^{L-1}_{++} : z_{-L}(p, e) = 0, x_i = x_i(p, e) \text{ for } i = 1, \ldots, m\}.$$

Fix a particular equilibrium price vector p^* in the economy $E(e)$. One way to answer local comparative statics qsts at this equilibrium is to apply the classical implicit function theorem. If $D_p z_{-L}(p^*, e)$ is invertible, then the implicit function theorem provides several immediate predictions: the equilibrium price p^* is locally unique; locally, on neighbourhoods W of e and V of p^*, the equilibrium price set is described by the graph of a C^1 function $p : W \to \mathbf{R}^{L-1}_{++}$; and local comparative statics are given by the formula

$$Dp(e) = -\left[D_p z_{-L}(p^*, e)\right]^{-1} D_e z_{-L}(p^*, e).$$

If this analysis can be performed for each equilibrium, then there are only finitely many equilibria, because the equilibrium set is compact. Moreover, for each equilibrium $(x, p) \in E(e)$ there is a neighborhood U of (x, p) for which $E(\cdot) \cap U$ has a unique, C^1 selection on a neighbourhood W of e, with the comparative statics derived from the preceding formula. Call such a correspondence *locally* C^1 *at* e. The following definition offers a convenient way to summarize these properties.

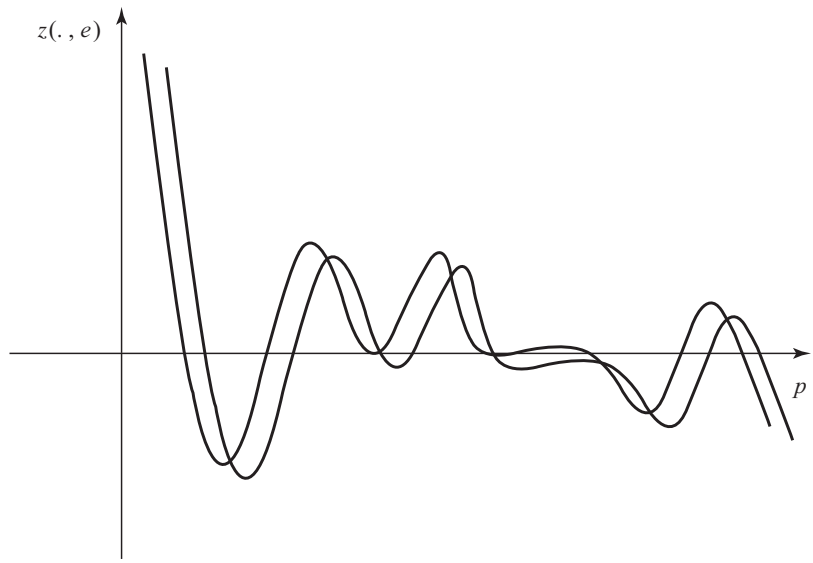
Definition 1 The economy $E(e)$ is 'regular' if it has finitely many equilibria, and E is locally C^1 at e.

An alternative way to describe the problem uses the language of differential topology. For a C^1 function f: $\mathbf{R}^m \to \mathrm{R}^n$, $y \in \mathbf{R}^n$ is a *regular value* of f if $Df(x)$ has full rank for every $x \in f^{-1}(y)$. Notice that this is precisely the condition identified above, for the case of equilibrium prices, under which local uniqueness and local comparative statics could be derived from the implicit function theorem. Whenever 0 is a regular value of $z\_{\_L}(\,,e)$, the corresponding economy $E(e)$ is regular. For a fixed function f, a given value y may fail to be a regular value, but almost every other value is regular: this is the conclusion of Sard's theorem. Dually, the fixed value y may fail to be a regular value for a particular function f, but is a regular value for almost every other function. When the set of functions is limited to those drawn from a particular parameterized family, the conclusion remains valid for almost all members of this family provided the parameterization is sufficiently rich. This idea of a rich parameterization can be expressed by requiring y to be a regular value of the parameterized family, and this parametric version of Sard's theorem is typically called the transversality theorem. Fig. 1 depicts this idea for smooth excess demand functions.

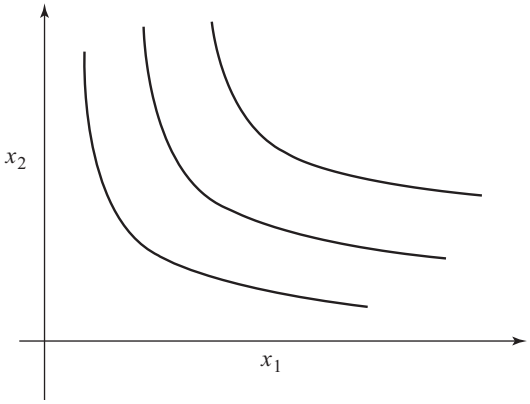
These observations suggest that, while extremely restrictive assumptions might be required to ensure that *every* economy is regular, generic regularity might follow simply from the differentiability of demand functions once the problem is framed this way. Straightforward calculations verify that 0 is a regular value of the excess demand function (viewed as a function of both prices and initial endowment parameters). From the transversality theorem we conclude that there is a subset $R^* \subset \mathbf{R}^{mL}_{++}$ of full Lebesgue measure such that for all $e \in R^*$, $E(e)$ is regular. With the use of additional properties of excess demand and equilibria, it is similarly straightforward to show that the set of regular economies is also open, giving a strong genericity result for regular economies.

This discussion follows Debreu's original development original development closely. This approach takes demand functions as primitives, and gives conditions on individual demand functions under which regularity is a generic feature of exchange economies. To take a step back and start with preferences as primitives, we seek conditions on preferences sufficient to guarantee that individual demand is suitably differentiable. Debreu (1972) addresses this point by introducing a class of 'smooth preferences', depicted in Fig. 2.

Definition 2 *The preference order* $\succcurlyeq$ *on* $\mathbf{R}^L_+$ *is* '*smooth*' *if it is represented by a utility function U such that*



Determinacy and Indeterminacy of Equilibria, Fig. 1 Generic determinacy for smooth excess demand



Determinacy and Indeterminacy of Equilibria, Fig. 2 Smooth preferences

- $U : \mathbf{R}^L_+ \to \mathbf{R}$ is C2 on $\mathbf{R}^L_{++}$
- for each $x \in \mathbf{R}^L_{++}$, $\{y \in \mathbf{R}^L_+ : y \tilde{\succ} x\} \subset \mathbf{R}^L_{++}$
- for each $x \in \mathbf{R}^L_{++}$, DU (x) $\gg 0$
- for each $x \in \mathbf{R}^L_{++}$, D2U (x) is negative definite on ker

  $DU(x) := \{z \in \mathbf{R}^L : DU(x)\, z = 0\}$

  Fairly straightforward arguments, again using the implicit function theorem, establish that individual demand functions derived from smooth preferences are C^1. Putting all of these results together yields:

Theorem 1 *Let $\succcurlyeq$ be a smooth preference order on $\mathbf{R}^L_+$ for each $i = 1, \ldots, m$. There exists an open set $R^* \subset \mathbf{R}^{mL}_{++}$ of full Lebesgue measure such that for all $e \in R^*$, $E(e)$ is regular.*

Determinacy and Indeterminacy: A New Approach to Many Problems

Behind this result for equilibria in finite exchange economies is a broad, powerful, and simple principle that has found many important and ingenious applications in the 35 years since Debreu's original 1970 paper. To cast the problem more generally, take a parameterized family of equations, captured by a function f: $\mathbf{R}^m \times \mathbf{R}^k \to \mathbf{R}^n$. This describes a problem with m variables and k parameters simultaneously entering n different equations. Imagine that for each parameter value $r \in \mathbf{R}^k$,

$$E(r) := \{x \in \mathbf{R}^m : f(x, r) = 0\}$$

gives the set of objects of interest. Moreover, imagine that the equations are sufficiently independent in determining the solutions, in the sense that 0 is a regular value of f. Counting the number of equations and unknowns produces three distinct cases, corresponding in turn to three different sorts of applications.

In the canonical case exemplified by the simple exchange economy described above, the number of relevant endogenous variables, m, is equal to the number of equations, n. In this case, 0 being a regular value of f characterizes exactly the case in which the equations are sufficiently independent that the loose 'counting equations and unknowns' heuristic corresponds with the precise technical result of generic determinacy. One prominent illustration of this case is given by two-period incomplete markets models with real assets, that is, assets that pay off in bundles of commodities. In these models, there are as many distinct budget equations as there are states. If we let S denote the number of states, this means there are $S + 1$ distinct Walras's Law statements, leading to $S + 1$ redundant market clearing equations. Because asset payoffs are in real terms, all budget constraints are homogeneous of degree 0 in state prices. This generates $S + 1$ distinct normalizations of state prices, compensating exactly for the drop in independent market clearing equations determining equilibrium. Generic determinacy in this case is established by Geanakoplos and Polemarchakis (1987).

When $m < n$, there are fewer equations than unknowns, and the regularity of the system of equations means that it is generically overdetermined. In this case, generically it is impossible to satisfy the equations simultaneously, that is, generically $E(r)$ is empty. As a simple example of this argument, consider the prevalence of trade at equilibrium in an Edgeworth box economy. One market-clearing condition in one (normalized) price characterizes equilibria, and standard arguments show that this excess demand function has 0 as a regular value. In fact, varying

the endowment of the first agent alone is enough. How often does equilibrium involve trade in some goods? With only two agents, trade occurs in equilibrium if and only if $x_2 \neq e_2$, so the additional two equations $x_2(p, e) - e_2 = 0$ characterize endowment and price combinations for which there is no trade in equilibrium. A simple calculation shows that 0 is a regular value of $f(p, e) := (z_{-2}(p, e), x_2(p, e) - e_2)$. Fixing the endowment profile e, however, this is a problem with three equations in a single variable, so there must be a set $R^{**} \subset \mathbf{R}^{mL}_{++}$ of full Lebesgue measure such that for every $e \in R^{**}$, there are no solutions to the equation $f(p, e) = 0$. For every endowment profile $e \in R^{**}$, every equilibrium then must involve trade, as every equilibrium price solves the first equation $z_{-2}(p^*, e) = 0$, so cannot also involve no trade, $x^2(p, e) \neq e_2$. Similar logic but more involved calculations show that equilibrium allocations are generically inefficient in incomplete markets models, and generically constrained inefficient in multi-good incomplete markets models. Geanakoplos and Polemarchakis (1987) pioneered this approach to efficiency with incomplete markets.

Finally, when $m > n$, generically indeterminacy arises, as generically the solution set $E(r)$ is an $(m - n)$-dimensional manifold. (A subset $M \subset \mathbf{R}^m$ is a $d -$ dimensional C^ℓ manifold if for each x A M there exist open sets $V \subset \mathbf{R}^m$ and $W \subset \mathbf{R}^d$, where V is a neighbourhood of x, and a C^ℓ diffeomorphism $\varphi: V \to W$ such that $\varphi(V \cap M) = W$). In this case, generically there is a continuum of solutions, and the set of solutions is locally, up to diffeomorphism, a set of dimension $m - n$. An important example of this case is provided by two-period incomplete financial markets models with nominal assets. Here, asset payoffs are in nominal terms, in some specified unit of account. As in the case of real assets described above, there are $S + 1$ independent budget constraints when there are S possible states of nature, so there are $S + 1$ redundant market clearing equations. Because asset payoffs are nominal, however, budget constraints are not all homogeneous of degree zero, and price levels matter. With only two homogeneity conditions, one for period one prices and one relating all commodity and asset prices, this leaves $S - 1$ dimensions of indeterminacy in equilibria generically. The detailed result is established by Geanakoplos and Mas-Colell (1989).

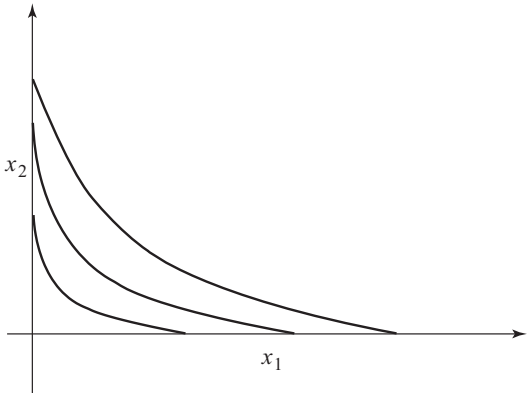
These three cases, and the generic properties of solution sets that follow, are collected below.

Theorem 2 *Let f: $\mathbf{R}^m \times \mathbf{R}^k \to \mathbf{R}^n$ be a C^ℓ function, where $\ell > \max\{m - n, 0\}$, and suppose 0 is a regular value of f.*

(a) Suppose m = n. There exists a set $R^* \subset \mathbf{R}^k$ of full Lebesgue measure such that, for every r $\in$ R*, E(r) contains only isolated points, E (r) is finite when compact, and E is locally C^1 at r.
(b) Suppose m < n. There exists a set $R^* \subset \mathbf{R}^k$ of full Lebesgue measure such that for every r $\in$ R*, E(r) is empty.
(c) Suppose m > n. There exists a set $R^* \subset \mathbf{R}^k$ of full Lebesgue measure such that for every r $\in$ R*, E(r) is an (m − n)-dimensional C^ℓ manifold.

The techniques pioneered by Debreu have found widespread applications, and have proven to be rmkably powerful. Nonetheless, the smoothness needed to study determinacy using the tools of differential topology does stem from assumptions that often carry real economic content. These assumptions restrict both the nature of admissible preferences and the nature of admissible constraints.

For example, to avoid problems arising when non-negativity constraints on consumption may become binding, these results rest on 'boundary' restrictions, both on endowments, because individual endowments are strictly positive, or on equilibrium consumption via boundary conditions on preferences that imply individual demands are strictly positive at all prices. Unless goods are aggregated extremely coarsely, neither pattern is supported by observations on consumer behaviour or characteristics. Relaxing the constraint on endowments turns out, perhaps surprisingly, to generate indeterminacy much more readily than relaxing the assumptions on positive consumptions, or incorporating other more general constraints on choices. Minehart (1997) shows by



Determinacy and Indeterminacy of Equilibria, Fig. 3 Preferences allowing boundary consumption

means of an example that for one natural case of restricted endowments, in which each agent is constrained to hold a single, individual-specific, good, an open subset of such parameters leads to indeterminacy in equilibrium. Highlighting the fact that the choice of parameterization can be important, Mas-Colell (1985) shows that this conclusion is not robust to perturbations in preferences; generic determinacy, in a topological sense, is restored by considering variations in preferences as well as constrained endowments. If the assumption that individual endowments of every good are positive is maintained, the restriction to positive individual demand for every good can be relaxed. For example, Mas-Colell (1985) provides generic determinacy results for exchange economies allowing for boundary consumptions; Fig. 3 depicts such preferences.

Smooth preferences, as defined by Definition 2 above, obviously rule out preferences with non-differentiabilities in level sets, a restriction that also has important behavioural content. Kinks have arisen as central manifestations of various behavioural phenomena, including loss aversion, ambiguity aversion, and reference dependence; examples include Kahneman and Tversky (1979), Tversky and Kahneman (1991), Koszegi and Rabin (2006), Sagi (2006), and Gilboa and Schmeidler (1989). Such kinks typically lead to excess demand functions that fail to be differentiable for some prices. Rader (1973), Pascoa and Werlang (1999), Shannon (1994), and Blume and Zame (1993) all develop methods to address such cases. With the exception of Blume and Zame (1993), these techniques can be roughly understood as expanding differential notions by adding to 'regularity' the condition that the function (for example, excess demand) is differentiable at every solution, and establishing that analogues of implicit function theorems, Sard's theorem or the transversality theorem remain valid for sufficiently nice non-smooth functions, such as Lipschitz continuous functions; in particular, see Shannon (1994, 2006). Blume and Zame (1993) instead use results that exploit the structure of algebraic sets to establish generic determinacy for utilities that are, roughly, finitely piecewise analytic, and need not be strictly concave. Examples in which determinacy has been studied using techniques along these various lines include asset market models with restricted participation (for example, see Cass et al. 2001) and models of ambiguity aversion (for example, see Rigotti and Shannon 2006). Fig. 4.

Determinacy in Infinite-Dimensional Economies

Many economic models require an infinite number of marketed commodities. Important examples include dynamic infinite horizon economies, continuous-time trading in financial markets, and markets with differentiated commodities. Such infinite-dimensional models present big obstacles to studying determinacy, starting with the fact that individual demand is not defined for most prices, precluding any straightforward parallel of Debreu's arguments for finite economies. In addition, the positive cone in most infinite-dimensional spaces has empty interior in the relevant topologies, meaning individual consumption sets are 'all boundaries', and existence of equilibrium typically requires conditions, such as uniform properness or variants, that effectively bound marginal rates of substitution. Thus boundary conditions akin to those in Debreu's smooth preferences are likely either to be impossible to satisfy or to contradict equilibrium existence in many important applications.

Provided there are finitely many agents and no market distortions, using the welfare theorems and Negishi's argument provides an alternative characterization of equilibria, replacing excess demand with 'excess savings'. Some version of this characterization of equilibria provides the framework for much of the existing equilibrium analysis in economies with infinitely many commodities, including the seminal work on existence of Mas-Colell (1986) and Aliprantis et al. (1987), and the approach to determinacy for discrete-time infinite horizon models with time separability pioneered by Kehoe and Levine (1985). To explain this, let X denote the commodity space. The efficient allocations are the solutions to a social planner's problem of the following form: given $\lambda \in \Lambda := \left\{\lambda \in \mathbf{R}^m_+ : \sum_{i=1}^m \lambda_i = 1\right\}$, choose a feasible allocation $x(\lambda)$ to solve:

$$\max \quad \sum_{i=1}^{m} \lambda_i U_i(x_i) s.t \sum_{i=1}^{m} x_i \leq \overline{e}.$$

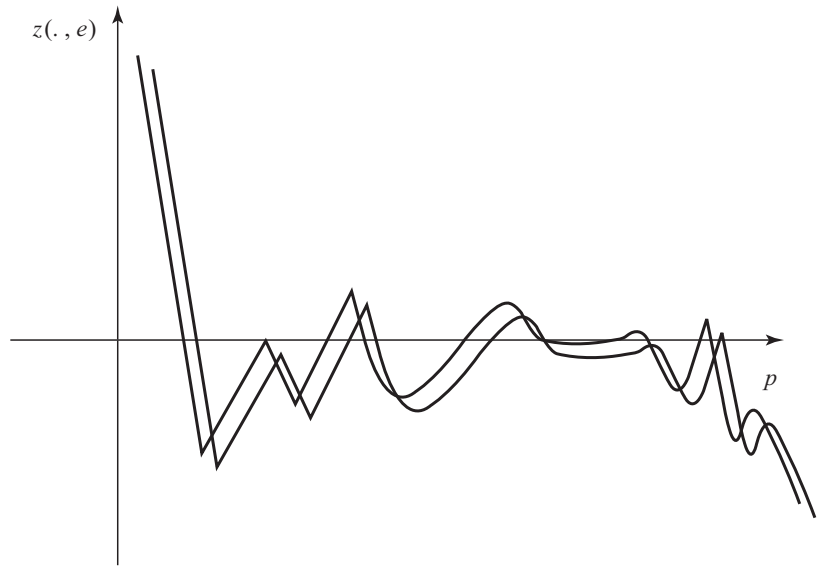
Under standard assumptions, the solution $x(\lambda)$ to this problem is well-defined and unique for each $\lambda \in \Lambda$, and a unique price $p(\lambda)$ supporting $x(\lambda)$ can be characterized. Equilibria then correspond to the solutions λ to the budget equations

$$p(\lambda) \cdot (x_2(\lambda) - e_2) = 0$$
$$\vdots$$
$$p(\lambda) \cdot (x_m(\lambda) - e_m) = 0$$

where Walras's Law accounts for the missing equation. In parallel with excess demand, define the *excess savings map* $s : \Lambda \times X_+^m \to \mathbf{R}^{m-1}$

$$s(\lambda, e) :$$
$$= (p(\lambda) \cdot (x_2(\lambda) - e_2), \ldots, p(\lambda) \cdot (x_m(\lambda) - e_m)).$$

Through this construction, the question of determinacy for infinite-dimensional economies can be cast in close parallel to finite economies, with the only change that the set of parameters is now infinite-dimensional. This raises several technical issues, most importantly the choice between topological and measure-theoretic notions of genericity due to the impossibility of defining a suitable analogue of Lebesgue measure in infinite-dimensional spaces (see Hunt et al. 1992, and Anderson and Zame 2001, for a discussion of these issues). This construction also makes imperative the need to link conditions on excess savings used to imply determinacy with conditions on preferences since, in contrast with excess demand, excess savings depends on artificial and unobservable



Determinacy and Indeterminacy of Equilibria, Fig. 4 Generic determinacy for non-smooth excess demand

constructs. Somewhat surprisingly, Shannon (1999) and Shannon and Zame (2002) show that generic determinacy follows from conditions on preferences that closely resemble Debreu's (1972) smooth preferences, after suitable renormalization. As in the finite case, these conditions can roughly be understood as strengthened notions of concavity, requiring that near feasible bundles utility differs from a linear approximation by an amount quadratic in the distance to the given bundle. These notions of concavity thus rule out preferences displaying local or global substitutes. Shannon and Zame (2002) provide a simple geometric argument showing that the excess spending mapping is Lipschitz continuous. Generic determinacy then follows by arguments similar to those sketched above for other problems with non-differentiabilities, making use of Shannon (2006) on comparative statics and a version of the transversality theorem for this setting. The direct, geometric nature of these arguments render them applicable in a wide range of examples, including models of continuous-time trading, trading in differentiated commodities, and trading over an infinite horizon.

Bibliography

Aliprantis, C., D.J. Brown, and O. Burkinshaw. 1987. Edgeworth equilibria. *Econometrica* 55: 1108–1138.

Anderson, R., and W.R. Zame. 2001. Genericity with infinitely many parameters. *Advances in Theoretical Economics* 1.

Blume, L., and W.R. Zame. 1993. The algebraic geometry of competitive equilibrium. In *Essays in general equilibrium and international trade: In memoriam trout rader*, ed. W. Neuefeind. New York: Springer-Verlag.

Cass, D., P. Siconolfi, and A. Villanacci. 2001. Generic regularity of competitive equilibrium with restricted participation on financial markets. *Journal of Mathematical Economics* 36: 61–76.

Debreu, G. 1970. Economies with a finite set of equilibria. *Econometrica* 38: 387–392.

Debreu, G. 1972. Smooth preferences. *Econometrica* 40: 603–615.

Dierker, E. 1972. Two remarks on the number of equilibria of an economy. *Econometrica* 40: 951–953.

Geanakoplos, J., and A. Mas-Colell. 1989. Real indeterminacy with financial assets. *Journal of Economic Theory* 47: 22–38.

Geanakoplos, J., and H. Polemarchakis. 1987. Existence, regularity and constrained suboptimality of competitive portfolio allocations when the asset market is incomplete. In *Essays in honor of kenneth J arrow,* vol 3, ed. W.P. Heller, R.M. Starr, and D.M. Starrett. Cambridge: Cambridge University Press.

Gilboa, I., and D. Schmeidler. 1989. Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics* 18: 141–153.

Hunt, B.R., T. Sauer, and J.A. Yorke. 1992. Prevalence: A translation invariant 'almost every' on infinite dimensional spaces. *Bulletin (New Series) of the American Mathematical Society* 27: 217–238.

Kahneman, D., and A. Tversky. 1979. Prospect theory: An analysis of decision under risk. *Econometrica* 47: 263–291.

Kehoe, T., and D. Levine. 1985. Comparative statics and perfect foresight in infinite horizon economies. *Econometrica* 53: 433–452.

Koszegi, B., and M. Rabin. 2006. A model of reference-dependent preferences. *Quarterly Journal of Economics* 121: 1133–1165.

Mas-Colell, A. 1985. *The theory of general economic equilibrium: A differentiable approach*. Cambridge: Cambridge University Press.

Mas-Colell, A. 1986. The price equilibrium existence problem in topological vector lattices. *Econometrica* 54: 1039–1054.

Minehart, D. 1997. A note on the generic finiteness of the set of equilibria in an exchange economy with constrained endowments. *Mathematical Social Sciences* 34: 75–80.

Pascoa, M.R., and S. Werlang. 1999. Determinacy of equilibrium in nonsmooth economies. *Journal of Mathematical Economics* 32: 289–302.

Rader, J.T. 1973. Nice demand functions. *Econometrica* 41: 913–935.

Rigotti, L., and C. Shannon. 2006. Sharing risk and ambiguity. Discussion paper. UC Berkeley.

Sagi, J. 2006. Anchored preference relations. *Journal of Economic Theory* 130: 283–295.

Shannon, C. 1994. Regular nonsmooth equations. *Journal of Mathematical Economics* 23: 147–166.

Shannon, C. 1999. Determinacy of competitive equilibria in economies with many commodities. *Economic Theory* 14: 29–87.

Shannon, C. 2006. A prevalent transversality theorem for Lipschitz functions. *Proceedings of American Mathematical Society* 134: 2755–2765.

Shannon, C., and W.R. Zame. 2002. Quadratic concavity and determinacy of equilibrium. *Econometrica* 70: 631–662.

Tversky, A., and D. Kahneman. 1991. Loss aversion in riskless choice: A reference dependent model. *Quarterly Journal of Economics* 106: 1039–1061.

Determinism

Shaun Hargreaves-Heap and Martin Hollis

The proposition that every event has a cause sounds clear and simple. It is neither. On a very strong reading it asserts a grand inevitability about the workings of the universe, which leaves only one course of history possible. Many economists will associate determinism, taken in this sense, with Marx (1858):

> In the social production which men carry on they enter into definite relations that are indispensable and independent of their will; these relations of production correspond to a definite stage of development of their material powers of production. The sum total of these relations of production constitutes the economic structure of society – the real foundation, on which rise legal and political superstructures and to which correspond definite forms of social consciousness. . . . It is not the consciousness of men that determines their existence, but, on the contrary, their social existence determines their consciousness. (*A Contribution to the Critique of Political Economy*, Preface)

Since anyone who agrees is plainly a determinist, it is easy to presume that those who disagree with Marx are not. For, on this account, determinism seems opposed to freedom, because it excludes all individual voluntarism. Indeed, the passage contrasts so starkly with neoclassical analyses of choice, where the emphasis is on the role of individuals, that the opposition between determinism and individual freedom appears to capture a vital difference between marxian and neoclassical economists. But this would be too casual an appeal to a popular distinction between freedom and determinism, doing justice to neither school.

The first point to note is that determinism need not be as specific as in the economic variety pressed by Marx above. The proposition that every event has a cause, on a weaker reading, claims only that every event is predictable from its antecedents. The prediction need not be underpinned by Marx's relations and forces of production nor, more generally, by any set of natural laws, which involve real necessities. According to Hume (1748), prediction requires causal laws only in the sense of correlations or patterns which we have come to notice and expect. If he is right in his analysis of causation, then all economists, who seek empirical regularities in order to predict and explain effects from their antecedents, are determinists.

For example, neoclassical economists use causal laws to explain action. The neoclassicist conceives of economic man in such a way that his actions can in principle be predicted, given his environment, preferences and stock of information. The antecedent cause of an action is preference combined with environment and information. This is plain in the simple model of consumer choice or production, where the agent is no more than a mechanical throughput between environment and outcome. It also applies, however, in more complex models, where outcomes are not uniquely determined. Stochastic models are often referred to as non-deterministic, and mixing strategies can produce varying outcomes in the same game. Nevertheless, they are examples of determinism on a weaker, Humean, definition. The only difference is that what is being determined (or predicted) is not unique events but the probability distributions of events.

This may surprise those who take neoclassical economics to be wedded to ideas of free choice which are incompatible with determinism. After all, it is easy to assume that a caused action cannot be a chosen action. But many philosophers (compatibilists) deem this a mistake. Free actions are those which translate the agent's wants into effect; those where the agent is compelled to act against his wishes are unfree. Since *both* kinds of action are caused, however, the traditional dispute between free will and determinism has been misconceived (see e.g. Hume 1748; Mill 1843; Ayer 1954). Indeed, the point carries over to the more striking claim that actions can be free only if they are also determined. Lack of freedom is a matter of interference between the agent's wants, his action and its intended outcome. Typical obstacles to freedom are coercion, ignorance and randomness. Free agents therefore need to be able to count on their actions being in conformity with

their wants and on consequences being as foreseen. Where they cannot count on these causal sequences, they cannot be sure that they will satisfy their preferences by their actions. The advance of a deterministic science of economics is thus a positive help in making the outcome less clouded and opaque and hence in choosing the actions which satisfy preferences.

This compatibilist line reconciles freedom and determinism in a way which lends weight to the popular distinction between positive and normative economics. Normative economics is concerned with what is good or bad, right or wrong. It includes value judgements about which causes of action are morally justified and which outcomes morally desirable. It seeks to set the boundary between coerced and free actions. But these are normative distinctions within the realm of caused actions and outcomes. Positive economics, by contrast, views the same realm solely for purposes of explanation. It asks why economic agents behave as they do and not whether they should. This kind of question is distinct from the others and will only be hampered by letting normative disputes obscure it. Indeed, by being left to get on with advancing deterministic explanations, and thus increasing knowledge of how to bring about desired ends, it offers an ever more powerful service to those individuals or policy-makers, whose normative judgements are to be respected.

Compatibilism is thus a tempting doctrine for any economist who shares Enlightenment hopes that a determinist, predictive science is an aid to progress in the pursuit of human happiness (or any other prescribed goal). It is also a useful doctrine for those who maintain the methodological unity of the sciences and the modelling of the social sciences on a positivist account of the natural sciences. But its reconciliation of freedom with determinism does not go unchallenged. There is bound to be a suspicion that, in assimilating human beings to other complex creatures or objects in nature and in embracing a scientific method designed for nature, a special quality of free human action has been lost. The suspicion is hard to focus precisely but Austin (1961) offers a sharp challenge. It is not enough, he holds, for an action to be free that, in other conditions or with other preferences, the agent *would have* acted differently. Free will requires that the agent *could have* chosen differently in the same conditions. 'Could have' does not reduce to 'would have, if . . .'. This implies that an economic science, which treats economic decision as a throughput between environment and outcome, *via* preferences which are exogenously given, must somehow be denying free will.

Compatibilists retort that an alternative choice in the same conditions and with the same preferences would be a very puzzling phenomenon. In special cases a rational agent might deliberately adopt a randomizing strategy. But, in general, it looks as if free will is being supposed to require an unpredictable choice of action and hence an inexplicable one. If free actions cannot be explained in terms of what an agent in those circumstances, with those wants and beliefs, predictably did, then they are beyond the scope of scientific enquiry.

To dispose of the retort one needs a rival to causal explanation. An obvious candidate is rational or intentional explanation, although it may turn out not to be genuinely a rival. This kind of explanation focuses less on the situation and more on the agent's own understanding of it. The agent intends a certain result and has reason to believe that the chosen action is the likeliest way to achieve it. Explanation becomes the rational reconstruction of this process of decision. Provided that agents' judgements are reconstructed as not automatic but the work of their own understanding, perhaps the agent *could have done otherwise* in the same conditions, as Austin required. But a shift of emphasis from environment to psychology will not be philosophically significant, unless psychological explanations are non-causal. So the obvious counter-move is to point out that psychology too has often been regarded as a law-governed deterministic science.

Yet that may be to miss the point that the social world depends on agents' beliefs about it, whereas the natural world does not depend on what atoms believe. Granted a distinction between regularities in nature and rules or norms in human society, one might suggest that the social world is furnished

and activated in ways which call for a different approach to psychology. At any rate, whether or not the workings of the mind are causal and law-like, the beliefs of agents affect the outcomes which economists observe. This peculiarity of social life is itself enough to open some interesting possibilities. For example, Keynes can be read as harnessing the point to show how different but plausible beliefs in the face of uncertainty would lead to different outcomes. The more recent literature finding rational expectations consistent with multiple equilibria takes it further. Choices become genuine, in Austin's sense, because what happens genuinely depends on what agents expect to happen.

Uncertainty and freedom become, so to speak, two sides of the same coin (cf. Shackle 1969). The theme is developed in the post-Keynesian reminders that historical time differs from logical time. In logical time the series of events is complete and only human ignorance distinguishes the known past from the unknown future. Natural science theories commonly conceive the world in this timeless way (although perhaps adding direction through an idea of irreversible change). Economics, on the other hand, can or even should accommodate the thought that agents can make a future discontinuous with the past. They are not just discovering what a supreme intelligence could have predicted from the start, given the first state of the world and the forces acting in it. (This is a reference to Laplace 1820.) If so, economic theories need to recognize the fact that economic events occur in historical time, in a way which natural science theories can abstract from.

One consequence is to undermine the positive/normative distinction. Economic agents are affected by beliefs, including beliefs about how the economy works. Economic theories (unless kept secret) thus affect outcomes. Economic science ceases to be an impartial description and, since rival theories involve differing normative commitments, the material for an allegedly positive science is hopelessly corrupted by normative elements in its data and by its own feedback through the beliefs of its subject matter. Whether an observed correlation amounts to a causal law threatens to depend on whether the positive economist can persuade people that it does! In place of the image of the economist as technician, we have an image of the economist as ideologue *malgré lui,* and the vision of economics as a moral science is restored.

This article has separated two ways of construing the relevance of determinism to economics. The more familiar starts from the popular contrast between determinism and freedom and with the tendency in economics to associate the former with Marx. This creates an impression that neoclassical economics rejects determinism, especially given the volume of neoclassical work on stochastic models. But if determinism claims only that outcomes are governed by causal laws and that events are accordingly predictable from their antecedents, the popular contrast disappears. Marxism and neoclassicism become alternative deterministic theories. There is still room for argument, however, about the success of compatibilism. To insist that choice implies 'could have acted otherwise' and not merely 'would have, if . . .' is to reinstate the dispute. It then matters whether economic agents have an open future, whereas (random factors aside) atoms do not. If so, uncertainty in economics differs crucially from anything implied by an Uncertainty Principle in physics. In particular, economic theories may affect what people believe and hence what economic science tries to describe and explain. In that case economics cannot help being a normative science.

See Also

- ▶ Dialectical Materialism
- ▶ Dialectical Reasoning
- ▶ Economic Interpretation of History

Bibliography

Ayer, A.J. 1954. Freedom and necessity. In *Philosophical essays*, ed. A.J. Ayer. London: Macmillan.

Austin, J.L. 1961. Ifs and cans. In *Philosophical papers*, ed. J.L. Austin. Oxford: Clarendon Press.

Hume, D. 1748. *An enquiry concerning human understanding*. Sect. VII, pts I and II.

Laplace, P. 1820. *Théorie analytique des probabilités*. Paris.
Marx, K. 1858. *A contribution to the critique of political economy*. Moscow: Progress Publishers, 1970.
Mill, J.S. 1843. *A system of logic*. Book VI. London: J.W. Parker.
Shackle, G.L.S. 1969. *Decision, order and time*. Cambridge: Cambridge University Press.

Deterministic Evolutionary Dynamics

William H. Sandholm

Abstract

We review the literature on deterministic evolutionary dynamics in game theory. We describe the micro-foundations of dynamic evolutionary models and offer some basic examples. We report on stability theory for evolutionary dynamics, and we discuss the senses in which evolutionary dynamics support and fail to support traditional game-theoretic solution concepts.

Keywords

Asymptotic stability; Convergence; Deterministic evolutionary dynamics; Evolutionarily stable strategies; Evolutionary dynamics; Extensive form games; Game theory; Imitative dynamics; Index theory; Lyapunov functions; Markov processes; Mean dynamic; Poincaré–Hopf theorem; Population games; Predictions; Refinements of Nash equilibrium; Replicator dynamic; Revision protocols

JEL Classifications

C7

Introduction

Deterministic evolutionary dynamics for games first appeared in the mathematical biology literature, where Taylor and Jonker (1978) introduced the *replicator dynamic* to provide an explicitly dynamic foundation for the static evolutionary stability concept of Maynard Smith and Price (1973). But one can find precursors to this approach in the beginnings of game theory: Brown and von Neumann (1950) introduced differential equations as a tool for computing equilibria of zero-sum games. In fact, the replicator dynamic appeared in the mathematical biology literature long before game theory itself: while Maynard Smith and Price (1973) and Taylor and Jonker (1978) studied game theoretic models of animal conflict, the replicator equation is equivalent to much older models from population ecology and population genetics. These connections are explained by Schuster and Sigmund (1983), who also coined the name 'replicator dynamic', borrowing the word 'replicator' from Dawkins (1982).

In economics, the initial phase of research on deterministic evolutionary dynamics in the late 1980s and early 1990s focused on populations of agents who are randomly matched to play normal form games, with evolution described by the replicator dynamic or other closely related dynamics. The motivation behind the dynamics continued to be essentially biological: individual agents are preprogrammed to play specific strategies, and the dynamics themselves are driven by differences in birth and death rates. Since that time the purview of the literature has broadened considerably, allowing more general sorts of large population interactions, and admitting dynamics derived from explicit models of active myopic decision making.

This article provides a brief overview of deterministic evolutionary dynamics in game theory. More detailed treatments of topics introduced here can be found in the recent survey article by Hofbauer and Sigmund (2003), and in books by Maynard Smith (1982), Hofbauer and Sigmund (1988, 1998), Weibull (1995), Vega-Redondo (1996), Samuelson (1997), Fudenberg and Levine (1998), Cressman (2003), and Sandholm (2007).

Population Games

Population games provide a general model of strategic interactions among large numbers of

anonymous agents. For simplicity, we focus on games played by a single population, in which agents are not differentiated by roles; allowing for multiple populations is mostly a matter of introducing more elaborate notation.

In a single-population game, each agent from a unit-mass population chooses a strategy from the finite set $S = \{1, \ldots, n\}$, with typical elements i and j. The distribution of strategy choices at a given moment in time is described by a *population state* $x \in X = \left\{x \in \mathbf{R}^n_+ : \sum_{i \in S} x_i = 1\right\}$. The *payoff* to strategy i, denoted $F_i : X \to \mathbf{R}$, is a continuous function of the population state; we use the notation $F : X \to \mathbf{R}^n$ to refer to all strategies' payoffs at once. By taking the set of strategies S as fixed, we can refer to F itself as a *population game*.

The simplest example of a population game is the most commonly studied one: random matching to play a symmetric normal form game $A \in R^{n \times n}$, where A_{ij} is the payoff obtained by an agent choosing strategy i when his opponent chooses strategy j. When the population state is $x \in X$, the expected payoff to strategy i is simply the weighted average of the elements of the ith row of the payoff matrix: $F_i(x) = \sum_{j \in S} A_{ij} x_j = (Ax)_i$. Thus, the population game generated by random matching in A is the linear population game $F(x) = Ax$.

Many models of strategic interactions in large populations that arise in applications do not take this simple linear form. For example, in models of highway congestion, payoff functions are convex: increases in traffic when traffic levels are low have virtually no effect on delays, while increases in traffic when traffic levels are high increase delays substantially (see Beckmann et al. 1956; Sandholm 2001). Happily, allowing nonlinear payoffs extends the range of possible applications of population games without making evolutionary dynamics especially more difficult to analyse, since the dynamics themselves are nonlinear even when the underlying payoffs are not.

Foundations of Evolutionary Dynamics

Formally, an *evolutionary dynamic* is a map that assigns to each population game F a differential equation $\dot{x} = V^F(x)$ on the state space X. While one can define evolutionary dynamics directly, it is preferable to derive them from explicit models of myopic individual choice.

We can accomplish this by introducing the notion of a *revision protocol* $\rho : \mathbf{R}^n \times X \to R^{n \times n}_+$. Given a payoff vector $F(x)$ and a population state x, a revision protocol specifies for each pair of strategies i and j a non-negative number $\rho^{ij}(F(x), x)$, representing the rate at which strategy i players who are considering switching strategies switch to strategy j. Revision protocols that are most consistent with the evolutionary paradigm require agents to possess only limited information: for example, a revising agent might know only the current payoffs of his own strategy i and his candidate strategy j.

A given revision protocol can admit a variety of interpretations. For one all-purpose interpretation, suppose each agent is equipped with an exponential alarm clock. When the clock belonging to an agent playing strategy i rings, he selects a strategy $j \in S$ at random, and then switches to this strategy with probability proportional to $\rho_{ij}(F(x), x)$. While this interpretation is always available, others may be simpler in certain instances. For example, if the revision protocol is of the imitative form $\rho_{ij} = x_j \times \hat{\rho}_{ij}$, we can incorporate the x_j term into our story by supposing that the revising agent selects his candidate strategy j not by drawing a strategy at random, but by drawing an opponent at random and observing this opponent's strategy.

A population game F and a revision protocol ρ together generate an ordinary differential equation $\dot{x} = V^F(x)$ on the state space X. This equation, which captures the population's *expected* motion under F and ρ, is known as the *mean dynamic* or *mean field* for F and ρ:

$$\begin{aligned} \dot{x} &= V_i^F(x) \\ &= \sum_{j \in S} x_j \rho_{ji}(F(x), x) \\ &\quad - x_i \sum_{j \in S} \rho_{ji}(F(x), x). \end{aligned} \tag{M}$$

The form of the mean dynamic is easy to explain. The first term describes the 'inflow' into

strategy i from other strategies; it is obtained by multiplying the mass of agents playing each strategy j by the rate at which such agents switch to strategy i, and then summing over j. Similarly, the second term describes the 'outflow' from strategy i to other strategies. The difference between these terms is the net rate of change in the use of strategy i.

To obtain a formal link between the mean dynamic (M) and our model of individual choice, imagine that the population game F is played not by a continuous mass of agents but rather by a large, finite population with N members. Then the model described above defines a Markov process $\{X_t^N\}$ on a fine but discrete grid in the state space X. The foundations for deterministic evolutionary dynamics are provided by the following finite horizon deterministic approximation theorem: Fix a time horizon $T < \infty$. Then the behaviour of the stochastic process $\{X_t^N\}$ through time T is approximated by a solution of the mean dynamic (M); the approximation is uniformly good with probability close to 1 once the population size N is large enough. (For a formal statement of this result, see Benaïm and Weibull 2003.)

In cases where one is interested in phenomena that occur over very long time horizons, it may be more appropriate to consider the infinite horizon behaviour of the stochastic process $\{X_t^N\}$. Over this infinite time horizon, the deterministic approximation fails, as a correct analysis must explicitly account for the stochastic nature of the evolutionary process. For more on the distinction between the two time scales, see Benaïm and Weibull (2003).

Examples and Families of Evolutionary Dynamics

We now describe revision protocols that generate some of the most commonly studied evolutionary dynamics. In the table below, $\overline{F}(x) = \sum_{i \in S} x_i F_i(x)$ represents the population's average payoff at state x, and $B^F(x) = argmax_{y \in X} y'F(x)$ is the best response correspondence for the game F.

A common critique of evolutionary analysis of games is that the choice of a specific revision protocol, and hence the evolutionary analysis that follows, is necessarily arbitrary. There is surely some truth to this criticism: to the extent that one's analysis is sensitive to the fine details of the choice of protocol, the conclusions of the analysis are cast into doubt. But much of the force of this critique is dispelled by this important observation: *evolutionary dynamics based on qualitatively similar revision protocols lead to qualitatively similar aggregate behaviour*. We call a collection of dynamics generated by similar revision protocols a 'family' of evolutionary dynamics.

To take one example, many properties that hold for the replicator dynamic also hold for dynamics based on revision protocols of the form $\rho_{ij} = x_j \hat{\rho}_{ij}$ where $\hat{\rho}_{ij}$ satisfies

$$\operatorname{sgn}\Big((\hat{\rho}_{ki} - \hat{\rho}_{ik}) - \big(\hat{\rho}_{kj} - \hat{\rho}_{jk}\big)\Big) = \operatorname{sgn}\big(F_i - F_j\big) \quad \text{for all} \quad k \in S.$$

(In words: if i earns a higher payoff than j, then the net conditional switch rate from k to i is higher than that from k to j for all $k \in S$.) For reasons described in Section 3, dynamics generated in this way are called 'imitative dynamics'. (See Björnerstedt and Weibull, 1996, for a related formulation.) For another example, most properties of the pairwise difference dynamic remain true for dynamics based on protocols of the form $\rho_{ij} = \phi(F_i - F_j)$, where $\phi : \mathbf{R} \to \mathbf{R}_+$ satisfies sign-preservation:

$$\operatorname{sgn}(\phi(d)) = \operatorname{sgn}\big([d]_+\big).$$

Dynamics in this family are called 'pairwise comparison dynamics'. For more on these and other families of dynamics, see Sandholm (2007, ch. 5).

Rest Points and Local Stability

Having introduced families of evolutionary dynamics, we now turn to questions of prediction: if agents playing game F follow the revision protocol ρ (or, more broadly, a revision protocol from

D

a given family), what predictions can we make about how they will play the game? To what extent do these predictions accord with those provided by traditional game theory?

A natural first question to ask concerns the relationship between the rest points of an evolutionary dynamic V^F and the Nash equilibria of the underlying game F. In fact, one can prove for a very wide range of evolutionary dynamics that if a state $x^* \in X$ is a Nash equilibrium (that is, if$x \in B(x)$), then x^* is a rest point as well.

One way to show that $NE(F) \subseteq RP(V^F)$ is to first establish a *monotonicity* property for V^F: that is, a property that relates strategies' growth rates under V^F with their payoffs in the underlying game (see, for example, Nachbar 1990; Friedman 1991; and Weibull 1995). The most general such property, first studied by Friedman (1991) and Swinkels (1993), we call 'positive correlation':

$$\text{If } x \notin = RP\left(V^F\right), \quad \text{then} \quad F(x)'V^F(x) > 0. \quad \text{(PC)}$$

Property (PC) is equivalent to requiring a positive correlation between strategies' growth rates $V_i^F(x)$ and payoffs $F_i(x)$ (where the underlying probability measure is the uniform measure on the strategy set S). This property is satisfied by the first three dynamics in Table 1, and modifications of it hold for the remaining two as well. Moreover, it is not difficult to show that if V^F satisfies (PC), then all Nash equilibria of F are rest points of V^F: that is, $NE(F) \subseteq RP(V^F)$, as desired (see Sandholm 2007, ch. 5).

In many cases, one can also prove that every rest point of V^F is a Nash equilibrium of F, and hence that $NE(F) = RP(V^F)$. In fact, versions of this statement are true for all of the dynamics introduced above, with the notable exception of the replicator dynamic and other imitative dynamics. The reason for this failure is easy to see: when revisions are based on imitation, unused strategies, even ones that are optimal, are never chosen. On the other hand, if we introduce a small number of agents playing an unused optimal strategy, then these agents will be imitated. Developing this logic, Bomze (1986) and Nachbar (1990) show that, under many imitative dynamics, every Lyapunov stable rest point is a Nash equilibrium.

As we noted at the onset, the original motivation for the replicator dynamic was to provide a foundation for Maynard Smith and Price's (1973) notion of an evolutionarily stable strategy (ESS). Hofbauer et al. (1979) and Zeeman (1980) show that an ESS is asymptotically stable under the replicator dynamic, but that an asymptotically state need not be an ESS.

More generally, when is a Nash equilibrium a dynamically stable rest point, and under which dynamics? Under differentiable dynamics, stability of isolated equilibria can often be determined by linearizing the dynamic around the equilibrium. In many cases, the question of the stability of the rest point x^* reduces to a question of the negativity of certain eigenvalues of the Jacobian matrix $DF(x^*)$ of the payoff vector field. In non-differentiable cases, and in cases where the equilibria in question form a connected component, stability can sometimes be established by using another standard approach: the construction of suitable Lyapunov functions.

Deterministic Evolutionary Dynamics, Table 1

Revision protocol	Evolutionary dynamic	Name	Origin
$\rho_{ij} = x_j(K - F_i)$, or $\rho_{ij} = x_j(K + F_j)$, or $\rho_{ij} = x_j[F_j - F_i]_+$	$\dot{x}_i = x_i\left(F_i(x) - \overline{F}(x)\right)$	Replicator	Taylor and Jonker (1978)
$\rho_{ij} = \left[F_j - \overline{F}\right]_+$	$\dot{x}_i = \frac{\left[F_i(x) - \overline{F}(x)\right]_+}{-x_i \sum_{j \in S}\left[F_j(x) - \overline{F}(x)\right]_+}$	Brown–von Neumann–Nash (BNN)	Brown and von Neumann (1950)
$\rho_{ij} = [F_j - F_i]_+$	$\dot{x}_i = \frac{\sum_{j \in S} x_j\left[F_i(x) - F_j(x)\right]_+}{-x_i \sum_{j \in S}\left[F_i(x) - F_j(x)\right]_+}$	Pairwise difference (PD)	Smith (1984)
$\rho_{ij} = \frac{\exp\left(\eta^{-1}F_j\right)}{\sum_{k \in S}\exp(\eta^{-1}F_k)}$	$\dot{x}_i = \frac{\exp\left(\eta^{-1}F_j(x)\right)}{\sum_{k \in S}\exp(\eta^{-1}F_k(x))} x_i$	Logit	Fudenberg and Levine (1998)
$\rho_{ij} = B_i^F(x)$	$\dot{x} = B_i^F(x) - x_i$	Best response	Gilboa and Matsui (1991)

For an overview of work in these directions, see Sandholm (2007, ch. 6).

In the context of random matching in normal form games, it is natural to ask whether an equilibrium that is stable under an evolutionary dynamic also satisfies the restrictions proposed in the equilibrium refinements literature. Swinkels (1993) and Demichelis and Ritzberger (2003) show that this is true in great generality under even the most demanding refinements: in particular, any component of rest points that is asymptotically stable under a dynamic that respects condition (PC) contains a strategically stable set in the sense of Kohlberg and Mertens (1986). While proving this result is difficult, the idea behind the result is simple. If a component is asymptotically stable under an evolutionary dynamic, then this dynamic stability ought not to be affected by slight perturbations of the payoffs of the game. *A fortiori*, the *existence* of the component ought not to be affected by the payoff perturbations either. But this preservation of existence is precisely what strategic stability demands.

This argument also shows that asymptotic stability under evolutionary dynamics is a qualitatively stronger requirement than strategic stability: while strategic stability requires equilibria not to vanish after payoff perturbations, it does not demand that they be attracting under a disequilibrium adjustment process. For example, while all Nash equilibria of simple coordination games are strategically stable, only the pure Nash equilibria are stable under evolutionary dynamics.

Demichelis and Ritzberger (2003) establish their results using tools from index theory. Given an evolutionary dynamic V^F for a game F, one can assign each component of rest points an integer, called the *index*, that is determined by the behaviour of the dynamic in a neighbourhood of the rest point; for instance, regular, stable rest points are assigned an index of 1. The set of all indices for the dynamic V^F is constrained by the *Poincaré–Hopf theorem*, which tells us that the sum of the indices of the equilibrium components of V^F must equal 1. As a consequence of this deep topological result, one can sometimes determine the local stability of one component of rest points by evaluating the local stability of the others.

Global Convergence: Positive and Negative Results

To provide the most satisfying evolutionary justification for the prediction of Nash equilibrium play, it is not enough to link the rest points of a dynamic and the Nash equilibria of the underlying game, or to prove local stability results. Rather, one must establish convergence to Nash equilibrium from *arbitrary* initial conditions.

One way to proceed is to focus on a class of games defined by some noteworthy payoff structure, and then to ask whether global convergence can be established for games in this class under certain families of evolutionary dynamics. As it turns out, general global convergence results can be proved for a number of classes of games. Among these classes are *potential games*, which include common interest games, congestion games, and games generated by externality pricing schemes; *stable games*, which include zero-sum games, games with an interior ESS, and (perturbed) concave potential games; and *supermodular games*, which include models of Bertrand oligopoly, arms races, and macroeconomic search. A fundamental paper on global convergence of evolutionary dynamics is Hofbauer (2000); for a full treatment of these results, see Sandholm (2007).

Once we move beyond specific classes of games, global convergence to Nash equilibrium cannot be guaranteed; cycling and chaotic behaviour become possible. Indeed, Hofbauer and Swinkels (1996) and Hart and Mas-Colell (2003) construct examples of games in which all reasonable deterministic evolutionary dynamics fail to converge to Nash equilibrium from most initial conditions. These results tell us that general guarantees of convergence to Nash equilibrium are impossible to obtain.

In light of this fact, we might instead consider the extent to which solution concepts simpler than Nash equilibrium are supported by evolutionary dynamics. Cressman and Schlag (1998) and

Cressman (2003) investigate whether imitative dynamics lead to subgame perfect equilibria in reduced normal forms of extensive form games – in particular, generic games of perfect information. In these games, interior solution trajectories do converge to Nash equilibrium components, and only subgame perfect components can be interior asymptotically stable. But even in very simple games interior asymptotically stable components need not exist, so the dynamic analysis may fail to select subgame perfect equilibria. For a full treatment of these issues, see Cressman (2003).

What about games with strictly dominated strategies? Early results on this question were positive: Akin (1980), Nachbar (1990), Samuelson and Zhang (1992), and Hofbauer and Weibull (1996) prove that dominated strategies are eliminated under certain classes of imitative dynamics. However, Berger and Hofbauer (2006) show that dominated strategies need not be eliminated under the BNN dynamic. Pushing this argument further, Hofbauer and Sandholm (2006) find that dominated strategies can survive under any continuous evolutionary dynamic that satisfies positive correlation and *innovation*; the latter condition requires that agents choose unused best responses with positive probability. Thus, whenever there is some probability that agents base their choices on direct evaluation of payoffs rather than imitation of successful opponents, evolutionary dynamics may violate even the mildest rationality criteria.

Conclusion

Because the literature on evolutionary dynamics came to prominence shortly after the literature on equilibrium refinements, it is tempting to view the former literature as a branch of the latter. But, while it is certainly true that evolutionary models have something to say about selection among multiple equilibria, viewing them simply as equilibrium selection devices can be misleading. As we have seen, evolutionary dynamics capture the behaviour of large numbers of myopic, imperfectly informed decision makers. Using evolutionary models to predict behaviour in interactions between, say, two well-informed players is daring at best.

The negative results described in Section 6 should be understood in this light. If we view evolutionary dynamics as an equilibrium selection device, the fact that they need not eliminate strictly dominated strategies might be viewed with disappointment. But, if we take the result at face value, it becomes far less surprising: if agents switch to strategies that perform reasonably well at the moment of choice, that a strategy is never optimal need not deter agents from choosing it.

A similar point can be made about failures of convergence to equilibrium. From a traditional point of view, persistence of disequilibrium behaviour might seem to undermine the very possibility of a satisfactory economic analysis. But the work described in this entry suggests that in large populations, this possibility is not only real but is also one that game theorists are well equipped to analyse.

See Also

- ▶ Learning and Evolution in Games: Adaptive Heuristics
- ▶ Learning and Evolution in Games: An Overview
- ▶ Learning and Evolution in Games: ESS
- ▶ Nash Equilibrium, Refinements of
- ▶ Stochastic Adaptive Dynamics

Bibliography

Akin, E. 1980. Domination or equilibrium. *Mathematical Biosciences* 50: 239–250.

Beckmann, M., C. McGuire, and C. Winsten. 1956. *Studies in the economics of transportation*. New Haven: Yale University Press.

Benaïm, M., and J. Weibull. 2003. Deterministic approximation of stochastic evolution in games. *Econometrica* 71: 873–903.

Berger, U., and J. Hofbauer. 2006. Irrational behavior in the Brown–von Neumann–Nash dynamics. *Games and Economic Behavior* 56: 1–6.

Björnerstedt, J., and J. Weibull. 1996. Nash equilibrium and evolution by imitation. In *The rational foundations of economic behavior*, ed. K. Arrow et al. New York: St Martin's Press.

Bomze, I. 1986. Non-cooperative two-person games in biology: A classification. *International Journal of Games Theory* 15: 31–57.
Brown, G., and J. von Neumann. 1950. Solutions of games by differential equations. In *Contributions to the theory of games I*, ed. H. Kuhn and A. Tucker. Annals of Mathematics Studies 24. Princeton: Princeton University Press.
Cressman, R. 2003. *Evolutionary dynamics and extensive form games*. Cambridge, MA: MIT Press.
Cressman, R., and K. Schlag. 1998. The dynamic (in) stability of backwards induction. *Journal of Economic Theory* 83: 260–285.
Dawkins, R. 1982. *The extended phenotype*. San Francisco: Freeman.
Demichelis, S., and K. Ritzberger. 2003. From evolutionary to strategic stability. *Journal of Economic Theory* 113: 51–75.
Friedman, D. 1991. Evolutionary games in economics. *Econometrica* 59: 637–666.
Fudenberg, D., and D. Levine. 1998. *Theory of learning in games*. Cambridge, MA: MIT Press.
Gilboa, I., and A. Matsui. 1991. Social stability and equilibrium. *Econometrica* 59: 859–867.
Hart, S., and A. Mas-Colell. 2003. Uncoupled dynamics do not lead to Nash equilibrium. *American Economic Review* 93: 1830–1836.
Hofbauer, J. 2000. From Nash and Brown to Maynard Smith: Equilibria, dynamics, and ESS. *Selection* 1: 81–88.
Hofbauer, J., and W. Sandholm. 2006. *Survival of dominated strategies under evolutionary dynamics*. Mimeo: University College London and University of Wisconsin.
Hofbauer, J., P. Schuster, and K. Sigmund. 1979. A note on evolutionarily stable strategies and game dynamics. *Journal of Theoretical Biology* 27: 537–548.
Hofbauer, J., and K. Sigmund. 1988. *Theory of evolution and dynamical systems*. Cambridge: Cambridge University Press.
Hofbauer, J., and K. Sigmund. 1998. *Evolutionary games and population dynamics*. Cambridge: Cambridge University Press.
Hofbauer, J., and K. Sigmund. 2003. Evolutionary game dynamics. *Bulletin of the American Mathematical Society N.S.* 40: 479–519.
Hofbauer, J., and J. Swinkels. 1996. *A universal Shapley example*. Mimeo: University of Vienna and Northwestern University.
Hofbauer, J., and J. Weibull. 1996. Evolutionary selection against dominated strategies. *Journal of Economic Theory* 71: 558–573.
Kohlberg, E., and J.-F. Mertens. 1986. On the strategic stability of equilibria. *Econometrica* 54: 1003–1038.
Maynard Smith, J. 1982. *Evolution and the theory of games*. Cambridge: Cambridge University Press.
Maynard Smith, J., and G. Price. 1973. The logic of animal conflict. *Nature* 246: 15–18.
Nachbar, J. 1990. 'Evolutionary' selection dynamics in games: Convergence and limit properties. *International Journal of Games Theory* 19: 59–89.
Samuelson, L. 1997. *Evolutionary games and equilibrium selection*. Cambridge, MA: MIT Press.
Samuelson, L., and J. Zhang. 1992. Evolutionary stability in asymmetric games. *Journal of Economic Theory* 57: 363–391.
Sandholm, W. 2001. Potential games with continuous player sets. *Journal of Economic Theory* 97: 81–108.
Sandholm, W. 2007. *Population games and evolutionary dynamics*. Cambridge, MA: MIT Press.
Schuster, P., and K. Sigmund. 1983. Replicator dynamics. *Journal of Theoretical Biology* 100: 533–538.
Smith, M. 1984. The stability of a dynamic model of traffic assignment – An application of a method of Lyapunov. *Transportation Science* 18: 245–252.
Swinkels, J. 1993. Adjustment dynamics and rational play in games. *Games and Economic Behavior* 5: 455–484.
Taylor, P., and L. Jonker. 1978. Evolutionarily stable strategies and game dynamics. *Mathematical Biosciences* 40: 145–156.
Vega-Redondo, F. 1996. *Evolution, games, and economic behavior*. Oxford: Oxford University Press.
Weibull, J. 1995. *Evolutionary game theory*. Cambridge, MA: MIT Press.
Zeeman, E. 1980. Population dynamics from game theory. In *Global theory of dynamical systems*, ed. Z. Nitecki and C. Robinson. Berlin: Springer.

Deterrence (Empirical), Economic Analyses of

Steven D. Levitt and Thomas J. Miles

Abstract

Empirical economic analyses of deterrence attempt to test the central prediction of Becker's (1968) rational-actor model of criminal behaviour: that less crime occurs when the expected penalties are greater. When economists have broken the simultaneity of crime rates and crime-control policies, they have generally concluded that policing levels and the scale of incarceration reduce crime rates. Economists have made less progress in determining whether these reductions in crime are due to deterrence or incapacitation, but the research suggests that both effects are likely present. Evidence on the deterrent effect of capital punishment and particular victim precautions is far less convincing.

Keywords

Becker, G.; Crime and the city; Crime, economic theory of; Deterrence (empirical), economic analyses of; Deterrence (theory), economics of; Granger–Sims causality; Incapacitation vs deterrence; Public enforcement of law; Punishment

JEL Classifications

K42; K14

Empirical economic analyses of deterrence seek to test the central prediction of the economic or rational-actor model of criminal behaviour that Becker (1968) pioneered. In the Beckerian model, a potential offender compares the expected costs and benefits of criminal activity, and when the expected utility of crime exceeds the expected utility loss of any punishment, the actor engages in the criminal activity. Economists have attempted to confirm or refute this model by relating geographic and temporal variation in punishment regimes, which proxy for the expected cost of offending, to aggregate crime rates, which measure the frequency of criminal activity. This approach poses two challenges. First, criminal justice policies are endogenous to crime rates, because jurisdictions often devote greater resources to crime control when the incidence of crime is higher. Second, even if the econometrician breaks the simultaneity of crime rates and crime-control policies, the estimates typically do not reveal whether deterrence or incapacitation is the operative mechanism.

Estimates of the Causal Effect of Policing Levels on Crime Rates

The criminal justice policies that have most often received empirical evaluation are the scale of policing and imprisonment, which in the economic model of crime correspond to the probability of apprehension and the magnitude of the sanction, respectively. Early studies tried to infer the causal effect of police levels on crime rates by drawing cross-sectional comparisons across cities or states, but Fisher and Nagin (1978) showed that cross-sectional estimates suffer from simultaneity bias because jurisdictions with higher crime rates respond by employing more police. In the 1990s a second wave of literature emerged.

The new studies employed more sophisticated econometric strategies to break the simultaneity problem. For example, Marvell and Moody (1996) used Granger causality to identify the impact of policing levels on crime rates. A variable 'Granger causes' another when changes in the first variable generally precede changes in the second, and thus Granger causality refers to a temporal relationship between two variables rather than actual causation (Granger 1969). Marvell and Moody (1996) applied this technique to more than 20 years of state and city data and found that police Granger-caused lower crime, or that increases in police were associated with future declines in crime.

Levitt (1997) employed a different econometric strategy: an instrumental variables or 'natural experiment' approach. He argued that mayoral and gubernatorial elections were valid instruments, because they correlate with police but do not correlate with crime, except through the other explanatory variables in the crime equation. He showed that sizable increases in the police forces in major cities were concentrated in election years, perhaps because greater police generate electoral benefits for politicians. His estimate, that a ten per cent increase in the police force produced at most a ten per cent reduction in crime rates, was comparable in magnitude to Marvell and Moody's (1996). McCrary (2002) argued that, when properly measured, electoral cycles induced insufficient variation in the size of police forces to measure the impact of crime. However, Levitt (2002) showed that an alternative instrumental variable, the number of firefighters, also produces negative and sizable estimates of the impact of police on crime. Recently, Evans and Owens (2005) demonstrated that the federal subsidies from the Clinton Crime Bill stimulated police hiring and produced similar reductions in crime rates.

Other authors used more finely disaggregated data to identify the effect of police on crime. In

data with annual observations, any increase in crime and police occurring within a calendar year appears contemporaneous rather than sequential, and the short-term causal effect of police on crime is not observed. Corman and Mocan (2000) examined the short-term effect using nearly 30 years of monthly data from New York City and applying Granger causality techniques. They found that police hiring occurs approximately six months after a jump in crime and that the increase in police leads to reductions in crime as great as Levitt's (1997) largest estimate. Di Tella and Schargrodsky (2004) examined data decomposed to the level of city blocks. When the city of Buenos Aires reallocated police to temples and mosques in response to terrorist threats against them, Di Tella and Schargodsky observed that auto thefts immediately around those buildings declined abruptly but that the reduction in crime quickly decayed with distance.

Despite the use of different estimation procedures and different data-sets, the second wave of literature on policing and crime produced quite similar estimates of the crime-reducing effect of police levels. The marginal reduction in crime associated with hiring an additional police officer in large urban environments roughly equals the marginal cost.

Estimates of the Causal Effect of Incarceration Rates on Crime Rates

Empirical analyses of the crime-reducing effect of prisons evolved in a similar manner to studies of policing. Early efforts failed to recognize or address the simultaneity problem and prematurely concluded that imprisonment has neither deterrent nor incapacitating effects (see Zimring and Hawkins 1991). In the 1990s researchers again applied more sophisticated empirical strategies that attempted to break the simultaneity problem. Marvell and Moody (1994) applied Granger causality techniques to a repeated cross-section of states and found that a ten per cent increase in the prison population produced nearly a two per cent fall in crime rates.

Levitt (1996) disentangled the simultaneity of crime and incarceration by using lawsuits challenging conditions in overcrowded prisons as instrumental variables. He showed that, when the suits produced court orders to reduce overcrowding, states typically complied by releasing prisoners who otherwise would have been incarcerated. His estimates implied that the reduction in crime from incarcerating an additional prisoner was two to three times larger than that predicted by Marvell and Moody (1994).

Although these studies indicate that imprisonment reduces crime, the relevance of their estimated parameters to social policy evaluation of present incarceration levels has already diminished. The prison population has grown so rapidly since the mid-1990s that its margin lies well outside the range in which the parameter estimates were generated. For most reasonable set of assumptions, the current scale of incarceration appears at or above the socially optimal level.

Estimates Distinguishing Deterrence from Incapacitation

Although economists' understanding of the causal relationships among policing, incarceration, and crime has improved, they have made less progress on the question of whether the declines in crime are due to deterrence or incapacitation. Determining the operative effect is crucial for evaluating the economic model of crime and for designing crime-control policy.

A few empirical economic studies attempted to assess the relative importance of deterrence and incapacitation by exploring responses to increased punishments. Kessler and Levitt (1999) studied the effect of a California referendum that provided sentence enhancements for certain serious crimes. The sentence enhancement imposed an additional incapacitating effect only upon completion of the standard prison term, and any decline in crime before that date was arguably attributable to deterrence. Kessler and Levitt found that the rate of crimes covered by the referendum fell relative to other states and that the rate of crimes not covered by the referendum did not

change. After the expiration of the standard prison terms, the rate of the affected crimes continued to fall, and this further decline indicated that the full impact of the sentence enhancements included both deterrent and incapacitating effects.

Another effort to distinguish deterrence from incapacitation proceeded from the observation that criminals do not specialize in particular types of offences, but instead are generalists who participate in potentially wide range of offences. Levitt (1998a) noted that, if deterrence is the operative mechanism, a longer sentence for a particular type of crime implies that generalist criminals should substitute to other kinds of crime. If the primary effect is instead incapacitation, then a longer sentence for a particular crime should lower the rate of alternative offences. Using arrest rate data, Levitt (1998a) found mixed evidence for deterrence.

Levitt (1998b) evaluated the responsiveness of criminal activity to the transition from the juvenile to the adult criminal justice system as another means of distinguishing deterrence from incapacitation. In states where the criminal justice system is substantially more punitive than the juvenile system, deterrence predicts that juveniles should reduce their criminal activity immediately upon reaching the age of majority (before there is time for incapacitation to become a factor). States where the adult system was especially punitive relative to the juvenile system experience sharp declines in crime at the age of majority relative to states where the transition to the adult system is most lenient, consistent with deterrence.

Other Empirical Analyses of Deterrence

Capital punishment seemingly offers a direct test of the deterrence hypothesis, because the alternative sentence for a death-eligible offender is typically life imprisonment, and any crime-reducing effect of capital punishment is therefore arguably attributable to deterrence. Ehrlich (1975, 1977a, b) produced some of the earliest and most contested claims of capital punishment's deterrent effect. Cameron (1994) reviews the large literature on the death penalty, and criticisms of Ehrlich's conclusions focus on the sensitivity of the estimates to the time period, the states, and the control variables included in the analysis. Recently, a number of studies examined the relationship between the death penalty and crime rates using repeated cross-sections of states in the period since the Supreme Court's 1976 reinstatement of capital punishment. These studies use data disaggregated at the monthly (Mocan and Gittings 2003) or county-level (Dezhbakhsh et al. 2003) and study the impact on different kinds of homicide (Shepherd 2004). All claim deterrent effects at least as large as Ehrlich's original estimates, despite their continuing sensitivity to minor specification changes. In contrast, Katz et al. (2003) used state-level panel data covering the period 1950–90 and detected no effect of the death penalty on crime rates. Unlike the literature on policing and incarceration, the use of higher frequency data and additional control variables has broadened, rather than narrowed, the range of estimated impacts of the death penalty.

Although most empirical economic analyses of deterrence evaluate the role of public law enforcement, a few studies consider the responsiveness of crime to the precautions taken by potential victims. A victim's precaution may have a general deterrent effect only if the prospective offender cannot observe it before deciding to commit the crime. Otherwise, the observation of a precaution may induce the offender to substitute to a more vulnerable victim but have no effect on the total rate of offending (see Clotfelter 1978; Shavell 1991). Ayres and Levitt (1998) analysed a particular kind of anti-theft device for automobiles as an unobservable precaution. The device contained a radio transmitter that allows police with special equipment to track the vehicle, but its lack of outward indications made it unobservable to potential offenders. Ayres and Levitt found that, when the device became available in a city, vehicle thefts fell sharply and that it did not induce car thieves to substitute to other types of crimes or to other geographic areas.

Another purported unobservable precaution that received extensive empirical analysis is

surreptitious gun possession. Lott and Mustard (1997) and Lott (1998) claimed that laws relaxing the requirements for concealed weapons permits had a general deterrent effect on crime rates, but numerous researchers challenged the Lott findings. Ayres and Donohue (1999) found that in more recent years the law correlated either positively or not at all with crime rates, and Duggan (2000) showed that crime rates in states that adopted concealed-weapons laws began to decline before the passage of the laws. Other researchers argued that additional tests of the hypothesis failed to confirm it. Ludwig (1998) found that that passage of these laws was associated with large declines in the victimization of juveniles, a group not permitted to carry concealed weapons under these laws. Kovandzic and Marvell (2003) reported no relationship between the number of concealed weapons permits issued and violent crime rates in a single state.

See Also

- ▶ Becker, Gary S. (Born 1930)
- ▶ Causality in Economics and Econometrics
- ▶ Crime and the City
- ▶ Deterrence (Theory), Economics of
- ▶ Difference-in-Difference Estimators
- ▶ Granger–Sims Causality
- ▶ Law, Public Enforcement of
- ▶ Natural Experiments and Quasi-Natural Experiments

Bibliography

Ayres, I., and J. Donohue III. 1999. Nondiscretionary concealed weapons laws: A case study of statistics, standards of proof, and public policy. *American Law and Economics Review* 1: 436–470.

Ayres, I., and S. Levitt. 1998. measuring positive externalities from unobservable victim precautions: An empirical analysis of Lojack. *Quarterly Journal of Economics* 113: 43–77.

Becker, G. 1968. Crime and punishment: An economic approach. *Journal of Political Economy* 76: 169–217.

Cameron, S. 1994. A review of econometric evidence on the effects of capital punishment. *Journal of Socio-Economics* 23: 197–214.

Clotfelter, C. 1978. Private security and the public safety. *Journal of Urban Economics* 5: 388–402.

Corman, H., and H. Mocan. 2000. A time-series analysis of crime and drug use in New York City. *American Economic Review* 90: 584–604.

Dezhbakhsh, H., P. Rubin, and J. Shepherd. 2003. Does capital punishment have a deterrent effect? New evidence from postmoratorium panel data. *American Law and Economics Review* 5: 344–376.

Di Tella, R., and E. Schargrodsky. 2004. Do police reduce crime? Estimates using the allocation of police forces after a terrorist attack. *American Economic Review* 94: 115–133.

Duggan, M. 2000. More guns, more crime. *Journal of Political Economy* 109: 1086–1114.

Ehrlich, I. 1975. The deterrent effect of capital punishment: A question of life and death. *American Economic Review* 65: 397–417.

Ehrlich, I. 1977a. The deterrent effect of capital punishment: Reply. *American Economic Review* 67: 452–458.

Ehrlich, I. 1977b. Capital punishment: Further thoughts and additional evidence. *Journal of Political Economy* 85: 741–788.

Evans, W., and E. Owens. 2005. Flypaper COPS. Mimeo. Department of Economics, University of Maryland. Online. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.417.401&rep=rep1&type=pdf. Accessed 20 Oct 2005.

Fisher, F., and D. Nagin. 1978. On the feasibility of identifying the crime function in a simultaneous equations model of crime and sanctions. In *Deterrence and incapacitation: Estimating the effects of criminal sanctions on crime rates*, ed. A. Blumstein, D. Nagin, and J. Cohen. Washington, DC: National Academy of Sciences.

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

Katz, L., S. Levitt, and E. Shustorovich. 2003. Prison conditions, capital punishment, and deterrence. *American Law and Economics Review* 5: 318–343.

Kessler, D., and S. Levitt. 1999. Using sentence enhancements to distinguish between deterrence and incapacitation. *Journal of Law and Economics* 17: 343–363.

Kovandzic, T., and T. Marvell. 2003. Right-to-carry concealed handguns and violent crime: Crime control through gun decontrol? *Criminology and Public Policy* 2: 363–396.

Levitt, S. 1996. The effect of prison population size on crime rates: Evidence from prison overcrowding litigation. *Quarterly Journal of Economics* 111: 319–325.

Levitt, S. 1997. Using electoral cycles in police hiring to estimate the effect of police on crime. *American Economic Review* 87: 270–290.

Levitt, S. 1998a. Why do increased arrest rates appear to reduce crime: Deterrence, incapacitation, or measurement error? *Economic Inquiry* 36: 353–372.

Levitt, S. 1998b. Juvenile crime and punishment. *Journal of Political Economy* 106: 1156–1185.

Levitt, S. 2002. Using electoral cycles in police hiring to estimate the effects of police on crime: Reply. *American Economic Review* 92: 1244–1250.
Lott, J. Jr. 1998. *More guns, less crime*. Chicago: University of Chicago Press.
Lott, J. Jr., and D. Mustard. 1997. Crime, deterrence, and right-to-carry concealed handguns. *Journal of Legal Studies* 26: 1–68.
Ludwig, J. 1998. Concealed-gun-carrying laws and violent crime: Evidence from state panel data. *International Review of Law and Economics* 18: 239–254.
Marvell, T., and C. Moody. 1994. Prison population growth and crime reduction. *Journal of Quantitative Criminology* 10: 109–140.
Marvell, T., and C. Moody. 1996. Specification problems, police levels, and crime rates. *Criminology* 34: 609–646.
McCrary, J. 2002. Using electoral cycles in police hiring to estimate the effect of police on crime: Comment. *American Economic Review* 92: 1236–1243.
Mocan, H., and R. Gittings. 2003. Getting off death row: Commuted sentences and the deterrent effect of capital punishment. *Journal of Law and Economics* 46: 453–478.
Shavell, S. 1991. Individual precautions to prevent theft: Private versus socially optimal behavior. *International Review of Law and Economics* 11: 123–132.
Shepherd, J. 2004. Murders of passion, execution delays, and the deterrence of capital punishment. *Journal of Legal Studies* 33: 283–322.
Zimring, F., and G. Hawkins. 1991. *The scale of imprisonment*. Chicago: University of Chicago Press.

Deterrence (Theory), Economics of

Isaac Ehrlich

Abstract

In both its classical and modern versions the economic theory of crime is predicated on 'the deterrence hypothesis' – the assumption that potential and actual offenders respond to both positive and negative incentives, and that the volume of offences in the population is influenced by law enforcement and other means of crime prevention. This article traces the evolution of the modern approach to crime from the traditional focus on the interaction between offenders and law enforcers to the development of a more comprehensive 'market model' under both partial and general equilibrium settings. Theoretical extensions also emphasize alternative criteria for optimal law enforcement.

Keywords

Beccaria, C.; Becker, G.; Bentham, J.; Bribery; Deterrence; Deterrence hypothesis; Ehrlich, I.; Expected utility hypothesis; Pecuniary vs. non-pecuniary penalties; Public choice; Public enforcement of law; Punishment; Rent seeking

JEL Classifications

K4

The persistence of criminal activity throughout human history and the challenges it poses for determining optimal law-enforcement activity have already attracted the attention of utilitarian philosophers and early economists like Beccaria, Paley and Bentham. It was not until the late 1960s, however, especially following the seminal work by Becker (1968), that economists reconnected with the subject, using the modern tools of economic theory and econometrics.

In both its utilitarian and modern versions the economic approach to crime is predicated on what the new literature calls 'the deterrence hypothesis' – the assumption that potential and actual offenders respond to incentives, and that the volume of offences in the population is therefore influenced by law enforcement and other means of crime prevention. By its common connotation, deterrence generally refers to the threat of a criminal sanction, or any other form of punishment having some moderating effect on the willingness to engage in criminal activity. To interpret this hypothesis so narrowly misses, however, the basic idea on which it is founded (Ehrlich 1979). The hypothesis relates to the role of both negative incentives (such as the prospect of apprehension, conviction and punishment) and positive incentives (such as opportunities for gainful employment in legitimate relative to illegitimate occupations) as deterrents to actual or would-be offenders. It follows that not just conventional law

enforcement matters in influencing the flow of offences but external market and household conditions as well, to the extent that these affect prospective gains and losses from illegitimate activity. For this approach to provide a useful approximation of the complicated reality of crime, it is not necessary that all those who commit specific crimes respond to incentives, nor is the degree of individual responsiveness prejudged; it is sufficient that a significant number of potential offenders so behave on the margin. By the same token, the theory does not preclude a priori any category of crime, or any class of incentives, as non-conforming. Indeed, economists have applied the deterrence hypothesis to a myriad of illegal activities, from tax evasion, drug abuse and fraud to skyjacking, robbery and murder.

The Economic Approach

In Becker's analysis the equilibrium volume of crime reflects the interaction between offenders and the law-enforcement authority, and the focus is on optimal probability, severity, and type of criminal sanction – the implicit 'prices' society imposes on criminal behaviour – in view of their impact on offenders and the relative social costs associated with their imposition. Subsequent theoretical work has focused on more complete formulations of specific components of the system and their micro foundations – primarily the supply of offences, the production of specific law-enforcement activities, and alternative criteria for achieving socially optimal law enforcement. A later evolution has aimed to expand the analytical setting within which crime is analysed to address the interaction between potential offenders (supply), consumers and potential victims (private actual or indirect 'demand'), and deterrence and prevention by public authorities (government intervention). This 'market model of crime' (Ehrlich 1981, 1996) has been further explored in recent years to include interactions between criminal activity and the general economy. For the specific articles on which the following discussion is based, see Ehrlich and Liu (2006, vols. 1 and 2).

Supply

The extent of participation in crime is generally modelled as an outcome of the allocation of time among competing legitimate and illegitimate activities by potential offenders. Since illegitimate activity carries the distinct risk of apprehension and punishment for illegitimate behaviour, individuals are assumed to act as expected-utility maximizers. This may generally lead many offenders to be multiple-job holders – being part-time offenders, or going in and out of criminal activity (see Ehrlich 1973, and the empirical documentation in Reuter et al. 1990). While the mix of pecuniary and non-pecuniary benefits varies across different crime categories, which attract persons of different earning opportunities or attitudes towards risk and moral values ('preferences'), the basic opportunities affecting choice are identified in all cases as the perceived probabilities of apprehension, conviction and punishment, the marginal penalties imposed, and the expected net return on illegal over legal activity. Net returns from crime rise with the level of community wealth, which enhances the potential loot from property crime, and fall with the probability of finding employment in the legitimate labour market and the prospective legitimate wages. Entry into criminal activity and the extent of involvement in crime are thus shown to be related inversely to deterrence variables and directly to the differential return it can provide over legitimate activity. Moreover, a one per cent increase in the probability of apprehension is shown to exert a larger deterrent effect than corresponding increases in the conditional probabilities of conviction and of any specific punishment if convicted (Ehrlich 1975). Essentially due to conflicting income and substitution effects, however, sanction severity can have more ambiguous effects on active offenders: a strong preference for risk may weaken (Becker 1968) or even reverse (Ehrlich 1973) the deterrent effect of sanctions, and the results are even less conclusive if one assumes that the length of time spent in crime, not just the moral obstacle to entering it, generates disutility (Block and Heineke 1975).

The results become less ambiguous at the *aggregate* level, however, as one allows for heterogeneity of potential offenders due to differences in legitimate employment opportunities or preferences for risk and crime: a more severe sanction can reduce the crime rate by deterring the entry of potential offenders even if it has little effect on actual ones (Ehrlich 1973). In addition to heterogeneity across individuals in personal opportunities and preferences, the literature has also addressed the role of heterogeneity in individuals' perceptions about probabilities of apprehension, as affected by learning from past experience (Sah 1991). As a result, current crime rates may react, in part, to past deterrence measures. A different type of heterogeneity that can affect variability in crime rates across different crime categories and geographical units is identified by Glaeser et al. (1996) and Glaeser and Sacerdote (1999) as stemming from the degree of social interaction: that is, the extent to which potential offenders are influenced by the behaviour of their neighbours.

Private 'Demand'

The incentives operating on offenders often originate from, and are partially controlled by, consumers and potential victims. Transactions in illicit drugs or stolen goods, for example, are patronized by consumers who generate a direct demand for the underlying offence. But even crimes that inflict pure harm on victims are affected by an indirect (negative) demand, which is derived from a positive demand for safety (Ehrlich 1981). By their choice of optimal self-protection (lowering the risk of becoming a victim) or self-insurance (reducing the potential loss if victimized) through use of locks, guards, safes, and alarms, or selective avoidance of crime-prone areas (Bartel 1975; Shavell 1991; Cullen and Levitt 1999), potential victims influence the marginal costs to offenders, and thus the implicit return on crime. And since optimal self-protection generally increases with the perceived risk of victimization (the crime rate), private protection and public enforcement will be interdependent. The interaction between the two and its impact on possible fluctuations in the equilibrium volume of offences is explored in Clotfelter (1977), and Philipson and Posner (1996).

Public Intervention

Since crime, by definition, causes a net social loss, and crime control measures are largely a public good, collective action is needed to augment individual self-protection. Public intervention typically aims to 'tax' illegal returns through the threat of punishment, or to 'regulate' offenders via incapacitation and rehabilitation programmes. All control measures are costly. Therefore, the 'optimum' volume of offences as determined by the law-enforcement authority acting as a social planner cannot be nil, but must be set at a level where the marginal cost of each measure of enforcement or prevention equals its marginal benefit.

To assess the relevant net social loss, however, one must adopt a criterion for public choice. Becker (1968) and Stigler (1970) have chosen variants of aggregate income measures as the relevant social welfare function to be maximized, requiring the minimization of the sum of social damages from offences and the social cost of law-enforcement activities. This approach can lead to powerful propositions regarding the optimal magnitudes of probability and severity of punishments for different crimes and different offenders, or, alternatively, the optimal level and mix of expenditures on police, courts and corrections. The analysis reaffirms the classical utilitarian proposition that the optimal severity of punishment should 'fit the crime', and thus be assessed essentially by its deterrent value, as the marginal social cost is higher for offences causing greater marginal social damage. Moreover, it makes a strong case for the desirability of monetary fines as a deterring sanction, since fines involve pure transfer payments between offenders and the rest of society. Different criteria for public choice, however, yield different implications regarding the optimal mix of probability

and severity of punishment, as is the case when the social welfare function is expanded to include concerns for the 'distributional consequences' of law enforcement on offenders and victims in addition to their aggregate income, in which case even fines can become socially costly. These considerations can be ascribed to aversion to risk (as in Polinsky and Shavell 1979), or to aversion towards *ex post* inequality under the law as a result of the 'lottery' nature of law enforcement, by which only offenders caught and convicted for crime pay for the damage caused by all offenders, including the luckier ones who escape apprehension and conviction (as in Ehrlich 1982). A positive analysis of enforcement must also address the behaviour of the separate agencies constituting the criminal justice system: police, courts, and prison authorities. For example, Landes's analysis (1971) of the courts, which focuses on the interplay between prosecutors and defence teams, explains why settling cases out of court may be an efficient outcome of many court proceedings.

The optimal enforcement policy arising from the income-maximizing criterion can be questioned from yet another angle: a public-choice perspective. The optimization rule invoked in the aforementioned papers assumes that enforcement is carried out by a social planner. In practice, public law enforcement can facilitate the interests of rent-seeking enforcers who are amenable to malfeasance and bribes. Optimal social policy needs to control malfeasance by properly remunerating public enforcers (Becker and Stigler 1974) or, where appropriate, setting milder penalties (Friedman 1999).

Market Equilibrium

In Ehrlich's (1981, 1996) 'market model', the equilibrium flow of offences results from the interaction between aggregate supply of offences, direct or derived demand for offences (through self-protection), and optimal public enforcement, which operates like a tax on criminal activity. The model derives the equilibrium volume of offences as well as the equilibrium net return, or premium, to offenders from illegitimate over legitimate activity as a result of the interaction between the relevant aggregate supply, 'demand', and government net taxation of crime in a competitive setting. One important application concerns a comparison of deterrence, incapacitation and rehabilitation as instruments of crime control. This is because the efficacy of deterring sanctions cannot be assessed merely by the elasticity of the aggregate supply of offences schedule, as it depends on the elasticity of the private demand schedule as well. Likewise, the efficacy of rehabilitation and incapacitation programmes cannot be inferred solely from knowledge of their impact on individual offenders (see Cook 1975). It depends crucially on the elasticities of the market supply and demand schedules, as these determine the extent to which successfully rehabilitated offenders will be replaced by others responding to the prospect of higher net returns. This market setting has also been applied by Viscusi (1986), who links observed net returns on specific crimes to underlying parameters of the model, and in works by Schelling (1967), Buchanan (1973), and Garoupa (2000), who analyse various aspects of organized crime by viewing it in a monopolistic rather than a competitive setting.

Crime and the Economy

The 'market model' has been developed largely in a static, partial-equilibrium setting in which the general economy affects the illegal sector of the economy, but not vice versa. More recently, the model has been extended to deal with the interaction between the two under static and dynamic conditions as well. For example, Ehrlich (1973) argues that income inequality, serving as a proxy for relative earning opportunities in illegal versus legal activities, induces time allocation in favour of illegal activity. A number of subsequent studies interpreted this relation to imply that the volume of offences can be lowered through subsidies to legitimate employment by workers with low legitimate

earning capacity. Using a general-equilibrium setting, Imrohoroglu et al. (2000) show, however, that, to the extent that subsidies must be paid for by raising taxes on legitimate production, such income distribution policies have an ambiguous effect on crime. The subsidy raises the opportunity cost of crime to apprehended offenders, but it also works as a disincentive to legitimate production because of an increased tax rate, which lowers the tax revenue available for crime detection.

The choice between legitimate and illegitimate activity may have not just static effects on the economy's level of output but dynamic growth effects as well if it affects productive human capital formation, which serves as an engine of productivity growth. Bureaucratic corruption is a case in point. As Ehrlich and Lui (1999) argue, this is because, whenever government intervenes in private economic activity, bureaucrats have an opportunity to engage in rent seeking by collecting explicit or implicit bribes, which rise with their bureaucratic status. The return on corruption is thus higher the greater is one's investment in becoming a bureaucrat or attaining higher bureaucratic status, which competes with investment in productive human capital. The analysis explains why corruption is a barrier to growth especially in less developed countries, and why under benevolent autocratic regimes the rate of economic growth can be as high as under democratic regimes.

Investigating and implementing alternative versions of a comprehensive model of crime based on micro foundations remains an intriguing challenge for future research.

See Also

- ▶ Bentham, Jeremy (1748–1832)
- ▶ Deterrence (Empirical), Economic Analyses of
- ▶ Econometrics
- ▶ Equality
- ▶ Expected Utility Hypothesis
- ▶ General Equilibrium
- ▶ Labour Economics
- ▶ Law, Public Enforcement of
- ▶ Microfoundations
- ▶ Rent Seeking
- ▶ Risk
- ▶ Uncertainty
- ▶ Unemployment

Bibliography

Bartel, A. 1975. An analysis of firm demand for protection against crime. *Journal of Legal Studies* 4: 433–478.

Becker, G. 1968. Crime and punishment: An economic approach. *Journal of Political Economy* 76: 169–217.

Becker, G., and G. Stigler. 1974. Law enforcement, malfeasance, and compensation of enforcers. *Journal of Legal Studies* 3: 1–18.

Becker, G., and W. Landes. 1974. *The economics of crime and punishment*. New York: Columbia University Press.

Block, M., and J. Heineke. 1975. A labor theoretic analysis of the criminal choice. *American Economic Review* 65: 314–325.

Buchanan, J. 1973. A defense of organized crime? In *The economics of crime and punishment: A conference sponsored by American Enterprise Institute for Public Policy Research*. Washington, DC: American Enterprise Institute for Public Policy Research.

Clotfelter, C. 1977. Public services, private substitutes, and the demand for protection against crime. *American Economic Review* 67: 867–877.

Cook, Philip J. 1975. The correctional carrot: Better jobs for parolees. *Policy Analysis* 1: 11–55.

Cullen, J., and S. Levitt. 1999. Crime, urban flight, and the consequences for cities. *Review of Economics and Statistics* 81: 159–169.

Ehrlich, I. 1973. Participation in illegitimate activities: Theoretical and empirical investigation. *Journal of Political Economy* 81: 521–565. Reprinted with supplements as Participation in illegitimate activities: An economic analysis. In *The economics of crime and punishment*, eds. Becker, G., and W. Landes. New York: Columbia University Press.

Ehrlich, I. 1975. The deterrent effect of capital punishment: A question of life and death. *American Economic Review* 65: 397–417.

Ehrlich, I. 1979. The economic approach to crime. In *Criminology review yearbook*, ed. S. Messingerand and E. Bittner. Beverly Hills: Sage Publications.

Ehrlich, I. 1981. On the usefulness of controlling individuals: An economic analysis of rehabilitation, incapacitation and deterrence. *American Economic Review* 71: 307–322.

Ehrlich, I. 1982. The optimum enforcement of laws and the concept of justice: A positive analysis. *International Review of Law and Economics* 2: 3–27.

Ehrlich, I. 1996. Crime, punishment, and the market for offenses. *Journal of Economic Perspectives* 10(1): 43–67.
Ehrlich, I., and G. Brower. 1987. On the issue of causality in the economic model of crime and law enforcement: Some theoretical considerations and experimental evidence. *American Economic Review, Papers and Proceedings* 77(2): 99–106.
Ehrlich, I., and Z. Liu. 2006. *The economics of crime*, International library of critical writings in economics. Cheltenham/Northampton: Edward Elgar.
Ehrlich, I., and F. Lui. 1999. Bureaucratic corruption and endogenous economic growth. *Journal of Political Economy* 107: S270–S293.
Friedman, D. 1999. Why not hang them all: The virtues of inefficient punishment. *Journal of Political Economy* 107(6, Part 2): S259–S269.
Garoupa, N. 2000. The economics of organized crime and optimal law enforcement. *Economic Inquiry* 38: 278–288.
Glaeser, E., B. Sacerdote, and J. Scheinkman. 1996. Crime and social interactions. *Quarterly Journal of Economics* 111: 507–548.
Glaeser, E., and B. Sacerdote. 1999. Why is there more crime in cities? *Journal of Political Economy* 107(6, Part2): S225–S258.
Imrohoroglu, A., A. Merlo, and P. Rupert. 2000. On the political economy of income redistribution and crime. *International Economic Review* 41: 1–25.
Karpoff, J., and J. Lott Jr. 1993. The reputational penalty firms bear from committing criminal fraud. *Journal of Law and Economics* 36: 757–802.
Landes, W. 1971. An economic analysis of the courts. *Journal of Law and Economics* 14: 61–107.
Philipson, T., and R. Posner. 1996. The economic epidemiology of crime. *Journal of Law and Economics* 39: 405–433.
Polinsky, A., and S. Shavell. 1979. The optimal trade-off between the probability and magnitude of fines. *American Economic Review* 69: 880–891.
Reuter, P., R. MacCoun, and P. Murphy. 1990. *Money from crime: A study of the economics of drug dealing in Washington, D.C.* Santa Monica: The RAND Corporation R-3894-RF.
Sah, R. 1991. Social osmosis and patterns of crime. *Journal of Political Economy* 99: 1272–1295.
Schelling, T. 1967. Economic analysis of organized crime. Appendix D. In *Task force on organized crime: The president's commission on law enforcement and administration of justice*. Washington, DC: US Government Printing Office.
Shavell, S. 1991. Individual precautions to prevent theft: Private versus socially optimal behavior. *International Review of Law and Economics* 11: 123–132.
Stigler, G. 1970. The optimum enforcement of laws. *Journal of Political Economy* 78: 526–535.
Viscusi, W. 1986. The risks and rewards of criminal activity: A comprehensive test of criminal deterrence. *Journal of Labor Economics* 4: 317–340.

Development Economics

Debraj Ray and Clive Bell

Abstract

This article surveys the current state of development economics, a subject that studies growth, inequality, poverty and institutions in the developing world. The article is organized around a view that emphasizes the role of history in creating development traps or slow progress. This 'non-convergence' viewpoint stands in contrast to a more traditional view, also discussed, based on the notion of economic convergence (across individuals, regions or countries). Some specific research areas in development economics receive closer scrutiny under this overall methodological umbrella, among them political economy, credit markets, legal issues, collective action and conflict.

Keywords

Adverse selection; Aspirations gap; Collective action; Complementarity; Convergence; Convexity; Credit; Development economics; Expectations; Human capital; Inequality; Insurance; Land rights; Limited liability; Measurement error; Micro-credit; Moral hazard; Multiple equilibria; Poverty traps; Property rights; Selection bias; Standards of living; Underdevelopment

JEL Classifications

O1

What we know as the 'developing world' is approximately the group of countries classified by the World Bank as having 'low' and 'middle' incomes. An exact description is unnecessary and not too revealing; suffice it to observe that these countries make up over five billion of the world's population, leaving out the approximately one

billion who are part of the 'high' income 'developed world'. Together, the low-and middle-income countries generate approximately 6 trillion (2001) dollars of national income, to be contrasted with the 25 trillion generated by high-income countries. An index of income that controls for purchasing power would place these latter numbers far closer together (approximately 20 trillion and 26 trillion, according to the *World Development Report,* World Bank 2003), but the per capita disparities are large and obvious, and to those encountering them for the first time, still extraordinary.

Development economics, a subject that studies the economics of the developing world, has made excellent use of economic theory, econometric methods, sociology, anthropology, political science, biology and demography, and has burgeoned into one of the liveliest areas of research in all the social sciences. My limited approach in this brief article is one of deliberate selection of a few conceptual points that I consider to be central to our thinking about the subject. The reader interested in a more comprehensive overview is advised to look elsewhere (for example, at Dasgupta 1993; Hoff et al. 1993; Ray 1998; Bardhan and Udry 1999; Mookherjee and Ray 2001; Sen 1999).

I begin with a traditional framework of development, one defined by conventional growth theory. This approach develops the hypothesis that given certain parameters, say savings or fertility rates, economies inevitably move towards a steady state. If these parameters are the same across economies, then in the long run all economies converge to one another. If in reality we see the utter lack of such convergence – which we do (see, for example, Quah 1996; Pritchett 1997) – then such an absence must be traced to a presumption that the parameters in question are *not* the same. To the extent that history plays any role at all in this view, it does so by affecting these parameters – savings, demographics, government interventionism, 'corruption' or 'culture'.

This view is problematic for reasons that I attempt to clarify below. Indeed, the bulk of this article is organized around the opposite presumption: that two societies with the same fundamentals can evolve along very different lines – going forward – depending on past expectations, aspirations or actual history.

To some extent, the distinction between evolution and parameter is a semantic one. By throwing enough state variables ('parameters') into the mix, one might argue that there is no difference at all between the two approaches. Formally, that would be correct, but then 'parameters' would have to be interpreted so broadly as to be of little explanatory value. Ahistorical convergence and historically conditioned divergence express two fundamentally different world views, and there is little that semantic jugglery can do to bring them together.

Development from the Viewpoint of Convergence

Why are some countries poor while others are rich? What explains the success stories of economic development, and how can we learn from the failures? How do we make sense of the enormous inequalities that we see, both within and across questions? These, among others, are the 'big questions' of economic development.

It is fair to say that the model of economic growth pioneered by Robert Solow (1956) has had a fundamental impact on 'big-question' development economics. An entire literature, including theory, calibration and empirical exercises, emanates from this starting point (see, for example, Lucas 1990; Mankiw et al. 1992; Barro 1991; Parente and Prescott 2000; Banerjee and Duflo 2005). Solow's path-breaking work introduced the notion of *convergence*: countries with a low endowment of capital in relation to labour will have a high rate of return to capital (by the 'law' of diminishing returns). Consequently, a given addition to the capital stock will have a larger impact on per capita income. It follows that, if we suitably control for parameters such as savings rates and population growth rates, poorer countries will tend to grow faster and hence will catch up or *converge* to the levels of well-being enjoyed by their richer counterparts. According to this view, development is largely a matter of getting some economic and

demographic parameters right and then settling down to wait.

It is true that savings and demography are not the only factors that qualify the argument. Anything that systematically affects the marginal addition to per capita income must be controlled for, including variables such as investment in 'human capital' or harder-to-quantify factors such as 'political climate' or 'corruption'. A failure to observe convergence must be traced to one or another of these 'parameters'.

Convergence relies on diminishing returns to 'capital'. If this is our assumed starting point, the share of capital in national income gives us rough estimates of the concavity of production in capital. The problem is that the resulting concavity understates observed variation in cross-country income by orders of magnitude. For instance, Parente and Prescott (2000) calibrate a basic Cobb–Douglas production function by using reasonable estimates of the share of capital income (0.25), but then huge variations in the savings rate do not change world income by much. For instance, doubling the savings rate leads to a change in steady-state income by a factor of 1.25, which is inadequate to explain an observed range of around 20:1 (in purchasing-power-parity incomes). Indeed, as Lucas (1990) observes, the discrepancy actually appears in a more primitive way, at the level of the production function. For the same simple production function to fit the data on per capita income differences, a poor country would have to have enormously higher rates of return to capital; say, 60 times higher if it is one-fifteenth as rich. This is implausible. And so begins the hunt for other factors that might explain the difference. What did we not control for, but should have?

This describes the methodological approach. The convergence benchmark must be pitted against the empirical evidence on world income distributions, savings rates, or rates of return to capital. The two will usually fail to agree. Then we look for the parametric differences that will bridge the model to the data.

'Human capital' is often used as a first port of call: might differences here account for observed cross-country variation? The easiest way to slip differences in human capital into the Solow equations is to renormalize labour. Usually, this exercise does not take us very far. Depending on whether we conduct the Lucas exercise or the Prescott–Parente variant, we would still be predicting that the rate of return to capital is far higher in India than in the United States, or that per capita income differences are only around half as much (or less) as they truly are. The rest must be attributed to that familiar black box – 'technological differences'. That slot can be filled in a variety of ways: externalities arising from human capital, incomplete diffusion of technology, excessive government intervention, within-country misallocation of resources, and so on. All these – and more – are interesting candidates, but by now we have wandered far from the original convergence model; and if that model still continues to illuminate, it is by way of occasional return to the recalibration exercise, after choosing plausible specifications for each of these potential explanations.

This model serves as a quick and ready fix on the world, and it organizes a search for possible explanations. Taken with the appropriate quantity of salt, and viewed as a first pass, such an exercise can be immensely useful. Yet playing this game too seriously reveals a particular world view. It suggests a fundamental belief that the world economy is ultimately a great leveller, and that if the levelling is not taking place we must search for that explanation in parameters that are somehow structurally rooted in a society.

While the parameters identified in these calibration exercises go hand in hand with underdevelopment, so do bad nutrition, high mortality rates, or lack of access to sanitation, safe water and housing. Yet there is no ultimate causal chain: many of these features go hand in hand with low income in self-reinforcing interplay. By the same token, corruption, culture, procreation and politics are all up for serious cross-examination: just because 'cultural factors' (for instance) seems more weighty an 'explanation', that does not permit us to assign them the status of a truly exogenous variable. In other words, the convergence predicted by technologically diminishing returns to inputs should not blind us to the possibility of non-convergent behaviour when all variables are

treated as they should be – as variables that potentially make for underdevelopment, but also as variables that are profoundly affected by the development process.

Development from the Viewpoint of Non-convergence

This leads to a different way of asking the big questions, one that is not grounded in any presumption of convergence. The starting point is that two economies with the same fundamentals can move apart along very different paths. Some of the best-known economists writing on development in the first half of the 20th century were instinctively drawn to this view: Young (1928), Nurkse (1953), Leibenstein (1957), and Myrdal (1957) among them.

Historical legacies need not be limited to a nation's inheritance of capital stock or GDP from its ancestors. Factors as diverse as the distribution of economic or political power, legal structure, traditions, group reputations, colonial heritage and specific institutional settings may serve as initial conditions – with a long reach. Even the accumulated baggage of unfulfilled aspirations or depressed expectations may echo into the future. Factors that have received special attention in the literature include historical inequalities, the nature of colonial settlement, the character of early industry and agriculture, and early political institutions.

Expectations and Development

Consider the role of expectations. Rosenstein-Rodan (1943) and Hirschman (1958) (and several others following them) argued that economic development could be thought of as a massive *coordination failure,* in which several investments do not occur simply because other complementary investments are similarly depressed in the same bootstrapped way. Thus one might conceive of two (or more) equilibria *under the very same fundamental conditions,* 'ranked' by different levels of investment.

Such 'ranked equilibria' rely on the presence of a *complementarity*, a particular form of externality in which the taking of an action by an agent increases the marginal benefit to other agents from taking a similar action. In the argument above, sector-specific investments lie at the heart of the complementarity: more investment in one sector raises the return to investment in some related sector.

Once complementarities – and their implications for equilibrium multiplicity – enter our way of thinking, they seem to pop up everywhere. Complementarities play a role in explaining how technological inefficiencies persist (David 1985; Arthur 1994), why financial depth is low (and growth volatile) in developing countries (Acemoglu and Zilibotti 1997), how investments in physical and human capital may be depressed (Romer 1986; Lucas 1988), why corruption may be self-sustaining (Kingston 2005; Emerson 2006), the growth of cities (Henderson 1988; Krugman 1991), the suddenness of currency crises (Obstfeld 1994), or the fertility transition (Munshi and Myaux 2006); I could easily go on. Even the traditional Rosenstein-Rodan view of demand complementarities has been formally resurrected (Murphy et al. 1989).

An important problem with theories of multiple equilibria is that they carry an unclear burden of history. Suppose, for instance, that an economy has been in a low-level investment trap for decades. Nothing in the theory prevents the very same economy from abruptly shooting into the high-level equilibrium today. There is a literature that studies how the past might weigh on the present when a multiple equilibria model is embedded in real time (see, for example, Adserà and Ray 1998; Frankel and Pauzner 2000). When we have a better knowledge of such models we will be able to make more sense of some classical issues, such as the debate on balanced versus unbalanced growth. Rosenstein-Rodan argued that a 'big push' – a large, balanced infusion of funds – is ideal for catapulting an economy away from a low-level equilibrium trap. Hirschman argued, in contrast, that certain 'leading sectors' should be given all the attention, the resulting imbalance in the economy provoking salubrious cycles of private investment in the complementary sectors. To my knowledge, we still lack good theories to examine such debates in a satisfactory way.

Aspirations, Mindsets and Development

The aspirations of a society are conditioned by its circumstances and history, but they also determine its future. There is scope, then, for a self-sustaining failure of aspirations and economic outcomes, just as there is for ever-progressive growth in them (Appadurai 2004; Ray 2006).

Typically, the aspirations of an individual are generated and conditioned by the experiences of others in her 'cognitive neighbourhood'. There may be several reasons for this: the use of role models, the importance of relative income, the transmission of information, or peer-determined setting of internal standards and goals. Such conditioning will affect numerous important socio-economic outcomes: the rate of savings, the decision to migrate, fertility choices, technology adoption, adherence to norms, the choice of ethnic or religious identity, the work ethic, or the strength of mutual insurance motives.

As an illustration, consider the notion of an *aspirations gap.* In a relatively narrow economic context (though there is no need to restrict oneself to this) such a gap is simply the difference between the standard of living that is aspired to and the standard of living that one already has. The former is not exogenous; it will depend on the ambient standards of living among peers or near-peers, or perhaps other communities.

The aspirations gap may be filled, or neglected, by deliberate action. Investments in education, health, or income-generating activities are obvious examples. Does history, via the creation of aspirations gaps, harden existing inequalities and generate poverty traps? Or does the existence of a gap spur individuals on ever harder to narrow the distance? As I have argued in Ray (1998, sections 3.3.2 and 7.2.4) and Ray (2006), the effect could go either way. A small gap may encourage investments, a large gap stifle it. This leads not only to history-dependence, but also a potential theory of the connections between income inequality and the rate of growth.

These remarks are related to Duflo's (2006) more general (but less structured) hypothesis that 'being poor almost certainly affects the way people think and decide'. This 'mindset effect' can manifest itself in many ways (an aspirations gap being just one of them), and can lead to poverty traps. For instance, Duflo and Udry (2004) find that certain within-family insurance opportunities seem to be inexplicably forgone. In broadly similar vein, Udry (1996) finds that men and women in the same household farm land in a way that is not Pareto-efficient (gains in efficiency are to be had by simply reallocating inputs to the women's plots). These observations suggest a theory of the poor household in which different sources of income are treated differently by members of the household, perhaps in the fear that this will affect threat points in some intra-household bargaining game. This in itself is perhaps not unusual, but the evidence suggests that poverty itself heightens the salience of such a framework.

Markets and History Dependence

I now move on to other pathways for history dependence, beginning with the central role of inequality. According to this view, historic inequalities persist (or widen) because each individual entity – dynasty, region, country – is swept along in a self-perpetuating path of occupational choice, income, consumption and accumulation. The relatively poor may be limited in their ability to invest productively, both in themselves and in their children. Such investments might include both physical projects, such as starting a business, and 'human projects', such as nutrition, health and education. Or the poor may have ideas that they cannot profitably implement, because implementation requires start-up funds that they do not have. Yet, faced with a different level of initial inequality, or jolted by a one-time redistribution, the very same economy may perform very differently. The ability to make productive investments is now distributed more widely throughout the population, and a new outcome emerges with not just lower inequality, but higher aggregate income. These are different steady states, and they could well be driven by distant histories (see, for example, Dasgupta and Ray 1986; Banerjee and Newman 1993; Galor and Zeira 1993; Ljungqvist 1993; Ray and Streufert 1993; Piketty 1997; Matsuyama 2000).

The intelligent layperson would be unimpressed by the originality of this argument.

That the past systematically preys on the present is hardly rocket science. Yet theories based on convergence *would rule out such obvious arguments.* Under convergence, the very fact that the poor have limited capital in relation to labour allows them to grow faster and (ultimately) to catch up. Economists are so used to the convergence mechanism that they sometimes do not appreciate just how unintuitive it is.

That said, it is time now to cross-examine our intelligent layperson. For instance, if all individuals have access to a well-functioning capital market, they should be able to make an efficient economic choice with no heed to their starting position, and the shadows cast by past inequalities must disappear (or at least dramatically shrink). For past wealth to alter current investments, imperfections in capital or insurance markets must play a central role.

At the same time, such imperfections are not sufficient: the concavity of investment returns would still guarantee convergence. A first response is that 'production functions' are simply not concave. A variety of investment activities have substantial fixed costs: business start-ups, nutritional or health investments, educational choices, migration decisions, crop adoptions. Indeed, it is hard to see how the presence of such non-convexities could *not* be salient for the ultra poor. Coupled with missing capital markets, it is easy to see that steady state traps, in which poverty breeds poverty, are a natural outcome (see, for example, Majumdar and Mitra 1982; Galor and Zeira 1993). Surveys of the economic conditions of the poor (Fields 1980; Banerjee and Duflo 2007) are eminently consistent with this point of view.

A related source of non-convexity arises from limited liability. A highly indebted economic agent may have little incentive to invest. Similarly, poor agents may enter into contracts with explicit or implicit lower bounds on liability. These bounds can create poverty traps (Mookherjee and Ray 2002a).

Investment activities that go past these minimal thresholds are potentially open to 'convexification'. There are various stopping points for human capital acquisition, and a household can hold financial assets which are, in the end, scaled-down claims on other businesses. According to this point of view, dynasties that make it past the ultra-poor thresholds will exhibit ergodic behaviour (as in Loury 1981; Becker and Tomes 1986) and so the prediction is roughly that of a two-class society: the ultra poor are caught in a poverty trap and the remainder enjoy the benefits of convergence. History would matter in determining the steady-state proportions of the ultra-poor.

But this sort of analysis ignores the endogenous non-convexities brought about by the price system. For instance, even if there are many different education levels, the wage payoff to each such level will generally be determined by the market. There is good reason to argue (see, for example, Ljungqvist 1993; Freeman 1996; Mookherjee and Ray 2002b, 2003) that the price system will sort individuals into different occupational choices, and that there will be persistent inequality across dynasties located at each of these occupational slots. Thus an augmented theory of history dependence might predict a particular proportion of the ultra-poor trapped by physical non-convexities (low nutrition, ill-health, debt, lack of access to primary education), as well as a persistently unequal dispersion of dynasties across different occupational choices, induced by the pecuniary externalities of relative prices.

Note that it is precisely the high-inequality, high-poverty steady states that are correlated with low average incomes for society as a whole, and it is certainly possible to build a view of underdevelopment from this basic premise. The argument can be bolstered by consideration of economy-wide externalities; for instance, in physical and human capital (Romer 1986; Lucas 1988; Azariadis and Drazen 1990).

History, Aggregates and the Interactive World

Theories such as these might yield a useful model for the interactive world economy. Take, for instance, the notion of aspirations. Just as domestic aspirations drive the dynamics of accumulation *within* countries, there is a role, too, for national aspirations that are driven by inter-country

disparities in consumption and wealth, with implications for the international distribution of income. Even the simplest growth framework that exhibits the usual features of convexity in its technology and budget constraints could give rise in the end to a bipolar world distribution. Countries in the middle of that distribution would tend to accumulate faster, be more dynamic and take more risks as they see the possibility of full catch-up within a generation or less. One might expect the greatest degree of 'country mobility' in this range. In contrast, societies that are far away from the economic frontier may see economic growth as too limited and too long-term an instrument, leading to a failure, as it were, of 'international aspirations'. Groups within these societies may well resort to other methods of potential economic gain, such as rent-seeking or conflict. (The aggregate impact of such activities would reinforce the slide.)

Of course, an entirely mechanical transplantation of the aspirations model to an international context is not a good idea. Countries are not individual units: a more complete theory must take into account the aspirations of various groups in the different countries, and the domestic and international components that drive such aspirations.

Next, consider the role of markets. Once again, tentatively view each country as a single economic agent in the framework of section "Markets and History Dependence". The non-convexities to be considered are at the level of the country as a whole – Young's increasing returns on a grand scale, or economy-wide externalities as in Lucas–Azariadis–Drazen. This reinterpretation is fairly standard, but, less obviously, the occupational choice story stands up to reinterpretation as well. To see this, note that the pattern of production and trade in the world economy will be driven by patterns of comparative advantage across countries. But in a dynamic framework, barring non-reproducible resources such as land or mineral endowments, *every* endowment is potentially accumulable, so that comparative advantage becomes endogenous. Thus we may view countries as settling into subsets of occupational slots (broadly conceived), producing an incomplete range of goods and services in relation to the world list, and engaging in trade.

For instance, suppose that country-level infrastructure can be tailored to either high-tech or low-tech production, but not both. If both high-tech and low-tech are important in world production and consumption, then *one* country has to focus on low-tech and *another* on high-tech. Initial history will constrain such choices, if for no other reason than the fact that existing infrastructure (and national wealth) determines the selection of future infrastructure. This is not to say that no country can break free of those shackles. For instance, as the whole world climbs up the income scale, natural non-homotheticities in demand will push commodity compositions increasingly in favour of high-quality goods. As this happens, more countries will be able to make the transition. But on the whole, if national infrastructure is more or less conducive to some (but not the full) range of goods, the non-convergence model that we discussed for the domestic economy must apply to the world economy as well.

This raises an obvious question: what is so specific about 'national infrastructure'? Why is it not possible for the world to ultimately rearrange itself so that every country produces the same or similar mix of goods, thus guaranteeing convergence? Do current national advantages somehow manifest themselves in future advantages as well, thus ensuring that the world economy settles into a permanent state of global inequality? Might economic underdevelopment across countries, at least in this relative sense, always stay with us?

To properly address such questions we have to drop the tentative assumption that each country can be viewed as an individual unit. In a more general setting, there are individuals within countries, and then there is cross-country interaction. The former are subject to the forces of occupational structure (and possible fixed costs), as discussed in section "Markets and History Dependence". The latter are subject to the specificities, if any, of 'national infrastructure', determining whether countries as a whole have to specialize (at least to some degree). The relative importance of within-country versus cross-country

inequalities will rest, in large part, on considerations such as these.

I have not brought in international political economy so far (though see below); yet, as frameworks go, this is not a bad one to start thinking about the effects of globalization. It is certainly preferable to a view of the world as a set of disconnected, autarkic growth models.

Institutions and History

In many developing countries, the early institutions of colonial rule were directly set up for the purposes of surplus extraction. There would be variation, of course, depending on whether the areas were sparsely or densely populated to begin with, or whether there was widespread availability of mineral deposits. Resource deposits certainly favoured large-scale extractive industry (as in parts of South America), while soil and weather conditions might encourage plantation agriculture, often with the use of slave labour (as in the Caribbean). On the other hand, a high pre-existing population density would favour extraction of a different hue: the setting up of institutional systems to acquire rents (the British colonial approach in large parts of India).

It has been argued, perhaps most eloquently by Sokoloff and Engerman (2000), that initial institutional modes of production and extraction in distant history had far-reaching effects on subsequent development. In their words, scholars 'have begun to explore the possibility that initial conditions, or factor endowments broadly conceived, could have had profound and enduring impacts on long-run paths of institutional and economic development' (2000, p. 220). The inequalities generated by such initial conditions may subsequently be inimical to development in a variety of ways (via the market-based pathways discussed earlier, for instance). In contrast, where initial settlements did not go hand in hand with systems of tribute, land grants, or large-scale extractive industries (as in several regions of North America), one might expect broad-based development to occur.

This is consistent with the market-based processes considered earlier. But a principal strand of the Sokoloff–Engerman argument, as also the lines of reasoning pursued in Robinson (1998), Acemoglu et al. (2001, 2002), and Acemoglu (2006), emphasizes political economy. In the words of Sokoloff and Engerman,

> [I]nitial conditions had lingering effects . . . because government policies and other institutions tended to reproduce them. Specifically, in those societies that began with extreme inequality, elites were better able to establish a legal framework that insured them disproportionate shares of political power, and to use that greater influence to establish rules, laws, and other government policies that advantaged members of the elite relative to nonmembers contributing to persistence over time of the high degree of inequality . . . In societies that began with greater equality or homogeneity among the population, however, efforts by elites to institutionalize an unequal distribution of political power were relatively unsuccessful . . . (Sokoloff and Engerman 2000, pp. 223–4)

The elite – erstwhile collectors of tribute, land-grant recipients, plantation owners and the like – may survive long after the initial institutions that spawned them are gone. Such survival may nevertheless be compatible with the maximization of aggregate surplus, provided that the elite are the most efficient of the economic citizenry in the generations to come. But there is absolutely no reason why this should be the case. A new generation of entrepreneurs, economic and political, may be waiting to take over in the wings. It is an open question as to what will happen next, but the elite may well engage in policy that has as its goal not economic efficiency but the crippling of political opposition. Some evidence of this reluctance to let go may be seen in literature that argues that more unequal societies redistribute less (see Perotti 1994, 1996; the survey by Bénabou 1996).

There are other routes. The elite may be unable to avoid an oppositional showdown. A theory of bad policy may then have to be replaced by a model of social unrest and conflict generated by initial inequality. While this mechanism is clearly different, the end result is the same. The channelling of resources to ongoing conflict will surely inhibit the accumulation of productive resources (Benhabib and Rustichini 1996; González 2007). There may also be effects running through legal systems (see, for example, La Porta et al. 1997, 1998) or the varying nature of different colonial systems (see,

for example, Bertocchi and Canova 2002). There may be effects running through the insecurity of property rights or fear of elite expropriation (see, for example, Binswanger et al. 1995).

We do not yet have a systematic exploration of these mechanisms, nor an accounting of their relative importance. But there is some reduced-form evidence that historical institutions affect growth in the manner described by Sokoloff and Engerman. The problem in establishing an empirical assertion of this sort is fairly obvious: good institutions and good economic outcomes may simply be correlated via variables we fail to observe or measure, or any observed causality may simply run from outcomes to institutions. Acemoglu et al. (2001) propose a novel instrument for (bad) institutions: the mortality rate among European settlers (bishops, sailors and soldiers to be exact). This is a clever idea that exploits the following theory: only areas that could be settled by the Europeans developed egalitarian, broad-based institutions. In the other areas, the same Europeans settled for slavery, dictatorship, highly unequal land grants and unbridled extraction instead. (The implied instrument is more convincing when the analysis is combined with controls for the general disease environment, which could have a direct effect on performance.)

The Acemoglu–Johnson–Robinson results, which show that early institutions have an effect on current performance, are provocative and interesting. It bears reiteration, though, that IV estimates are suggestive of an institutional impact on development, but one just cannot be sure of what the mechanism is. By relinquishing more immediate institutional effects on the grounds of, say, endogeneity, it becomes much harder to identify the structural pathways of influence. This appears to be an endemic problem with large, sweeping cross-country studies that attempt to detect an institutional effect. Good instruments are hard to find, and when they exist, their effect could be the echo of one or more of a diversity of underlying mechanisms.

Iyer (2005) and Banerjee and Iyer (2005) consider a somewhat different channel of influence. Both these papers study the differential impact of colonial rule within a single country, India. Iyer studies British annexations of parts of India, and the effect today on public goods provision across annexed and non-annexed parts. There is obvious endogeneity in the areas chosen for annexation (a similar observation applies, in passing, to countries 'selected' for colonization). Iyer instruments annexation by exploiting the so-called Doctrine of Lapse, under which the British annexed states in which a native ruler died without a biological heir. Banerjee and Iyer study the effect of variations in the land revenue systems set up by the British, starting from the latter half of the 18th century. In particular, they distinguish between landlord-based institutions, in which large landlords were used to syphon surplus to the British, and other areas based on rent payments, either directly from the cultivator or via village bodies. While these institutions of extraction no longer exist (India has no agricultural income tax), the authors argue that divided, unequal areas in the past cannot come together for collective action. Dispossessed groups are more worried about insecurity of tenure and fear of expropriation than about the absence of public goods, investment (public or private) or development expenditure.

Institutions and the Interactive World

In section "History, Aggregates and the Interactive World", we applied market-based theories of occupational choice and persistent inequality to the interactive world economy, (tentatively) treating each country as an economic agent. Recall the main assumption for such an interpretation to be sensible: that countries must face infrastructural constraints that limit full diversification. With these constraints in place, there will be persistent inequality in the world income distribution, with countries in 'occupational niches' that correspond to their infrastructural choices.

Bring to this story the role of institutional origins. Then a particular institutional history may be more suited to particular subsets of occupations, driving the country in question into a determinate slot in the world economy. From that point on, the persistent cross-country inequalities generated by the market-based theory will continue to link past institutions to subsequent growth. In short, initial

institutional differences may be correlated with subsequent performance, but the *magnitude* of that under-or over-performance is not to be entirely traced to initial history. Distant history could simply have served as a marker for some countries to supply a particular range of occupations, goods and services. Today's inequality may well be driven, not by that far-away history but simply by the world equilibrium path that follows on those initial conditions. If all goods are needed, there must be banana producers, sugar manufacturers, coffee growers, and high-tech enclaves, but there cannot be too little or too many of any of them.

The 'inefficient political power' argument used in section "Institutions and History" can also be transplanted to international interactions. It may well be that a large part of such interactions – protection of international property rights, restrictions on technology transfer, or barriers to trade – is used to deter the entry of developing countries onto a level playing field in which they can successfully compete with their compatriots in developed countries. It would certainly be naive to disregard this point of view altogether.

Looked at this way, our view of history fits in well with the entire debate on globalization. One might view one side of this debate as emphasizing the convergence attributes of globalization: outsourcing, the establishment of international production standards, technology transfer, political accountability and responsible macroeconomic policies may all be invoked as foot soldiers in the service of convergence.

On the other side of the battle lines are equally formidable opponents. A skewed playing field can only keep tipping, so goes the argument. The protection of intellectual property is just a way of maintaining or widening existing gaps in knowledge. Technology transfers are inappropriate because the input mix is not right. Non-convexities and increasing returns are endemic.

My goal here is not to take sides on this debate (though like everyone, *I* do have an opinion) but to clarify it from a 'non-convergence perspective' that has so far received more attention within the closed economy. There is a strong parallel between globalization (and those contented or discontented with it, to borrow a phrase from Joseph Stiglitz 2002) and the questions of convergence and divergence in closed economies.

Digging Deeper: The Microeconomics of Development

There is no getting away from the big questions, even if they cannot be fully answered with the knowledge and tools we have to hand. The issues we have discussed (and our intuitive first-takes on them) determine our world view, the cognitive canvas on which we arrange our overall thoughts. But only the most hard-bitten macroeconomist would feel no trepidation about taking these models literally, and applying them without hesitation across countries, regions and cultures.

The microeconomics of development enables us to dig below the macro questions, unearthing insight and structure with far more confidence than we can hope to have at the world or cross-country level. From the viewpoint of economic theory, the assumptions made can be more carefully motivated and are open to careful testing. From the viewpoint of empirical analysis, it is far easier to find instruments or natural experiments, or, for that matter, to conduct one's own experiments. There is the philosophical problem of scaling up the results, of using a well-controlled finding to predict outcomes elsewhere. In the end, the choice between the fuzzy, imprecise big picture and the small yet carefully delineated canvas is perhaps a matter of taste.

I need hardly add that my selectivity continues unabated: there is a whole host of issues, and I can but touch on a fraction of them. I focus deliberately on four important topics that are relevant to my overall theme of history dependence, and that have been the subject of much recent attention.

The Credit Market

As we have seen, a failure of the credit market to function is at the heart of market-based arguments for divergence.

The fundamental reason for imperfect or missing credit markets is that individuals cannot be counted upon (for reasons of strategy or luck) to

fully repay their loans. If borrowers do not have deep pockets, or if a well-defined system for enforcing repayment is missing, then it stands to reason that lenders would be reluctant to advance those loans in the first place. There is little point in asserting that a carefully chosen risk premium will deal with these risks: the premium itself affects the default probability. Therefore some borrowers will be shut out of the market, *no matter what rate of interest they are willing to pay.* Such a market will typically clear by rationing access to credit, and not by an adjustment of the rate of interest.

Three fundamental features characterize different theories of imperfect credit markets. There is classical adverse selection, in which borrower (or project) characteristics may systematically adjust with the terms of the loan contract on offer. Stiglitz and Weiss (1981) initiate this literature for credit markets, arguing that the higher the interest rate, the more likely it is that the borrower pool will be contaminated by riskier types. Then there is the moral hazard problem (see, for example, Aghion and Bolton 1997), in which the borrower must expend effort *ex post* to increase the chances of project success. Moral hazard also ties into 'debt overhang', in which existing indebtedness makes it less credible that a borrower will put in sustained effort in the project. Finally, there is the enforcement problem (see, for example, Eaton and Gersovitz 1981), in which a borrower may be tempted to engage in strategic default. Ghosh et al. (2001) survey some of the literature.

The poor are particularly affected, not because they are intrinsically less trustworthy, but because in the event of a project failure they will not have the deep pockets to pay up. The poor may well possess collateral – a small plot of land or their labour – but such collateral may be hard to adequately monetize: a formal sector bank may be unwilling to accept a small rural plot as collateral, much less bonded labour; but other lenders (a rural landlord, for instance) might. It is therefore not surprising to see interlinkages in credit transactions for the poor: a small farmer is likely to borrow from a trader who trades his crop, while a rural tenant is likely to borrow from his landlord. Even when the entire market looks competitive, these niches may create pockets of exploitative local monopoly (Ray and Sengupta 1989; Floro and Yotopoulos 1991; Floro and Ray 1997; Mansuri 1997; Genicot 2002).

In short, the very fact of their limited wealth puts the relatively poor under additional constraints in the credit market. This is why imperfect capital markets serve as a starting point for many of the models that study market-based history dependence.

The direct empirical evidence on the existence of credit constraints is surprisingly sparse, which is obviously not to say that they do not exist, but only to point out that this is an area for future research. Existing literature in a development context largely uses the existence of (presumably undesirable) consumption fluctuations in households to infer the lack of perfect financial markets (see Morduch 1994; Townsend 1995; Deaton 1997). A direct test for credit constraints yields positive results for Indian firms (Banerjee and Duflo 2004), though it is unclear how general this finding is (see, for example, Hurst and Lusardi 2004). There is a sizeable literature dealing with the impact of credit constraints on outcomes such as health (Foster 1995), education (Jacoby and Skoufias 1997) or the acquisition of production inputs such as bullocks (Rosenzweig and Wolpin 1993).

Chiappori and Salanie (2000) and Karlan and Zinman (2006) are two examples of specific tests for different frictions, such as adverse selection and enforcement. Udry's seminal (1994) paper on credit and insurance markets in northern Nigeria may be viewed as singling out enforcement as perhaps the most important binding constraint. The importance of enforcement constraints is, of course, not peculiar to credit or insurance; Fafchamps (2004) develops the point for a variety of markets in sub-Saharan Africa. For more on insurance, see Townsend (1993, 1995), Ligon (1998), Fafchamps (2003), and Fafchamps and Lund (2003). Coate and Ravallion (1993), Ligon et al. (2002), Kocherlakota (1996), and Genicot and Ray (2003) develop some of the associated theory with limited enforcement.

Finally, there is a literature on *micro-credit,* the lending of relatively small amounts to the very

poor; Armendáriz and Morduch (2005) is a good starting point.

Collective Action for Public Goods

There is a growing literature on the political economy of development. Unlike some mainstream approaches in political science and political economy, this literature appears to largely eschew voting models. In my view this is not a bad thing. Perhaps the most important criticism of voting models is that even in vigorous democracies, most policies are not subject to referenda among the citizenry at large. Certainly, there are periodic elections, and the sum total of enacted policies – and the package of future promises – are then up for voter scrutiny, but, nevertheless, there is a large and significant gap between voting and the enactment of a *particular* policy. Between that policy and the voter falls the shadow of collective action, lobbies, capture and influence, cynical trade-offs across special interests, and covert or open conflict. For countries with a non-democratic history, these considerations are expanded by orders of magnitude.

An important literature concerns the determinants of collective action for the provision of public goods, and how poverty or inequality affects the ability to engage in such action. The relationship here is complex. There are two potential reasons why inequality in a community may enhance collective action. First, the elite in a high-inequality community might largely internalize all the benefits from the resulting public good, and therefore pay for it (Olson 1965). Good examples involve military alliances (Sandler and Forbes 1980), technology adoption (Foster and Rosenzweig 1995) or even 'top-down interventions' by local rulers or elites (Banerjee et al. 2007). Second, the elite has a low opportunity cost of money, while the poor have a low opportunity cost of labour; in some situations, the two resources can be usefully combined for collective action (an alliance for violent conflict, as in Esteban and Ray 2007a, is a good example). But there are many situations in which inequality can dampen effective collective action: when all agents supply similar inputs – say effort – but their impact or cost of provision is nonlinear (Khwaja 2004; Ray et al. 2007), when there are unequally distributed private endowments (Baland and Platteau 1998; Bardhan et al. 2006), when different individuals in the same community want different things by virtue of their social differences or inequality (Alesina et al. 1999; Banerjee et al. 2001; Miguel and Gugerty 2005; Alesina and La Ferrara 2005), or when inequalities in wealth erode the informational basis of collective action (Esteban and Ray 2006).

The importance of this area of research cannot be overemphasized. Several of the fundamental accompaniments of development require state intervention at a basic level: health, education, social safety nets and infrastructure. This is especially so in poor countries, where privatized health and education are often ruled out by the sheer force of economic necessity. Yet states often are set upon by numerous claims that compete for their attention. How are these claims resolved? The theory and practice of collective action demands more research.

Moreover, while it can be argued (as above) that inequality within a community might go either way in affecting that community's ability to obtain public goods, there is no escaping the fact that at the level of the entire society, high inequality serves to fracture and divide. Simply put, the very rich want state policy that is different from what the very poor desire, and rare is the society that has them in the same camp, and demanding the same things of their government. In the world of the median voter, one might simply resolve these issues by looking at the median voter's ideal policy, but even in this rarefied scenario there are complex issues that deserve our consideration. Political alliances can often redefine the median voter (Levy 2004) and even without alliances it is unclear just who the median voter is (Bénabou 2000). When we return to the 'real world' of collective action, these issues are magnified considerably. In that world, each citizen does not have an endowment of one vote. The real endowments are labour and money. How these commodities combine (or compete) is fundamental to our understanding of political economy and – via this channel – our views on persistent history-dependence.

Conflict

A more sinister expression of collective action is conflict. In the second half of the 20th century and well into the first decade of the 21st the loss of human life from conflicts in developing countries was immense; the costs are beyond measurement. Even the narrow economic costs of conflict can be extremely large (Hess 2003).

That conflict contributes to economic regress is not surprising. But given our focus on history dependence, it is of equal interest to consider the causal chain running from underdevelopment to conflict. That chain has a natural and simple foundation: poverty reduces the opportunity cost of engaging in conflict. The grabbing of resources, often in an organized way, is often a far more lucrative alternative to the steady process of wealth accumulation. It is certainly a quicker alternative. (One might argue that there is less to gain as well, but this effect is attenuated in unequal societies.)

This unfortunate observation has substantial empirical support. For instance, Miguel et al. (2004) use rainfall as an instrument for economic growth in 41 African countries and derive a striking negative effect of growth on civil conflict: a negative growth shock of five percentage points raises the likelihood of civil conflict by 50 per cent; see also Dube and Vargas (2006) and Hidalgo et al. (2007), both of which also instrument for economic shocks to find significant effects on conflictual outcomes. Collier and Hoeffler (1998), Sambanis (2001), Fearon and Laitin (2003), and Do and Iyer (2006) all establish strong correlations between economic adversity and conflict, the last of these countries establishing this over regions in a single country (Nepal).

Yet conflict is demonstrably wasteful, and if warring parties could sit down at the negotiating table, why would societies engage in it? This is a classical question to which there are a number of possible answers. First, there may be a Prisoner's Dilemma-like quality to conflictual incidents, in the sense that one party can precipitate attacks while the other remains passive (Leventoglu and Slantchev 2005). Second, while conflict generates waste, there is no reason to believe that *every* group is thereby made worse off by it. It is entirely possible that a group prefers conflict to a peaceful outcome: the former involves a smaller pie, but the group may obtain a larger share of it (Esteban and Ray 2001). Third, while one should be able to find a system of taxes and transfers that Pareto-dominate the conflict outcome, for various reasons – lack of commitment, a sparse informational base for the levying of taxes, dynamics with rapid power shifts – it may not be possible to implement that system (Fearon 1995; Powell 2004, 2006). Fourth, it is certainly possible that conflict is over indivisible resources such as political power or religious hegemony. It may then be absurd to imagine that side A compensates side B with suitable transfers in exchange for political power: the lack of credibility involved is only too apparent. Finally, conflict may be endemic because both parties to it have incomplete information regarding chances of success, though this view has come under increasing criticism from political scientists (see, for example, Fearon 1995).

The next question of relevance concerns ethnic and social divisions. Might the presence of potentially divisive markers (caste, religion, geography, ethnicity in general) exacerbate conflictual situations? For instance, Esteban and Ray (2007a) argue that non-economic ('ethnic') markers may play a salient role in the outbreak of conflict even when society exhibits high economic inequality and may look prima facie more ripe for a class war.

A standard tool for measuring ethnic and social divisions is that of *fractionalization,* roughly defined as the probability that two individuals drawn at random will come from two distinct groups. While fractionalization seems to have a negative effect on economic outcomes such as per capita GDP (Alesina et al. 2003), growth (Easterly and Levine 1997), or governance (Mauro 1995), its effect on civil conflict appears to be insignificant (Collier and Hoeffler 2004; Fearon and Laitin 2003). Of course, as Horowitz (2000) and others have observed, it is the presence of large cleavages that is potentially conflictual, whereas fractionalization continues to increase with diversity. The solution is to drop fractionalization

altogether. Montalvo and Reynal-Querol (2005) adapt Esteban and Ray's (1994) measure of polarization to show that measures of ethnic and religious polarization *do* indeed have a significant impact on conflict (see also Do and Iyer 2006). Obviously, more research is called for on questions such as these. For instance, it is unclear how polarization should enter an empirical specification: Esteban and Ray (2007b) argue that highly polarized societies may actually avoid a showdown through deterrence, though *conditional* on the outbreak of conflict, polarization must vary positively with the intensity of conflict.

The continuing study of conflict in development demands our highest priority. Certainly, the social waste of conflict dominates the inefficiency of misallocated resources that so many mainstream economists prefer to emphasize. Indeed, it is entirely possible that the much-maligned (and much-studied) inefficiencies of incomplete information are also of a lower order of magnitude. But, most of all, it is the chain of cumulative causation that must ultimately drive our interest, from underdevelopment to conflict and back again to continuing underdevelopment. Conflict is one channel through which history matters.

Legal Matters

Contract enforcement, property rights, and expropriation risks: these are a few instances of legal matters that are central to development. They bear closely on that much-used catchall phrase, 'institutional effects on development'. For instance, Acemoglu et al. (2001) as well as the recent survey by Pande and Udry (2007) clearly have the security of property rights high on the list when discussing 'institutions'. La Porta et al. (1997, 1998, 2002) and Djankov et al. (2003) begin with the premise that common (English commercial) law and civil (French commercial) law afford different degrees of protection and support to investors, creditors and litigants, and argue that it has had dramatic effects on a variety of indicators across countries: corruption, stock-market participation, corporate valuation, government interventionism, judicial efficiency – and presumably, via these, to economic indicators.

It is little surprise that the security of property rights is generally conducive to investment, and that long-term investment is especially encouraged by such security (see, for example, Demsetz 1967). Short-term efforts, in contrast, may well be enhanced by insecurity of tenure. Depending on the exact form that property rights assume, there may be further positive effects – for example, via access to credit – that arise from the ability to mortgage or sell property (Feder et al. 1988).

Empirical research into these matters is invariably assailed by questions of endogeneity and omitted variables. For instance, long-gestation investments may provoke – and permit – the establishment of property rights, and high-ability agents might use their ability to both invest and secure their rights. Nevertheless, the evidence on property rights is that by and large they are good for investment and production (Besley 1995; Banerjee et al. 2002; Do and Iyer 2003; Goldstein and Udry 2005), and even more obviously, property values where these are reasonably well-defined (Alston et al. 1996; Lanjouw and Levy 2002). Instances in which property titling creates better access to credit are, intriguingly enough, somewhat harder to come by Field and Torero (2006) and Dower and Potamites (2006), are two of the rarer examples that do document better access, but with some qualifications).

Indeed, economists have little trouble in finding numerous instances of changed (or changing) property rights regimes. This is because there is a plethora of situations in which the absence of well-defined rights is the rule rather than the exception. In rural societies the world over, land rights can be highly ambiguous, and land titles can be missing even when an unambiguous definition of property exists. If one adds to this the sizable proportion of land under tenancy, the effective security for cultivators becomes more tenuous still (and indeed this complicates matters, because their rights may be inversely related to those of the owner!). In non-rural settings, there are substantial uncertainties for those who operate in the informal sector (such as the periodic 'cleansing' of informal retailers from city pavements). If the above studies are to be taken seriously, there are

substantial production losses from such states of insecurity.

If imperfections of the law are so inimical to the fortunes of cultivators and producers (and especially for the small and the poor among them), why do we see such institutional 'failures' in equilibrium? The Coase–Posner view would presumably have none of this: in their view, legal systems would invariably develop to maximize social surplus. But of course, there could be several reasons for the persistence of 'inefficient institutions'. When side payments are not feasible or credible, economic agents often prefer a larger share of a reduced pie to a smaller share of a more efficient pie. For instance, domestic businesses that can rely on a trusted network of kin or extended family might prefer an ambiguous legal system, which prevents entry. Or workers might prefer imperfect enforceability of a work norm, so that efficiency wages need to be paid. Borrowers might prefer that loan repayment cannot be fully enforced, so that incentives to repay must be built into the loan contract. And when tenancy is widespread in agriculture, the very design of overall property rights to maximize efficiency can be a highly complex problem.

The last three examples possess another feature that is worth some emphasis: ambiguous property rights often have equity effects that do not go the same way that efficiency-minded economists would like them to go (see Weitzman 1974; Cohen and Weitzman 1975; Baland and Platteau 1996). The ambiguity of property rights can serve as insurance, buffer, or redistributive device. As examples, consider broad access to water resources or grazing land, or the efficiency-wage premia that may need to be paid to workers or borrowers.

Most importantly, the ambiguity of property rights slows down the emergence of an overt assetless class, and that has its own social value (it should not be forgotten that the flip side of unambiguous rights is exclusion). For example, Goldstein and Udry (2005) develop this point of view in the context of rural Ghana, arguing that the ambiguity in property rights prevented the outbreak of extreme poverty (and had an interesting efficiency effect in the bargain, as individuals were reluctant to leave the land fallow – an important investment – in the fear that this would signal a lack of need for land).

The political economy of rights is a messy business, but of central importance in development economics. Poverty in general enhances the social and political need for ambiguity, while to the extent that such ambiguity wears on efficiency, we have an extremely important instance of non-convergence. Sometimes such non-convergence assumes particularly dramatic form. In West Bengal (India) 'Operation Barga' provided widespread – and welcome – use rights to registered sharecroppers (see, for example, Banerjee et al. 2001). Those very use rights now lie at the heart of recent difficulties in converting agricultural land in India for use in industry. In the world of the second best, few policies have unambiguously one-directional effects.

A Concluding Note: Theory and Empirics

While I have tried to provide a conceptual overview in this article, recent research in development economics has been almost entirely empirical. A veritable explosion in computing power, the expansion of institutional data-sets and their increased availability in electronic form, and the growing ease of collecting one's own data have bred a new generation of development economists. Their empirical sensibilities are of a high order; they are extremely sensitive to issues of endogeneity, omitted variables, measurement error and biases induced by selection. They are constantly on the search for good instruments or natural experiments, and, when these are hard to find, they are adept at creating experiments of their own.

There is little doubt that we know little enough about the world we live in that it is often worth finding out the simple things, rather than continuing to engage in what some would term flights of theoretical fantasy. Are people really credit-rationed? Does rising income automatically make for better nutrition and health? If we had

the option to throw in more textbooks, or reduce class size, or add more teachers, or install monitoring devices to track teacher attendance, which policy should we implement? Do women leaders behave differently from men in the policies that they adopt? Do households behave as one frictionless unit? Or, if one is the big-picture sort, have countries indeed converged over the last 200, or 500, years? Are richer countries more democratic? How many excess female deaths have occurred in China or India because of gender bias? Are poorer countries more 'corrupt'? And so on. The list is practically endless.

The somewhat churlish theoretically minded economist might ask, why are well-trained statisticians unable to answer these questions? Why do we need economists, who are supposed, at the very least, to combine two observations to form a deduction? The answer, at one level, is very simple and not overly supportive of the churlish theorist's complaint. While the questions are straightforward, the answers are often extremely difficult to tease out from the data, and one needs a well-trained *economist,* not a statistician, to understand the difficulty and eliminate it. Because of the aforementioned econometric issues, not a single one of the questions asked above admits a straightforward answer. Development economists spend a lot of time thinking of inventive ways to get around these problems, and it is no small feat of creativity, dedication and extremely hard work to pull off a convincing solution.

It is true that the very desire to obtain a clean, unarguable answer – with its attendant desire to have control over the empirical environment – sometimes narrows the scope of the enquiry. There is often great reluctance to rely on theoretical structure (for such reliance would contaminate the near-lexicographic desire for an unambiguous result). This means that the question to be asked is often akin to that for a simple production function (for example, 'do students do better in exams if they are given more textbooks?') or is focused on the direct effect of some policy intervention ('does the provision of health check-ups improve health outcomes?'). So it is that a boring but well-identified empirical question will often be treated with a great deal more veneration (especially if a clever instrument or randomization device is involved) than a model that relies on intuitive but undocumented assumptions.

That said, it is also a fact that we know very little about the answers to some of the most basic questions, such as the ones we have listed above. The great contribution of empirical development microeconomics is that we are building up this knowledge, piece by piece. Whether the search for that knowledge is informed by theory or not, there will be enough theorists to attempt to put these observations together. There will be enough empirical researchers to keep generating the hard knowledge. Development economics is alive and well.

See Also

- ▶ Agriculture and Economic Development
- ▶ Dual Economies
- ▶ Emerging Markets
- ▶ Endogenous Growth Theory
- ▶ Growth and Institutions
- ▶ Poverty Traps

Bibliography

Acemoglu, D., and F. Zilibotti. 1997. Was Prometheus unbound by chance? Risk, diversification and growth. *Journal of Political Economy* 105: 709–751.

Acemoglu, D., S. Johnson, and J. Robinson. 2001. The colonial origins of comparative development: An empirical investigation. *American Economic Review* 91: 1369–1401.

Acemoglu, D., S. Johnson, and J. Robinson. 2002. Reversal of fortune: Geography and institutions in the making of the modern world income distribution. *Quarterly Journal of Economics* 118: 1231–1294.

Adserà, A., and D. Ray. 1998. History and coordination failure. *Journal of Economic Growth* 3: 267–276.

Aghion, P., and P. Bolton. 1997. A theory of trickle-down growth and development. *Review of Economic Studies* 64: 151–172.

Alesina, A., and E. La Ferrara. 2005. Ethnic diversity and economic performance. *Journal of Economic Literature* 43: 762–800.

Alesina, A., R. Baqir, and W. Easterly. 1999. Public goods and ethnic divisions. *Quarterly Journal of Economics* 114: 1243–1284.

Alesina, A., A. Devleeschauwer, W. Easterly, S. Kurlat, and R. Wacziarg. 2003. Fractionalization. *Journal of Economic Growth* 8: 155–194.

Alston, L., G. Libecap, and R. Schneider. 1996. The determinants and impact of property rights: Land titles on the Brazilian frontier. *Journal of Law, Economics, and Organization* 12: 25–61.

Appadurai, A. 2004. The capacity to aspire. In *Culture and public action*, ed. V. Rao and M. Walton. Stanford: Stanford University Press.

Armendáriz, B., and J. Morduch. 2005. *The economics of microfinance*. Cambridge, MA: MIT Press.

Arthur, W. 1994. *Increasing returns and path-dependence in the economy.* Ann Arbor: University of Michigan Press.

Azariadis, C., and A. Drazen. 1990. Threshold externalities in economic development. *Quarterly Journal of Economics* 105: 501–526.

Baland, J.-M., and J.-Ph. Platteau. 1996. *Halting the degradation of natural resources: Is there a role for rural communities?* Oxford: Clarendon Press.

Baland, J.-M., and J.-Ph. Platteau. 1998. Wealth inequality and efficiency on the commons. Part II: The regulated case. *Oxford Economic Papers* 50: 1–22.

Banerjee, A., and E. Duflo. 2005. Growth theory through the lens of development economics. In *Handbook of economic growth*, ed. P. Aghion and S. Durlauf. Amsterdam: North-Holland.

Banerjee, A., and E. Duflo. 2007. The economic lives of the poor. *Journal of Economic Perspectives* 21 (1): 141–167.

Banerjee, A., and L. Iyer. 2005. History, institutions and economic performance: The legacy of colonial land tenure systems in India. *American Economic Review* 95: 1190–1213.

Banerjee, A., and A. Newman. 1993. Occupational choice and the process of development. *Journal of Political Economy* 101: 274–298.

Banerjee, A., D. Mookherjee, K. Munshi, and D. Ray. 2001. Inequality, control rights and efficiency: A study of sugar cooperatives in Western Maharashtra. *Journal of Political Economy* 109: 138–190.

Banerjee, A., P. Gertler, and M. Ghatak. 2002. Empowerment and efficiency: Tenancy reform in West Bengal. *Journal of Political Economy* 110: 239–280.

Banerjee, A., Iyer, L., and Somanathan, R. 2007. Public action for public goods. Working Paper No. 12911. Cambridge, MA: NBER.

Bardhan, P., and C. Udry. 1999. *Development microeconomics*. New York: Oxford University Press.

Bardhan, P., Ghatak, M., and Karaivanov, A. 2006. *Wealth inequality and collective action*. Mimeo, London School of Economics.

Barro, R. 1991. Economic growth in a cross-section of countries. *Quarterly Journal of Economics* 106: 407–444.

Becker, G., and N. Tomes. 1986. Human capital and the rise and fall of families. *Journal of Labor Economics* 4: S1–39.

Bénabou, R. 1996. Inequality and growth. In *NBER macroeconomics annual*, ed. B. Bernanke and J. Rotemberg. Cambridge, MA: MIT Press.

Bénabou, R. 2000. Unequal societies: Income distribution and the social contract. *American Economic Review* 90: 96–129.

Benhabib, J., and A. Rustichini. 1996. Social conflict and growth. *Journal of Economic Growth* 1: 125–142.

Bertocchi, G., and F. Canova. 2002. Did colonization matter for growth? An empirical exploration into the historical causes of Africa's underdevelopment. *European Economic Review* 46: 1851–1871.

Besley, T. 1995. Property rights and investment incentives: Theory and evidence from Ghana. *Journal of Political Economy* 103: 903–937.

Binswanger, H., K. Deininger, and G. Feder. 1995. Power, distortions, revolt and reform in agricultural land relations. In *Handbook of development economics*, ed. J. Behrman and T. Srinivasan, vol. 3B. Amsterdam: North-Holland.

Chiappori, P.-A., and B. Salanie. 2000. Testing for asymmetric information in insurance markets. *Journal of Political Economy* 108: 56–78.

Coate, S., and M. Ravallion. 1993. Reciprocity without commitment: Characterization and performance of informal insurance arrangements. *Journal of Development Economics* 40: 1–24.

Cohen, J., and M. Weitzman. 1975. A Marxian view of enclosures. *Journal of Development Economics* 1: 287–336.

Collier, P., and A. Hoeffler. 1998. On economic causes of civil war. *Oxford Economic Papers* 50: 563–573.

Collier, P., and A. Hoeffler. 2004. Greed and grievance in civil war. *Oxford Economic Papers* 56: 563–595.

Dasgupta, P. 1993. *An inquiry into well-being and destitution*. Oxford: Clarendon Press.

Dasgupta, P., and D. Ray. 1986. Inequality as a determinant of malnutrition and unemployment: theory. *Economic Journal* 96: 1011–1034.

David, P. 1985. Clio and the economics of QWERTY. *American Economic Review* 75: 332–337.

Deaton, A. 1997. *The analysis of household surveys: A microeconometric approach to development policy.* Baltimore: Johns Hopkins Press, for the World Bank.

Demsetz, H. 1967. Towards a theory of property rights. *American Economic Review* 57: 347–359.

Djankov, S., R. La Porta, F. Lopez-de-Silanes, and A. Shleifer. 2003. Courts. *Quarterly Journal of Economics* 118: 453–517.

Do, Q.-T. and Iyer, L. 2006. *Poverty, social divisions and conflict in Nepal*. Mimeo, Harvard Business School.

Dorfman, R., P. Samuelson, and R. Solow. 1958. *Linear programming and economic analysis*. Tokyo: McGraw-Hill Kogashuka.

Dower, P., and E. Potamites. 2006. *Signaling creditworthiness: Land titles, banking practices and access*

D

to formal credit in Indonesia. Mimeo, Department of Economics, New York University.

Dube, O., and J. Vargas. 2006. Are all resources cursed? Coffee, oil and armed conflict in Colombia. Documentos de CERAC 002748. Bogota, Colombia.

Duflo, E. 2006. Poor but rational? In *Understanding poverty*, ed. A. Banerjee, R. Bénabou, and D. Mookherjee. New York: Oxford University Press.

Duflo, E., and C. Udry. 2004. Intrahousehold resource allocation in Cote d'Ivoire: Social norms, separate accounts and consumption choices. Working Paper No. 10498. Cambridge, MA: NBER.

Easterly, W., and R. Levine. 1997. Africa's growth tragedy: Policies and ethnic divisions. *Quarterly Journal of Economics* 111: 1203–1250.

Eaton, J., and M. Gersovitz. 1981. Debt with potential repudiation: Theoretical and empirical analysis. *Review of Economic Studies* 48: 289–309.

Emerson, P. 2006. Corruption, competition and democracy. *Journal of Development Economics* 81: 193–212.

Esteban, J., and D. Ray. 1994. On the measurement of polarization. *Econometrica* 62: 819–851.

Esteban, J., and D. Ray. 2001. Social rules are not immune to conflict. *Economics of Governance* 2: 59–67.

Esteban, J., and D. Ray. 2006. Inequality, lobbying and resource allocation. *American Economic Review* 96: 257–279.

Esteban, J., and D. Ray. 2007a. *On the salience of ethnic conflict*. Mimeo, Department of Economics, New York University.

Esteban, J., and D. Ray. 2007b. Polarization, fractionalization and conflict. *Journal of Peace Research*, forthcoming.

Fafchamps, M. 2003. *Rural poverty, risk, and development*. Cheltenham: Edward Elgar.

Fafchamps, M. 2004. *Market institutions and Sub-Saharan Africa: Theory and evidence*. Cambridge, MA: MIT Press.

Fafchamps, M., and S. Lund. 2003. Risk sharing networks in rural Philippines. *Journal of Development Economics* 71: 261–287.

Fearon, J. 1995. Rationalist explanations for war. *International Organization* 49: 379–414.

Fearon, J., and D. Laitin. 2003. Ethnicity, insurgency, and civil war. *American Political Science Review* 97: 75–90.

Feder, G., T. Onchan, Y. Chalamwong, and C. Hongladarom. 1988. *Land policies and farm productivity in Thailand*. Baltimore: Johns Hopkins University Press.

Field, E., and M. Torero. 2006. *Do property titles increase credit access among the urban poor? Evidence from a Nationwide Titling Program*. Mimeo, Department of Economics, Harvard University.

Fields, G. 1980. *Poverty, inequality and development*. London: Cambridge University Press.

Floro, M., and D. Ray. 1997. Vertical links between formal and informal financial institutions. *Review of Development Economics* 1: 34–56.

Floro, M., and P. Yotopoulos. 1991. *Informal credit markets and the new institutional economics: The case of Philippine agriculture*. Boulder: Westview.

Foster, A. 1995. Prices, credit constraints, and child growth in low-income rural areas. *Economic Journal* 105: 551–570.

Foster, A., and M. Rosenzweig. 1995. Learning by doing and learning from others: Human capital and technical change in agriculture. *Journal of Political Economy* 103: 1176–1209.

Frankel, D., and A. Pauzner. 2000. Resolving indeterminacy in dynamic settings: The role of shocks. *Quarterly Journal of Economics* 115: 283–304.

Freeman, S. 1996. Equilibrium income inequality among identical agents. *Journal of Political Economy* 104: 1047–1064.

Galor, O., and J. Zeira. 1993. Income distribution and macroeconomics. *Review of Economic Studies* 60: 35–52.

Genicot, G. 2002. Bonded labor and serfdom: A paradox of voluntary choice. *Journal of Development Economics* 67: 101–127.

Genicot, G., and D. Ray. 2003. Group formation in risk-sharing arrangements. *Review of Economic Studies* 70: 87–113.

Ghosh, P., D. Mookherjee, and D. Ray. 2001. Credit rationing in developing countries: An overview of the theory. In *Readings in the theory of economic development*, ed. D. Mookherjee and D. Ray. London: Basil Blackwell.

Goldstein, M., and C. Udry. 2005. *The profits of power: Land rights and agricultural investment in Ghana*. Mimeo, Department of Economics, Yale University.

González, F. 2007. Effective property rights, conflict and growth. *Journal of Economic Theory*, forthcoming.

Henderson, J. 1988. *Urban development: Theory, fact, and illusion*. Oxford: Oxford University Press.

Hess, G. 2003. The economic welfare cost of conflict: An empirical assessment. Working Paper Series No. 852, CESifo, Munich.

Hidalgo, F., S. Naidu, S. Nichter, and N. Richardson. 2007. *Occupational choices: Economic determinants of land invasions*. Berkeley: Mimeo, Department of Political Science, University of California.

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

Hoff, K., A. Braverman, and J. Stiglitz. 1993. *The economics of rural organization: Theory, practice and policy.* London: Oxford University Press.

Hurst, E., and A. Lusardi. 2004. Liquidity constraints, household wealth, and entrepreneurship. *Journal of Political Economy* 112: 319–347.

Iyer, L. 2005. The long-term impact of colonial rule: Evidence from India. Working Paper No. 05–041, Harvard Business School.

Jacoby, H., and E. Skoufias. 1997. Risk, financial markets, and human capital in a developing country. *Review of Economic Studies* 64: 311–335.

Karlan, D., and J. Zinman. 2006. *Observing unobservables: Identifying information asymmetries with a consumer credit field experiment*. Mimeo, Department of Economics, Yale University.

Khwaja, A. 2004. Is increasing community participation always a good thing? *Journal of the European Economic Association* 2: 427–436.

Kingston, C. 2005. *Social structure and cultures of corruption*. Mimeo, Department of Economics, Amherst College.

Kocherlakota, N. 1996. Implications of efficient risk sharing without commitment. *Review of Economic Studies* 63: 595–609.

Krugman, P. 1991. *Geography and trade*. Cambridge, MA: MIT Press.

La Porta, R., F. Lopez-de-Silanes, A. Shleifer, and R. Vishny. 1997. Legal determinants of external finance. *Journal of Finance* 52: 1131–1150.

La Porta, R., F. Lopez-de-Silanes, A. Shleifer, and R. Vishny. 1998. Law and finance. *Journal of Political Economy* 106: 1113–1155.

La Porta, R., F. Lopez-de-Silanes, A. Shleifer, and R. Vishny. 2002. Investor protection and corporate valuation. *Journal of Finance* 57: 1147–1170.

Lanjouw, J., and P. Levy. 2002. Untitled: A study of formal and informal property rights in urban Ecuador. *Economic Journal* 112: 986–1019.

Leibenstein, H. 1957. *Economic backwardness and economic growth*. New York: Wiley.

Leventoglu, B., and B. Slantchev. 2005. *The armed peace: A punctuated equilibrium theory of war.* San Diego: Mimeo, Department of Political Science, University of California.

Levy, G. 2004. A model of political parties. *Journal of Economic Theory* 115: 250–277.

Ligon, E. 1998. Risk-sharing and information in village economies. *Review of Economic Studies* 65: 847–864.

Ligon, E., J. Thomas, and T. Worrall. 2002. Mutual insurance and limited commitment: Theory and evidence in village economies. *Review of Economic Studies* 69: 209–244.

Ljungqvist, L. 1993. Economic underdevelopment: The case of missing market for human capital. *Journal of Development Economics* 40: 219–239.

Loury, G. 1981. Intergenerational transfers and the distribution of earnings. *Econometrica* 49: 843–867.

Lucas, R. 1988. On the mechanics of economic development. *Journal of Monetary Economics* 22: 3–42.

Lucas, R. 1990. Why doesn't capital flow from rich to poor countries? *American Economic Review* 80: 92–96.

Majumdar, M., and T. Mitra. 1982. Intertemporal allocation with a non-convex technology: The aggregative framework. *Journal of Economic Theory* 27: 101–136.

Mankiw, N., D. Romer, and D. Weil. 1992. A contribution to the empirics of economic growth. *Quarterly Journal of Economics* 107: 407–438.

Mansuri, G. 1997. Credit layering in rural financial markets: Theory and evidence from Pakistan. Ph.D. thesis, Boston University.

Matsuyama, K. 2000. Endogenous inequality. *Review of Economic Studies* 67: 743–759.

Mauro, P. 1995. Corruption and growth. *Quarterly Journal of Economics* 110: 681–712.

Miguel, E., and M. Gugerty. 2005. Ethnic diversity, social sanctions, and public goods in Kenya. *Journal of Public Economics* 89: 2325–2368.

Miguel, E., S. Satyanath, and E. Sergenti. 2004. Economic shocks and civil conflict: An instrumental variables approach. *Journal of Political Economy* 112: 725–753.

Montalvo, J., and M. Reynal-Querol. 2005. Ethnic polarization, potential conflict, and civil wars. *American Economic Review* 95: 796–813.

Mookherjee, D., and D. Ray. 2001. *Readings in the theory of economic development*. London: Basil Blackwell.

Mookherjee, D., and D. Ray. 2002a. Contractual structure and wealth accumulation. *American Economic Review* 92: 818–849.

Mookherjee, D., and D. Ray. 2002b. Is equality stable? *American Economic Review* 92: 253–259.

Mookherjee, D., and D. Ray. 2003. Persistent inequality. *Review of Economic Studies* 70: 369–394.

Munshi, K., and J. Myaux. 2006. Social norms and the fertility transition. *Journal of Development Economics* 80: 1–38.

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

Myrdal, G. 1957. *Economic theory and underdeveloped regions*. London: Duckworth.

Nurkse, R. 1953. *Problems of capital formation in underdeveloped countries*. New York: Oxford University Press.

Obstfeld, M. 1994. The logic of currency crises. *Cahiers Economiques et Monétaires (Banque de France)* 43: 189–213.

Olson, M. 1965. *The logic of collective action: Public goods and the theory of groups*. Cambridge, MA: Harvard University Press.

Pande, R., and C. Udry. 2007. Institutions and development: A view from below. In *Proceedings of the 9th World Congress of the Econometric Society*, ed. R. Blundell, W. Newey, and T. Persson. Cambridge: Cambridge University Press.

Parente, S., and E. Prescott. 2000. *Barriers to riches*. Cambridge, MA: MIT Press.

Perotti, R. 1994. Income distribution and investment. *European Economic Review* 38: 827–835.

Perotti, R. 1996. Growth, income distribution, and democracy: What the data say. *Journal of Economic Growth* 1: 149–187.

Piketty, T. 1997. The dynamics of the wealth distribution and the interest rate with credit rationing. *Review of Economic Studies* 64: 173–189.

Powell, R. 2004. The inefficient use of power: Costly conflict with complete information. *American Political Science Review* 98: 231–241.

Powell, R. 2006. War as a commitment problem. *International Organization* 60: 169–203.

Pritchett, L. 1997. Divergence, big time. *Journal of Economic Perspectives* 11 (3): 3–17.
Quah, D. 1996. Twin peaks: Growth and convergence in models of distribution dynamics. *Economic Journal* 106: 1045–1055.
Ray, D. 1998. *Development economics*. Princeton: Princeton University Press.
Ray, D. 2006. Aspirations, poverty and economic change. In *Understanding poverty*, ed. A. Banerjee, R. Bénabou, and D. Mookherjee. New York: Oxford University Press.
Ray, D., and K. Sengupta. 1989. Interlinkages and the pattern of competition. In *The Economic Theory of Agrarian Institutions*, ed. P. Bardhan. Oxford: Clarendon Press.
Ray, D., and P. Streufert. 1993. Dynamic equilibria with unemployment due to undernourishment. *Economic Theory* 3: 61–85.
Ray, D., J.-M. Baland, and O. Dagnielie. 2007. Inequality and inefficiency in joint projects. *Economic Journal* 117: 922–935.
Robinson, J. 1998. Theories of 'bad policy'. *Policy Reform* 1: 1–46.
Romer, P. 1986. Increasing returns and long-run growth. *Journal of Political Economy* 92: 1002–1037.
Rosenstein-Rodan, P. 1943. Problems of industrialization of eastern and southeastern Europe. *Economic Journal* 53: 202–211.
Rosenzweig, M., and K. Wolpin. 1993. Credit market constraints and the accumulation of durable production assets in low-income countries: Investments in bullocks. *Journal of Political Economy* 101: 223–244.
Sambanis, N. 2001. Do ethnic and nonethnic civil wars have the same causes? A theoretical and empirical inquiry (part 1). *Journal of Conflict Resolution* 45: 259–282.
Sandler, T., and J. Forbes. 1980. Burden sharing, strategy and the design of NATO. *Economic Inquiry* 18: 425–444.
Sen, A. 1999. *Development as freedom*. New York: Alfred A. Knopf.
Sokoloff, K., and S. Engerman. 2000. History lessons: Institutions, factor endowments, and paths of development in the new world. *Journal of Economic Perspectives* 14 (3): 217–232.
Solow, R. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70: 65–94.
Stiglitz, J. 2002. *Globalization and its discontents*. New York: W.W. Norton.
Stiglitz, J., and A. Weiss. 1981. Credit rationing in markets with imperfect information. *American Economic Review* 71: 393–410.
Townsend, R. 1993. Risk and insurance in village India. *Econometrica* 62: 539–591.
Townsend, R. 1995. Consumption insurance: An evaluation of risk-bearing systems in low-income economies. *Journal of Economic Perspectives* 99 (3): 83–102.
Udry, C. 1994. Risk and insurance in a rural credit market: An empirical investigation in Northern Nigeria. *Review of Economic Studies* 61: 495–526.
Udry, C. 1996. Gender, agricultural productivity and the theory of the household. *Journal of Political Economy* 104: 1010–1045.
Weitzman, M. 1974. Free access vs. private ownership as alternative systems for managing common property. *Journal of Economic Theory* 8: 225–234.
World Bank. 2003. *World development report*. London: Oxford University Press.
Young, A. 1928. Increasing returns and economic progress. *Economic Journal* 38: 527–542.

Development Planning

Amiya Kumar Bagchi

Conscious plans for development of the economy as a whole over an extended period (say, five or ten years) were drawn up for the first time in the Soviet Union in the 1920s. The socialist countries of Eastern Europe, and the People's Republic of China have since then been the most consistent practitioners of development planning. However, the practice of drawing up development plans soon spread from the Soviet Union to non-socialist countries and some of the plans promulgated by the respective governments were also implemented, with different degrees of success.

In socialist countries, the broad outlines of the development plan have to be approved by the highest authority which may be the praesidium of the supreme legislative and executive body or the Party Congress convened for the purpose. However, the political and administrative authorities at the lower levels of administration, such as the county or the province, transmit information upwards regarding both the availability of resources and the felt needs of development. The actual implementation of the plan and the detailing of the outputs to be produced and the inputs to be used for executing the plan are delegated to the lower level authorities. As we shall see later, in socialist countries an almost continuous debate has been conducted regarding the degree of devolution of administrative authority and

decentralization of economic decision-making. By and large, the central leadership in socialist countries have taken the decisions regarding the strategic and long-term variables, such as the rate and sectoral composition of investment, the degree of openness of the economy, and the allocation of resources as between different regions, while leaving the tactical or short-run production decisions to the lower-level authorities.

Long-term development plans have to be based on a depiction of the structure of the economy and its probable evolution under the influence of different types of intervention by the government. In the Soviet Union in the 1920s, inspiration for the construction of models of a planned economy was drawn mainly from the works of Karl Marx. (For an anthology of translations of Soviet writings on the subject, see Spulber 1964.) In particular, by drawing on the schemes of expanded reproduction constructed by Marx (1893, 1894), G.A. Fel'dman constructed a two-sector model of development by assuming the economy to be closed and dividing it into two vertically-integrated sectors, one producing capital goods and the other consumer goods (Fel'dman 1928; Domar 1957). Fel'dman assumed that capital goods were the only limiting factor of production. An analytically equivalent model was constructed by P.C. Mahalanobis (1953). One interesting result of the Fel'dman–Mahalanobis model is the demonstration that given a constant technology and a constant capital–output ratio, the long-term rate of growth of the economy is determined by the proportion of investment devoted to the expansion of the capital goods sector.

In Fel'dman's model, the capital goods sector included all the intermediate goods needed for producing the final goods, as did the consumer goods sector. But the actual calculation of the output of a particular intermediate good needed to sustain a desired level of a particular capital good or consumer good could be made only after all the direct and indirect uses of the corresponding intermediate good had been traced. In trying to solve this problem, the Soviet planners early evolved the method of the material balances, under which, once, let us say, a given volume of output of finished steel had been decided upon, all the inputs directly and indirectly needed to sustain that level of output in the way of iron ore, coal, limestone, blast furnace and steel-smelting facilities, transport services and power would be worked out. This would generally involve several iterations until the demands and supplies of the different inputs converged. These exercises would be carried out for all the major items entering planning – and these could run to several hundred items (Montias 1959).

Wassily Leontief later worked out what has come to be known as the input–output method of analysis, which can be regarded as the logical completion of the method of material balances. (For a succinct summary of the available elaborations of the input–output models used in plan exercises, see Taylor 1975.)

In the Soviet Union and other socialist countries, considerable attention was paid to the use of mathematical methods for solution of large-scale planning problems and for finding out least-cost methods of carrying out given projects or programmes. The Russian mathematician L.V. Kantorovich has been credited with the discovery of the method of linear programming though the first convenient algorithm for solving such a programme was invented by G.B. Dantzig (Kantorovich 1965). However, the real problems of planning in the Soviet Union and other socialist economies have centred on questions of the use of prices or simulation of planning by markets, on the degree of decentralization of decision-making, and on the level and composition of maintainable investment, rather than on questions of which techniques to use to draw up plans.

Socialist economies such as the Soviet Union and China, soon after the beginning of planning, attained very high rates of investment: the rate of investment during the first five-year plan in Russia went up, for example, from 15 per cent to 44 per cent of national income between 1928 and 1932 (Ellman 1975). In China the ratio of investment to national income went up to 25 per cent at the end of her first five-year plan (1953–7), and in the 1970s the investment–income ratio generally stayed above 30 per cent. One result of the drive

to raise the rate of investment and construction was that the huge labour surpluses in these countries which had been prevalent in pre-revolution days were mopped up after the first few years of planning.

The problems that the socialist countries typically faced were well summed up by Mao Zedong in his famous talk on the ten major relationships (Mao 1956). According to Mao, in the context of Chinese development, maintenance of a balance was crucial in the relationships (i) between heavy industry, light industry and agriculture; (ii) between industry in coastal regions and industry in the interior; (iii) between civil investment and defence construction; (iv) between the state, the units of production and the actual producers; (v) between the central and local authorities; (vi) between Han, that is, the majority nationality, and the minority nationalities; (vii) between Communist Party authorities and cadres and non-members of the Party; (viii) between different policies fostering revolution rather than counter-revolution; (ix) between rewarding the correct policy-executors and punishing the wrongdoers; and (x) between China and the foreign countries. The relations (vi), (vii), (viii) and (ix) are political questions of broad importance involving socialist legality, the correct treatment of counter-revolutionary elements, but also questions with a mainly Chinese orientation. But the other relations involve mainly questions of economic strategy and have appeared in many different contexts. It has been felt in many socialist countries that not enough attention was paid until recently to the aim of raising the standards of living of the people. Too many resources were devoted to the development of heavy industry and too few to the growth of light industries catering for mass consumption (cf. Kalecki 1969).

It was also felt that because of the highly centralized character of management, the stress on investment and a general atmosphere of scarcity within which managers were to achieve certain quantitative goals, there is a tendency at the enterprise level in a socialist economy to hoard resources and to invest too much (Kornai 1980). Moreover, it was thought that in the drive to raise the rate of industrialization, while keeping prices stable by ensuring the supply of an adequate quantity of agricultural goods at fixed prices to the non-agricultural sector, plans had tended to discriminate against the rural producers. The allegation that Soviet industrialization was mainly financed by Russian peasants has been called into question by recent research (Ellman 1975; Vyas 1979). However, in many socialist countries, including China, moves have been made in recent years to increase the incomes of agricultural producers significantly and prices of agricultural products have been raised drastically with the same end in view. In China, deliberate attempts have also been made to bring down the ratio of accumulation (investment) to national income, to increase the rate of growth of light industry, and to provide greater incentives to peasants and industrial enterprises to change their product mix in response to changing demand patterns, to economize on scarce resources and to bring about a greater degree of flexibility of management (Ma Hong 1983; Xue Muqiao 1981). But the chief instruments of adjustment and reform have been changes in prices paid to producers of specified goods, especially agricultural commodities, and political and administrative decentralization, rather than allowing producers to change their prices or their investment patterns independently of political authorities. The main underpinnings for an egalitarian distribution of income in the shape of a comprehensive public distribution and social security system and of stability in consumer prices have so far been maintained in all socialist economies. One reason for this is that there is no simple way in which an economy-wide reform can be instituted so that prices either equal prices of production or equalize supplies and demands in all markets but do not bring about other undesirable side-effects in the form of an increase in inequality of income distribution or unemployment.

Socialist economics have been concerned in recent years with making a transition from a regime of extensive to one of intensive growth, that is, from one where economic growth is accelerated by raising the rate of investment or the application of labour to one where it can be raised by increasing the productivity of agents of

production. Economic reforms are seen as one means of doing this. Increasing the rate of innovation and adaptation and absorption of imported technology are seen as other means of doing this. It is in the latter area that the relations between socialist countries and advanced capitalist economies become crucial. Socialist economies are striving to import improved technologies from the USA, the EEC countries and Japan without becoming dependent on them or becoming heavily indebted to them. On the other side, the advanced capitalist countries are trying to increase their markets in socialist countries without selling them technologies which could make them economically or militarily stronger than the capitalist countries in the future.

The problems that the non-socialist countries have faced in formulating credible development plans have been far more complex than those discussed above and their success in implementing them has been far more mixed.

While the Soviet theorists and Mao took a socialist system to be the environment in which a development plan was to be located, most other theorists were not explicit about the kind of system they had in mind when they proposed specific plans for development of the underdeveloped economies. Paul Rosenstein-Rodan's pioneering attempt to formulate appropriate plans for development of the Eastern European countries after World War II can be taken to be the genesis of what came to be known as the 'balanced growth' doctrine (Rosenstein-Rodan 1943). Ragnar Nurkse (1953) developed some of these ideas further in his writings. According to these theorists, in a poor underdeveloped economy, a credible development plan would have to consist of a programme for a simultaneous and balanced development of all the important sectors in the economy, so that expanding demands are met by matching supplies, and vice versa. Moreover, this process of balanced growth would lead to the realization of internal economies of scale and external effects arising from learning processes, and a decline in uncertainty faced by buyers, sellers and investors (see, in this connection, Dobb 1960, ch. 1). Maurice Dobb (1951), Nurkse and Lewis (1954) all stressed the necessity and possibility of mobilizing underemployed and unemployed labour for the purpose of capital formation in underdeveloped economies.

The balanced growth doctrine has the advantage that it can be embodied in specific development plans elaborated out of the Fel'dman–Mahalanobis models, and the input–output models devised by Leontief and his co-workers and later followers. But even before such models had been elaborated to take account of all the interconnections involved in a dynamic income generation process, it was clear that in a non-socialist economy, a development plan, however well-formulated, was likely to run into problems because of the lack of concordance between planners' goals and private sector goals and lead to political side-effects which could derail it before it had really had time to run its course.

It is useful to analyse some of these problems by using the four-sector model of development which Mahalanobis (1955) used as the scaffolding for drawing up the draft second five-year plan of India. In this model, the economy is divided into four vertically integrated sectors, the first producing capital goods by factory methods, the second producing consumer goods by factory methods, the third producing consumer goods by handicraft methods and the fourth producing services by labour-intensive methods. The idea behind this classification was that a designated proportion of the output of capital goods industries would be devoted to their own expansion in order to promote growth, while the handicraft and service sectors would meet much of the demand for consumer goods and services generated by increasing incomes and at the same time mobilize underemployed and unemployed labour thus minimizing the need to divert investible resources to the factory sector for production of more consumer goods.

However, one of the basic conditions for employment of more labour would be that the new workers can be fed and clothed (Nurkse 1953; Lewis 1954). It cannot be assumed that some automatic mechanism would spring up for diverting food from the farms to factories in urban or rural areas. Kalecki (1955) was one of the first to emphasize the importance of ensuring a smooth supply of wage goods and keeping the rate of

saving high by curbing consumption for financing development in Third World countries.

Most Third World countries were, however, characterized by various kinds of landlordism or other semi-feudal constraints such as debt bondage, the use of non-market coercion, etc., limiting farm output. The failure to carry out thoroughgoing land reforms which would vest the ownership and management of the land in the hands of the actual cultivators also meant that traders and moneylenders could continue to prosper by exacting extortionate margins on goods sold or bought and charging usurious interest rates on loans to the poor in the countryside. These conditions also facilitate political coalitions between landlords, traders and moneylenders blocking the process of reforms to endow the peasants with the incentive and wherewithal to produce more and meet the needs of industrialization.

As Kalecki (1955) realized, if the marketed surplus fails to go up, an increase in the rate of investment as envisaged by all development plans would soon meet an inflation barrier (since the income elasticity of the demand for food is high and its price elasticity is low). A rising output of farm products does not in itself guarantee a rising volume of marketed surplus. If the consumption of the suppliers of farm products rises proportionately more than farm output, then the marketed surplus will fall. With a landlord-dominated farm sector, traders and landlords generally command enough credit and other assets to ensure that the rest of society pays a stiffly rising price for farm products whenever the output of agriculture falters (say, because of adverse weather conditions or floods or pests). If the government can be persuaded to run a procurement programme so that it is committed to buying up any agricultural supplies coming on the market at a minimum price, but cannot force the landlords or traders to deliver the grain (or cotton or oilseeds) at that price, then a ratchet is put under the prices of farm products. Thus the physical rate of growth of farm output puts only an outside limit on the rate of growth of non-agricultural output: the actual limit (which is lower) is set by the ownership pattern of agricultural assets and by the conditions of sales of agricultural commodities. When the farm sector is dominated by landlords, the rate of growth of agricultural output interacts with such factors as luxury consumption of the rich, the tendency to speculation whenever the harvest is poor, the extremely skewed distribution of credit, and public support for farm prices to produce a constricting limit on industrial growth. In a socialist economy, with fixed prices of food grains, a comprehensive public distribution system and the abolition of speculation, a similar rate of agricultural growth would be consistent with a much higher rate of industrial development. (A non-socialist economy with a relatively egalitarian distribution of landholdings would pose lesser problems for growth than a landlord-dominated society.) Thus, referring back to the four-sector Mahalanobis model, it can be seen that mobilization of labour to produce labour-intensive consumer (or capital) goods would require as a precondition a durable solution of the problem of supply of the needed foodgrains and other agricultural goods.

It can also be seen that stepping up the rate of investment in the economy would require stepping up the rate of savings to an equivalent amount. Such a stepping up of saving would not normally occur on a voluntary basis in an underdeveloped economy which had been stagnating before the onset of development planning. So the government would have to tax the rich in order to release the necessary resources for investment and keep the demand for foodgrains and other goods with inelastic supplies within reasonable bounds (compare Kalecki 1955).

However, in a non-socialist economy the government generally fails to curb the increase in the purchasing power in the hands of the rich to an adequate extent. The rich then not only demand and commandeer more of the scarce resources which should go into investment, they also do not purchase sufficient amounts of the handicrafts or the labour-intensive consumer goods which, in the four-sector Mahalanobis model, are supposed to satisfy the increasing demands released in the economy. Thus excess capacity emerges (or continues) in many sectors of the economy (including capital goods turned out by government and private factories), with attendant

unemployment, even while there is excess demand in other sectors (see Bagchi 1970). In particular, the rich generally demand newer types of luxury goods produced in the advanced capitalist countries. If these cannot be produced at home, they will be imported from abroad. Since the failure to step up the rate of aggregate saving to an adequate extent or channel investment into the sectors which accelerate the growth of the economy in any case lead to balance of payments deficits, most Third World countries attempting to plan their development will also have foreign trade regimes characterized by exchange controls, high tariffs on permitted imports, and quantitative restrictions on imports and exports. Under these circumstances, restricted importables will normally fetch high premia in domestic currency and it will be profitable to smuggle them in or produce them behind the walls of the high tariffs and quantitative restrictions of various kinds, thus leading to further diversion of resources.

Some of the difficulties underdeveloped countries faced in obtaining enough foreign capital inflows for financing development were approached via the so-called 'two-gap' models of aid, trade and development (Chenery and Bruno 1962; Manne 1963; McKinnon 1964). In these models, on plausible assumptions about the desired rate and pattern of growth, a gap between ex ante exports and imports and a parallel gap between ex ante investment and savings are estimated. Since exports of most underdeveloped primary-commodity-producing countries are price and income-inelastic, and many of them also face non-price barriers in trade, whereas their planned investment is often relatively import-intensive, it was often found that the ex ante trade gap was larger than the ex ante investment-saving gap (Landau 1971). It was argued then that the planning authorities of the country concerned should plan to borrow or canvas for aid to cover the larger of the two gaps, and then development could proceed as planned.

Few countries were, however, in the happy position of being able to borrow or receive as aid whatever foreign capital inflow the planning exercises indicate as the optimum amount, even in the days when official grants and loans were less niggardly than they have become in the last decade or more. Moreover, the two-gap models themselves did not indicate the desirable or the feasible method of adjustment of the two gaps to each other ex post, and to the amount of foreign capital actually received. Even if the foreign aid or loans equalled the larger of the two gaps, the planning authorities could not leave the adjustment process to autonomous market forces, but had to adopt specific policies to bring about an appropriate adjustment process (Vanek 1967, ch. 6). When the foreign trade gap is dominant, for example, it is appropriate to allow savings to go down, in order to make the investment–savings gap rise to the export–import gap rather than stimulate (import-intensive) investment and increase the trade gap further. Under a wide variety of conditions, both policy-induced and market-induced adjustment processes would lead to a rise in consumption and a slowing down of investment (because of the uncertainty as regards the availability of imports and because of inventory accumulation as a result of excess capacity in import-constrained sectors). Thus where foreign capital inflows are a binding constraint, a negative relation may well be observed between inflows of aid and domestic savings effort (Rahman 1968; Griffin 1970).

Moreover, with overvalued foreign exchange and with a perceived disadvantage in investing in fields requiring new, foreign-controlled technology, there may also be hidden outflows of domestic capital to safe havens of hoarding or investment even while a substantial amount of foreign capital is coming in under official auspices.

Besides two-gap models, there were other advances in the understanding of development plans. It was realized that where the supply of foreign exchange was a constraint, planners might try to build up intersectoral linkages so as to provide for machines to produce machines or produce higher-order intermediate goods, and so attempt to accelerate economic development to the maximum possible extent (Raj and Sen 1961). Optimizing exercises involving time-lags could be carried out with the same class of models. However, the implementation of the indicated development plans by non-socialist developed countries would flounder on their inability (a) to buy the technology, which

was often patented or otherwise owned by transnational corporations, on reasonable terms, (b) to devise appropriate social and organizational mechanisms for absorbing and diffusing the technology, and for exacting the needed savings and allocation of investment out of the economy (see Bagchi 1982, ch. 9).

In the field of application of input–output analysis and social accounting matrices to development plan models there have also been significant advances. Although it was sometimes suggested that different clusters of industries of the economy (such as heavy industry, light industry and agriculture) could grow at very different rates, because the current input–output flow system regularly displayed gaps between some sectors and close ties as between others (Manne and Rudra 1965), it was realized that the flows of demands generated by the planning process would tie the growth patterns of different sectors tightly together, as we have seen already. Significant advances have been made in applying the social accounting matrices to plan models, and the implicit multipliers relating the growth of particular sectors or factor incomes to the rest of the economy have been utilized to predict the income generation and distributional implications of different patterns of plan expenditures (Pyatt and Round 1979; Taylor 1979).

However, it is one thing to devise models for development and another thing to implement them in underdeveloped countries with big landlords, propertied classes which are divided among themselves and which are continually attracted to the metropolitan centres by the lure of more modern life styles, safety of investments against threats of revolutions, and other considerations. Rosenstein-Rodan (1943) had conceived of the development plan as being carried out by a 'trust' which could internalize all external effects and all secondary effects of investment. In actual fact the limits of organization either through the market or in firms or governmental organizations, and the temptation to resort to opportunistic behaviour to the detriment of the collective good have been far more prevalent in non-socialist underdeveloped countries than in the socialist economies. The propertied, or more narrowly, capitalist groups have found it very difficult to evolve codes of cooperation without which confidence in the future and long-term investment become very fragile plants (see Axelrod 1984).

Even while the balanced growth doctrine was being evolved, Albert Hirschman had proposed exploiting the profitability-signalling property of disequilibrium situations to recommend a path of development along which imbalances were deliberately engineered (Hirschman 1958). In fact, as it turned out, capitalists more often reacted the 'wrong' way to disequilibria, by cornering scarce commodities, using political levers to raise barriers against entry into their favoured pastures, playing intertemporal arbitrage games to defeat the planners' intentions (see Bagchi 1966; Hirschman 1968). The obstacles to the execution of development plans in non-socialist countries had been foreseen in the 1950s by many Marxists, of whom Paul Baran was the most prominent, (Baran 1952, 1957), and by other social scientists such as Gunnar Myrdal (1957). In the general atmosphere of crisis in the world economy, there is sometimes an agreement between proponents at both extremes of the political spectrum that development planning is impossible in Third World countries. What both experience and analysis indicate, however, is that the implementation of development plans is likely to be fraught with contradictions. There will be imbalances between regions, increasing differentiation of peasantry, tensions between development of the public and private sectors, conflicts between interests of local development and interests of transnational corporations and their local collaborators, and questions will be raised and often resolved through bloody confrontations regarding the appropriate political regimes. It is through the mobilization of ordinary people to tackle these manifold contradictions and to fight the vested interests blocking the progress of development programmes that further advances will be made. National planning, in that sense, has been and will always be, intimately tied up with politics. But for most Third World countries development planning remains an essential part of the programme for charting their own future.

See Also

- ▶ Fel'dman, Grigorii Alexandrovich (1884–1958)
- ▶ Mahalanobis, Prasanta Chandra (1893–1972)
- ▶ Planned Economy

References

Axelrod, R. 1984. *The evolution of cooperation*. New York: Basic Books.

Bagchi, A.K. 1966. Shadow prices, controls and tariff protection in India. *Indian Economic Review, New Series* 1(1): 22–44.

Bagchi, A.K. 1970. Long-term constraints on India's industrial growth 1951–8. In *Economic development in South Asia*, ed. E.A.G. Robinson and M. Kidron. London: Macmillan.

Bagchi, A.K. 1982. *The political economy of underdevelopment*. Cambridge: Cambridge University Press.

Baran, P.A. 1952. The political economy of backwardness. *Manchester School of Economic and Social Studies* 20(1): 66–84.

Baran, P.A. 1957. *The political economy of growth*. New York: Monthly Review Press.

Chenery, H.B., and M. Bruno. 1962. Development alternatives in an open economy: The case of Israel. *Economic Journal* 72: 79–103.

Dobb, M.H. 1951. *Some aspects of economic development*. Delhi: Delhi School of Economics.

Dobb, M.H. 1960. *An essay on economic growth and planning*. London: Routledge & Kegan Paul.

Domar, E. 1957. A Soviet model of growth. In *Essays in the theory of economic growth*, ed. E. Domar. New York: Oxford University Press.

Eckstein, A. 1977. *China's economic revolution*. Cambridge: Cambridge University Press.

Ellman, M. 1975. Did the agricultural surplus provide the resources for the increase in investment in the USSR during the first five year plan? *Economic Journal* 85: 844–863.

Fel'dman, G.A. 1928. K teorii narodnogo dokhoda. *Planvoe khoziaistvo* 11, 12. Trans. as '*On the theory of growth rates of national income*', I and II, in Spulber (1964).

Griffin, K. 1970. Foreign capital, domestic savings and economic development. *Bulletin of the Oxford University Institute of Economics and Statistics* 32(2): 99–112.

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

Hirschman, A. 1968. The political economy of import-substituting industrialization in Latin America. *Quarterly Journal of Economics* 82(1): 1–32.

Kalecki, M. 1955. The problem of financing economic development. *Indian Economic Review* 2(3): 1–22.

Kalecki, M. 1969. *Introduction to the theory of growth in a socialist economy*. Oxford: Basil Blackwell.

Kantorovich, L.V. 1965. *The best use of economic resources*. Oxford: Pergamon.

Kornai, J. 1980. *Economics of shortage*, 2 vols. Amsterdam: North-Holland.

Landau, L. 1971. Saving functions for Latin America. In *Studies in development planning*, ed. H.B. Chenery et al. Cambridge, MA: Harvard University Press.

Lewis, W.A. 1954. Economic development with unlimited supplies of labour. *Manchester School of Economic and Social Studies* 22: 139–191.

Lockett, M., and C.R. Littler. 1983. Trends in Chinese enterprise management 1978–1982. *World Development* 11(8): 683–704.

Ma Hong. 1983. *New strategy for China's economy*. Beijing: New World Press.

Mahalanobis, P.C. 1953. Some observations on the process of growth of national income. *Sankhya* 12(4): 307–312.

Mahalanobis, P.C. 1955. The approach of operational research to planning in India. *Sankhya* 16(1 and 2): 3–130.

Manne, A. 1963. Key sectors of the Mexican economy. In *Studies in process analysis*, ed. A. Manne and H.M. Markowitz. New York: Wiley.

Manne, A., and A. Rudra. 1965. A consistency model of India's Fourth Plan. *Sankhya, Series B* 27(1 and 2): 57–144.

Mao Zedong. 1956. On the ten major relationships. In *Selected works of Mao Tse-tung*, vol. V. Peking: Foreign Languages Press, 1977.

Marx, K. 1893. *Capital: A critique of political economy*, vol. II. Trans. from the 2nd German edn of 1893, ed. F. Engels. Moscow: Foreign Languages Publishing House, 1957.

Marx, K. 1894. *Capital: A critique of political economy*, vol. III. Trans. from original German edn of 1894, ed. F. Engels. Moscow: Foreign Languages Publishing House, 1966.

McKinnon, R.I. 1964. Foreign exchange constraints in economic development and efficient aid allocation. *Economic Journal* 74: 388–409.

Montias, J.M. 1959. Planning with material balances in Soviet-type economies. *American Economic Review* 49: 963–985.

Myrdal, G. 1957. *Economic theory and underdeveloped regions*. London: Duckworth.

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

Nuti, D.M. 1981. Socialism on earth. *Cambridge Journal of Economics* 5(4): 391–403.

Pyatt, G., and R.I. Round. 1979. Accounting and fixed price multipliers in a social accounting matrix framework. *Economic Journal* 89(356): 850–873.

Rahman, M.A. 1968. Foreign capital and domestic savings: A test of Haavelmo's hypothesis with cross-country data. *Review of Economics and Statistics* 50: 137–138.

Raj, K.N., and A.K. Sen. 1961. Alternative patterns of growth under conditions of stagnant export earnings. *Oxford Economic Papers* 13(1): 43–52.

Rosenstein-Rodan, P.N. 1943. Problems of industrialisation of eastern and south-eastern Europe. *Economic Journal* 53: 202–211.
Spulber, N. (ed.). 1964. *Foundations of Soviet strategy for economic growth: Selected Soviet essays, 1924–1930*. Bloomington: Indiana University Press.
Taylor, L. 1975. Theoretical foundations and technical implications. In *Economy-wide models and development planning*, ed. C.R. Blitzer, P.B. Clark, and L. Taylor. London: Oxford University Press.
Taylor, L. 1979. *Macro models for developing countries*. New York: McGraw-Hill.
Vanek, J. 1967. *Estimating foreign resource needs for economic development: Theory, method and a case study of Colombia*. New York: McGraw-Hill.
Vyas, A. 1979. Primary accumulation in the USSR revisited. *Cambridge Journal of Economics* 3(2): 119–130.
Xue Muqiao. 1981. *China's socialist economy*. Beijing: Foreign Languages Press.

Dialectical Materialism

Roy Edgley

Dialectical materialism is what Engels in the Preface to the second edition of the *Anti–Dühring* calls 'the communist world outlook'. The term 'dialectical materialism' was probably first used by 'the father of Russian Marxism', Plekhanov, in 1891. It was unknown to Marx himself. Engels came close to coining it, and it was in fact Engels who was chiefly responsible for founding dialectical materialism: the relevant books are his *Anti-Dühring* (published 1877–8), *Dialectics of Nature* (written 1878–82, first published 1927) and *Ludwig Feuerbach and the End of Classical German Philosophy* (published 1886–8).

Marx's distinctive intellectual work was a theory of society, specifically of economics as the basis of society, and in particular, in his Capital, of the economics of capitalism. This social theory is known as 'historical materialism'. Dialectical materialism is distinguished from and related to historical materialism in various ways. For a start, it is a theory not simply about society but about reality as a whole, nature as well as society. The presupposition of dialectical materialism, in the words of the Preface to the second edition of the Anti-Dühring, is that 'in nature . . . the same . . . laws . . . force their way through as those which in history govern. . . events'. Thus the basic theories of dialectical materialism are formulated as laws of a completely universal application, governing 'nature, society, and thought' (Anti-Dühring, pt. I, ch. xiii). Second, in accordance with this claim of complete universality, dialectical materialism is generally regarded as philosophy, whereas historical materialism claims to be not philosophy but science, social science. Third, and further to its status as philosophy rather than science, it yields a very general account of the structural relations of the special sciences.

What we have here is a traditional rather than distinctively modern conception of philosophy and its relation to science. A philosophy is a 'world outlook', a synoptic view of the totality of things achieved in this case by revealing in the special sciences a common content, an underlying general conception of reality that they all share and express. This philosophy is therefore itself regarded as scientific, a kind of 'natural philosophy' exemplified in and supported by the findings of the special sciences as they investigate their own limited domains of reality.

Engels' case for dialectical materialism has a special political point for Marxism: namely to argue its scientificity. The case is that historical materialism shares with the natural sciences not, or not only, a method of inquiry but the same 'world outlook'. Historical materialism's claim to scientific status is of crucial importance to it. Marxism rejects as more or less unscientific both other (bourgeois) social theories and other forms of socialism such as ethical or utopian socialism. It seeks to recruit to its support the cognitive authority of science, distinguishing itself within the socialist movement as what Engels called 'scientific socialism'.

With the rise of the bourgeoisie, the Scientific Revolution and Enlightenment had seen the establishment of the natural sciences of astronomy, physics and chemistry. But it was not until the late 18th and early 19th centuries that the social sciences began to develop, in a process in which social theory sought to transform itself from

philosophy into science. When in the 1840s Marx and Engels embarked on their construction of a unified and comprehensive social science they rejected as models not only the existing (bourgeois) forms of social theory, such as classical political economy, but also the earlier forms of the modern natural sciences. In their view each major social revolution, basically in the dominant mode of production, involves also an ideological revolution, a revolution in world outlook. Thus in the transition from feudalism to capitalism the religion-dominated ideology of the Middle Ages had given way to a general conception of reality shaped decisively by natural science. A central element in this 'natural philosophy' of the bourgeois era was the so-called 'mechanical philosophy'. According to this, the objective reality investigated by science is a mechanism of matter in motion, a kind of cosmic clockwork, and understanding this reality is knowing the laws governing the mechanism. Between this and the new world outlook of the rising working class there would be both continuities and breaks, but even the breaks would be prepared in bourgeois society. Thus for Marx and Engels the natural sciences in the later part of the 18th century had already begun to change in a significant way, developing one of the most basic and characteristic aspects of the new communist point of view.

Newton has said that in the beginning God threw the planets round the sun, creating processes ruled by the laws of motion and gravity, processes of repetitive or cyclical movement in a system that itself remained essentially unchanged and unchanging. But the Kant–Laplace nebular hypothesis rejected this static conception and replaced it with a theory representing the present solar system as the latest stage in a long and continuing evolution. For Marx and Engels, what this showed was that 'Nature has a history' and that the natural sciences were themselves evolving from a static conception of nature towards a recognition of its historicity. Lyle's geology and Darwin's biology seemed to confirm this tendency.

The key to understanding this mode of non-cyclical (progressive) change, according to Marx and Engels, had already been prepared within philosophy, by Hegel. This key was the dialectic. They believed, however, that in Hegel the dialectic suffers a deformation characteristic of philosophy, especially bourgeois philosophy. Its form is idealist, not materialist. For Hegel, in other words, reality is ideal, the activity and product of spirit or mind, so that its dialectical nature is its nature as an essentially non-material process.

Dialectical materialism, then, results from the crossing of two bourgeois philosophies, Hegel's dialectical idealism and the mechanistic materialism of the Scientific Revolution and Enlightenment. Hegel's idealism is incompatible with materialism, and the mechanicism of traditional materialism is incompatible with dialectic. They are therefore rejected, leaving a conception of reality that is both dialectical and materialist.

In this unification of dialectic and materialism both doctrines are transformed. Traditional materialism, being non-dialectical, is reductive, a 'nothing-but' theory: it holds that reality is nothing but matter in motion, and thus that processes that appear to be otherwise are really not otherwise because they are 'reducible' to matter in motion. Ideas, for example, are reducible to and ultimately identical with material processes. On this view change itself, that is the development of difference and novelty, is really nothing but the continuation of the same basic processes and laws. The dialectical point of view, on the contrary, claims that concrete reality is a unity, but a differentiated unity in which the elements are all essentially interrelated and integrated but not reducible to one another. Indeed, differentiation means opposition and contradiction. Thus the material and ideal themselves are really different and opposed, but they exist and are related within a unity in which the material is basic: matter can exist without mind but not mind without matter. Epistemologically, then, physics yields, contrary to idealism, knowledge of an objective mind-independent reality, and forms the base of a unified system of the special sciences that, contrary to traditional materialism, are nevertheless not reducible to physics. Moreover, differentiation is not a static condition but an active process. Reality is a unity that is specifically contradictory, and it is the conflict of opposites within unity that

drives reality onwards in a historical process of progressive change. This change is both evolutionary and revolutionary, both quantitative and qualitative: its revolutionary or discontinuous moments yield genuine novelty, change of a qualitative kind. Mind itself on this view is such an emergent novelty.

This dialectical world outlook is standardly summarized in the form of three fundamental laws: (1) the law of the unity of opposites, according to which concrete reality is a unity in conflict, a unity that is contradictory; (2) the law of the negation of the negation, which says that in the conflict of opposites one term negates the other, but preserves something of the negated term and is then itself negated in a historical process that in this way rises to ever higher levels; (3) the law of the transformation of quantity into quality, which says that in the evolutionary process of gradual quantitative change contradictions intensify to the point at which a revolutionary qualitative change occurs. The popularized version of these laws represents dialectic as a triadic process of thesis, antithesis and synthesis.

Dialectic claims to revolutionize our thinking at all levels, including – even most particularly – the intellectually fundamental level of logic. Among its most controversial elements is its use of the logical category of contradiction. Dialectic presupposes the doctrine that there are contradictions in reality, and is thought to imply that therefore traditional formal logic, with its central principle of non-contradiction, must be superseded by a logic that permits contradictory propositions as true of this contradictory reality. The orthodox rejoinder has argued that two ideas can be contradictory but that such ideas cannot both be true, i.e. that reality itself cannot be contradictory. Hegel rejects this distinction between ideas and reality, but may be seen as ultimately accepting, through his idealism, the orthodox view that contradiction is a relation between ideas. What is distinctive, even outrageous, about dialectical materialism is that it takes the logical category of contradiction to be applicable to material reality.

What are the implications of dialectical materialism for economics? Economic theory, on this view, takes the form of laws in which major contradictions are identified within the processes of production, exchange, and distribution, and are used to explain historical change in society. In particular, these laws reveal how the gradual intensification of contradictions leads to crisis and ultimately to a revolution in which a qualitatively new economic system establishes itself.

But dialectical materialism has implications not only for the form of economic theory but also for the relation in which economics stands to the other social sciences, such as political science. First, the totalizing perspective of dialectic, according to which all things are so closely integrated that they can be understood only in their interrelation, rejects the conception of economics as a specialist social science capable of understanding its own domain of social phenomena independently of other domains and other social sciences. For the dialectic, economics is less a social science than an integral part or aspect of social science, of a comprehensive and unified theory about a unified, if contradictory, social totality. Second, however, materialism asserts that within the social totality economic processes have overriding importance. The general philosophical materialism associated with the rise of natural science contrasts matter with mind and ideas, and holds that matter is the most fundamental, or even the only ultimate, component of reality. In application to society in distinction from nature, materialism contrasts ideas and theory with practice and claims that the most fundamental aspect of any social system is its most material practice, its economic practice, and in particular its mode of (material) production. Thus for dialectical materialism, social structure and social change in general are explained ultimately in terms of economic structure and economic change. Economics is the most basic part of social science.

Indeed, under the sway of dialectical materialism Marxism has tended to exaggerate this doctrine to the point of vulgarization. In representing the scientificity of historical materialism as consisting in its sharing a world outlook with the natural sciences, dialectical materialism conceives

historical materialism as a natural science of society. This attempt to combine dialectic and materialism within the general perspective of natural science has been a standing temptation to leave within 'the communist world outlook' unreconstructed residues from the bourgeois world outlook. The result has been a variety of intellectual pressures converging on an influential distortion, namely the vulgar version of Marxism that Lenin labelled 'economism'.

On the side of dialectic, the orthodox view that contradiction, as a logical relation, is a relation between ideas seems incompatible with its application to material reality. In consequence, the category of contradiction has tended to be identified with that of conflict (conflict of forces) and its specifically logical and critical content evacuated. What this has helped to undermine is the possibility of conceiving the social science of historical materialism as social critique.

On the side of materialism, classical scientific materialism is reductive and determinist, and conceives of 'matter' as an inert substance subject to 'iron laws' of nature. For a Marxism under the influence of this tendency, the political and theoretical superstructure are epiphenomena of society's material base. Only that material base, the economy, and perhaps only its most material aspect, technology, has real causal agency. The effect of this on socialist strategy is anti-Marxist: concentration on working-class action within the economic base rather than its extension to politics and the state. In fact, even this limited activity is threatened as either impossible or unnecessary by the conception of the science of economics encouraged by a materialism of the natural science sort. Though it was Engels who was chiefly responsible for dialectical materialism, Marx himself sometimes lends support to this version of economics. In the Preface to the first German edition of Capital he refers to 'the natural laws of capitalist production' as 'tendencies working with iron necessity towards inevitable results'; and in the Afterword to the second German edition he speaks favourably of the reviewer who says that 'Marx treats the social movement as a process of natural history, governed by laws not only independent of human will, consciousness and intelligence, but rather, on the contrary, determining that will, consciousness and intelligence . . .'. Whatever space this leaves for socialist action, if any, it seems inadequate for anything as large in scale and conscious in purpose as revolutionary class war. Lenin, though a committed believer in dialectical materialism, found it necessary to argue persistently against the anti-revolutionary tendencies of economism.

Marx once declared that he was not a Marxist. It was among the first generation of his followers after Marx's death that Marxism took shape, in the period that culminated in the Russian Revolution. Those followers learned their Marxism chiefly from the two most famous books of the founders, Marx's Capital and Engels' Anti-Dühring, the former regarded as constituting the basic economic science of historical materialism, the latter the philosophy of Marxism, specifically dialectical materialism. Dialectical materialism was an essential component of that first-generation Marxism, the generation of the Second International. It became, and remained, equally central to Soviet communism and to the Communist Party orthodoxy established under Soviet leadership. Between the two world wars, as Soviet communism slid into the tyranny of Stalinist dictatorship and party bureaucracy, this first Marxist philosophy of dialectical materialism came under attack from within that part of the Marxist movement outside the USSR and began to give way to a second form of Marxist philosophy. This was Marxist humanism, since then the characteristic form of 'Western Marxism'. Its chief theorists were Lukacs, Korsch and Gramsci, followed by the thinkers of the Frankfurt School and by Sartre's attempt to fuse Marxism and Existentialism. They attacked the materialism of the natural sciences, and in emphasizing Marx's debt to Hegel and dialectic insisted on the necessary roles in social change of politics and ideology. Their revisions of Marxism found some confirmation in the rediscovery, in the 1920s and 1930s, of Marx's early writings, especially his Economic and Philosophical Manuscripts of 1844. In their turn, since the 1960s these Hegelianizing tendencies have themselves

come under attack, chiefly from Althusser and his followers. But 'diamat' (to use the abbreviated name of dialectical materialism common in the USSR) has remained characteristic mainly of Soviet communism and of the Communist Parties dominated by Russia.

See Also

- ▶ Dialectical Reasoning
- ▶ Economic Interpretation of History

Bibliography

Colletti, L. 1969. *Marxism and Hegel*. London: NewLeft Books, 1973.

Colletti, L. 1975. Marxism and the dialectic. *New Left Review* 93: 3.

Engels, F. 1877–8. *Anti-Dühring.* Moscow: Foreign Languages Publishing House, 1954.

Engels, F. 1886–8. *Ludwig Feuerbach and the end of classical German philosophy.* In *Selected Works*, ed. K. Marx and F. Engels, vol. 2. Moscow: Foreign Languages Publishing House, 1962.

Engels, F. 1927, written 1878–82. *Dialectics of nature.* Moscow: Progress Publishers, 1974.

Graham, L.R. 1973. *Science and philosophy in the soviet union*. London: Allen Lane.

Jordan, Z.A. 1967. *The evolution of dialectical materialism*. London: Macmillan.

Lefebvre, H. 1939. *Dialectical materialism*. London: Jonathan Cape, 1968.

Lenin, V.I. 1895–1916. *Philosophical notebooks,* vol. 38 of the *Collected works.* Moscow: Foreign Languages Publishing House, 1963.

Lenin, V.I. 1902. *What is to be done?* Moscow: Progress Publishers, 1969.

Lenin, V.I. 1908. *Materialism and empirico-criticism*. Moscow: Foreign Languages Publishing House, 1952.

Tse-tung, Mao. 1937. On contradiction. In *Selected works*, vol. 2. London: Lawrence and Wishart, 1954.

Marx, K. 1888, written 1845. *Theses on Feuerbach.* In *The German Ideology* (written 1845–6), ed. K. Marx and F. Engels, London: Lawrence & Wishart, 1970.

Norman, R., and S. Sayers. 1980. *Hegel, Marx and dialectic*. Brighton: Harvester.

Plekhanov, G.V. 1908. *Fundamental problems of marxism*. London: Lawrence & Wishart, 1969.

Sartre, J.-P. 1960. *Critique of dialectical reason*. London: New Left Books, 1976.

Stalin, J.V. 1924. *Problems of leninism*. Moscow: Foreign Languages Publishing House, 1945.

Wetter, G.A. 1952. *Dialectical materialism*. London: Routledge and Kegan Paul, 1958.

Dialectical Reasoning

Gareth Stedman Jones

Keywords

Capitalism; Civil society; Dialectical materialism; Dialectical reasoning; Engels, F.; Equal exchange; Hegel, G. W. F.; Marx, K. H.; Materialist conception of history; Political economy; Proudhon, P. J.; Ricardo's theory of value; Surplus labour; Use value

JEL Classifications

B4

This notoriously elusive and multifaceted notion assumed importance in the history of political economy because Marx's 'critique of political economy', *Capital,* and particularly its first draft, the *Grundrisse* of 1857–8, was presented in a dialectical form. Part of the difficulty of encapsulating the dialectic within any concise definition derives from the fact that it may be conceived as a method of thought, a set of laws governing the world, the immanent movement of history or any combination of the three. The dialectic originated in ancient Greek philosophy. The original meaning of '*dialogos*' was to reason by splitting in two. In one form of its development, dialectic was associated with reason. Starting with Zeno's paradoxes, dialectical forms of reasoning were found in most of the philosophies of the ancient world and continued into medieval forms of disputation. It was this form of reasoning that Kant attacked in his distinction between the logic of understanding which, applied to the data of sensation, yielded knowledge of the phenomenal world, and dialectic or the logic of reasoning, which proceeded independently of experience and purported to give knowledge of the transcendent order of things in themselves. In another form of dialectic, the focus was primarily upon process: either an ascending dialectic in which the existence of a higher reality is demonstrated, or a descending

form in which this higher reality is shown to manifest itself in the phenomenal world. Such conceptions were particularly associated with Christian eschatology, neo-platonism and illuminism, and typically patterned themselves into conceptions of original unity, division or loss, and ultimate reunification.

For practical purposes, however, the form in which the dialectic was inherited and modified by Marx was that in which it had been elaborated by Hegel. 'Hegel's dialectics is the basic form of all dialectics, but only *after* it has been stripped of its mystified form, and it is precisely this which distinguishes my method' (Marx, letter to Kugelmann, 6 March 1868).

In Hegel, the dialectic is a self-generating and self-differentiating process of reason (reason being understood both to be the process of cognition and the process of the world). The Hegelian Absolute actualizes itself by alienating itself from itself and then by restoring its self-unity. This corresponds to the three basic divisions of the Hegelian system: the *Logic*, the *Philosophy of Nature* and the *Philosophy of Mind.* It is free because self-determined. Its freedom consists in recognizing that its alienation into its other (nature) is but a free expression of itself. The truth is the whole and it unfolds through a dialectical progression of categories, concepts and forms of consciousness from the most simple and empty to the most complex and concrete. Each category reveals itself to the observer to be incomplete, lacking and contradictory; it thus passes over into a more adequate category capable of resolving the one-sided and contradictory aspects of its predecessor, though throwing up new contradictions in its turn. Against Kant, this process of dialectical reason is not concerned with the transcendent, but is immanent in reality itself. Reflective understanding is not false, but partial. It abstracts from reality and decomposes objects into their elements. Analytic understanding represents a localized standpoint which sets up an unsurpassable barrier between subject and object and thus cannot grasp the systematic interconnection between things or the total process of which it is a part. The absolute subject contains both itself and its other (both being and thought) which is revealed to be identical with itself. Human history, human thought are vehicles through which the absolute achieves self-consciousness, but humanity as such is not the subject of the process. Thus the absolute spirit dwells in human activity without being reducible to it, just as the categories of the *Logic* precede their embodiment in nature and history.

The character of the Marxian dialectic is yet harder to pin down than that of Hegel. In some well-known lines in the Post-Face to the Second Edition of *Capital* in 1873, Marx stated,

> I criticised the mystificatory side of the Hegelian dialectic nearly thirty years ago . . . [but] the mystification which the dialectic suffers in Hegel's hands by no means prevents him from being the first to present its general form of motion in a comprehensive and conscious manner. With him it is standing on its head. It must be inverted in order to discover the rational kernel within the mystical shell. (Marx 1873, pp. 102–3)

This statement has satisfied practically no one. How can a dialectic be inverted? How can a rational kernel be extracted from a mystical shell? To critics from empiricist, positivist or structuralist traditions, anxious to free Marx from the clutches of Hegelianism, the dialectic is intrinsically unworkable and must either be dropped or stated in quite other terms (for example, Bernstein 1899; Della Volpe 1950; Althusser 1965; Cohen 1978; Elster 1985). To a second group, the dialectical understanding of capitalism is only a particular instance of more general dialectical laws which govern reality as a whole, both natural and social (Engels, dialectical materialism). To a third group, the Hegelian roots of Marx's thought are not sufficiently emphasized in this statement; Marxism is only Hegelianism taken to its logical revolutionary conclusions in the discovery of the proletariat as the subject–object of history and the 'totality' as the distinguishing feature of its world-outlook (Lukács 1923 and much of 20th-century Western Marxism). This *Methodenstreit* cannot be discussed here. All that can be attempted is to give some sense to Marx's statement and in particular to indicate how it informed his critique of political economy.

Marx specifically criticized 'the mystificatory side of the Hegelian dialectic' in his 1843 *Critique of Hegel's Philosophy of Right* and in the

concluding section of the *1844 Manuscripts* (both of which were only published in the 20th century). In these texts, Marx followed Feuerbach in considering Hegelian philosophy to be the conceptual equivalent of Christian theology; both were forms of alienation of man's species attributes; Christianity transposed human emotion into a religious Godhead, while Hegel projected human thinking into a fictive subject, the Absolute Idea, which in turn then supposedly generated the empirical world. Employing Feuerbach's 'transformative method' (the origin of the inversion metaphor), subject and predicate were reversed and hence the correct starting point of philosophy was the finite, man. Nature similarly was not the alienated expression of Absolute Spirit, it was irreducibly distinct. Thus there could be no speculative identity of being and thought. Man, however, as a natural being, could interact harmoniously with nature, his inorganic body. Once the absolute spirit had been dismantled and the identity of being and thought eliminated, it could be argued that the barrier against the harmonious interpenetration of man and nature and the free expression of human nature, was not 'objectification', the division between subject and object constitutive of the finite human condition, but rather the inhuman alienation of man's species life activity in property, religion and the state. True Communism, humanism, meant the re-appropriation of man's essential powers, the generic use of his conscious life activity. In contrast to the predominant Young Hegelian position, therefore, which counterposed Hegel's revolutionary 'method' (the dialectic) to his 'conservative system', Marx argued that there was no incompatibility between the two. For while Hegel's dialectic ostensibly negated the empirical world, it covertly depended upon it. Not only was the moment of contradiction a prelude to the higher moment of reconciliation and the restoration of identity, but the ideas themselves were tacitly drawn from untheorized experience. The effect of the dialectical chain which embodied the world was not to subvert the existing state of affairs, but to sanctify it.

In the crucial period that followed, that of the *German Ideology* and the *Poverty of Philosophy*, in which the basic architecture of the 'materialist conception of history' was elaborated, the attack upon speculative idealism was made more radical. The generic notion of 'conscious life activity', 'praxis', was replaced by the more specific notion of production. Hegel and the Idealist tradition were given credit for emphasizing the active transformative side of human history, but castigated for recognizing this activity only in the form of thought. Thought itself was now made a wholly derivative activity. The fundamental activity was labour and what developed in history were the productive powers men employed in their interaction with nature, 'the productive forces'. Stages in the development of these productive forces were accompanied by successive 'forms of human intercourse', what became 'the relations of production'. Finally, 'man' as a generic being was dispersed into the struggle between different classes of men, between those who produced and those who owned and controlled the means of production.

In this new theorization of history, explicit references to Hegel were few and the dialectic scarcely mentioned. But Hegel re-entered the story as soon as Marx attempted to write up a systematic theory of the capitalist mode of production in 1857–8. To see why, we must briefly survey his economic writings up to that date.

Marx's 1843 critique of Hegel had led him to the conclusion that civil society was the foundation of the state and that the anatomy of civil society was to be found in political economy. However, if his preoccupation with political economy dated from this point, it was not that of an economist. In the 1844 Manuscripts what is to be found is a humanist critique of both political economy and civil society: not an alternative theory of the economy, but rather a juxtaposition between the 'economic' and the 'human', the former being judged in terms of the latter. No distinction is made between political economy and the economic reality it purports to address, the one is simply seen as the mirror of the other.

The first attempt to define capitalism as an economic phenomenon occurred in the *Poverty of Philosophy* (1847). However, whatever the significance of that work in other respects, it did not outline any specifically Marxian portrayal of the

capitalist economy. As in 1844 there was no internal critique of classical political economy. The main difference was that, whereas in 1844 Marx saw that economy through the eyes of Adam Smith, he now saw it through the eyes of Ricardo. In particular, he adopted what he took to be Ricardo's theory of value and belaboured Proudhon for positing as an ideal – the equivalence of value and price – what he considered to be the actual situation under capitalism. The only critique of Ricardo to be found there was a purely external historicist one: that Ricardo was the scientific expression of the epoch of capitalist triumph, but that that epoch had already passed away, that its gravediggers had already appeared and that its collapse was already at hand.

When Marx resumed his economic studies after the 1848 revolutions, Proudhonism was still the main object of attack. It occupied a major part of his unfinished economic manuscripts of 1850–1 and the attack on the Proudhonist banking schemes of Darimon took up the first part of the written-up notebooks of 1857–8, the *Grundrisse.* Proudhonism was the main object of attack because it could be taken for the predominant form of socialist or radical reasoning about the economy. Ricardo could again be utilized to attack such reasoning in order to argue that it represented a nostalgia for petty commodity production under conditions of equal exchange, a situation supposedly preceding modern capitalism rather than representing an emancipation from it. However, if the capitalist mode of production and its historical limits were to be grasped in theory, this would have to involve a critique of Ricardo himself.

The form this critique took, involved problematizing Ricardo's theory of value (or rather Marx's reading of it). Steedman (1979) has argued strongly that Marx misconstrued Ricardo's theory, though Ricardo's shifting of position between the three editions of the *Principles* and the fact that Marx only used the third edition makes his mistake an understandable one). On the one hand, it raised a question never posed by Ricardo: the source of profit in a system of equal exchange. On the other hand, it involved juxtaposing wealth in the form of productive forces, that is, as a collection of use values against the translation of all wealth into exchange values within capitalism. Ricardo, it was argued, possessed no criterion for distinguishing between the content – or the material elements – and the form of the economy, such as Marx possessed in the distinction between forces and relations of production. Ricardo never problematized the 'value form'; he linked the object of measurement with the measurement itself. For this reason, Ricardo was considered to possess no conception of the historicity of capitalism. Once the material could be distinguished from the social, the content from the form, the capitalist mode of production could be conceived as a dynamic system whose principle of movement could be located in the contradictory relationship between matter and form.

It is here that Hegel came in. We know that during the writing of the *Grundrisse* at the beginning of 1858, Marx re-read Hegel, in particular the *Science of Logic.* He wrote to Engels, 'I am getting some nice developments, e.g. I have overthrown the entire doctrine of profit as previously conceived. In the method of working, it was of great service to me that by mere accident I leafed through Hegel's *Logic* again' (Marx to Engels, 16 January 1858).

What Marx found so useful in his reading of Hegel's *Logic* at this time is not really mysterious. It suggested a way of elaborating the contradictory elements that Marx had discerned in the value form into a theorization of the trajectory of the capitalist mode of production as a whole. The point is emphasized by Marx in his Post-Face to *Capital*: the dialectic includes in its positive understanding of what exists a simultaneous recognition of its negation, its inevitable destruction; because it regards every historically developed form as being in a fluid state, in motion, and therefore grasps its transient aspect as well (1873, p. 103). The dialectic offered a means of grasping a structure in movement, a process – the subtitle of *Capital*, volume 1, was 'the process of capitalist production'. If capitalism could be represented as a process and not just a structure, then concomitantly its building blocks were not factors, but, as in Hegel, 'moments'. As Marx put it in the *Grundrisse*:

> When we consider bourgeois society in the long view and as a whole, then the final result of the process of social production always appears as the society itself i.e. the human being itself in its social relations. Everything that has a fixed form, such as the product etc., appears as merely a moment, a vanishing moment in this movement. The conditions and objectifications of the process are themselves equally moments of it, and its only subjects are the individuals, but individuals in mutual relationships, which they equally reproduce and produce anew. . .. in which they renew themselves even as they renew the world of wealth they create. (Marx 1857–8, p. 712)

Marx's attempt to utilize the *Logic* can be seen most clearly in the *Grundrisse.* There one can see the genesis of particular concepts which in *Capital* appear in more polished form. What is clear is that the *Logic* is used as a first means of setting terms in relation to each other. The text is littered with Hegelian expressions and turns of phrase; indeed, sometimes it appears as if lumps of Hegelian ratiocination have simply been transposed, undigested, to sketch the more intractable links in the chain. Here, for instance, is money striving to become capital: '. . . already for that reason, value which insists on itself as value preserves itself through increase; and it preserves itself precisely only by constantly driving beyond its quantitative barrier, which contradicts its character as form, its inner generality' (p. 270). But at the same time we can see Marx remind himself to correct the 'idealist manner of presentation, which makes it seem as if it were merely a matter of conceptual determination and of the dialectic of these concepts' (p. 151).

But the interest of dialectical logic for Marx was not simply that it offered him a way of outlining a structure in movement; more fundamentally it enabled him to depict contradiction as the motor of this movement. This was why the dialectic was 'in its very essence critical and revolutionary' (Marx 1873, p. 103), in that both in Hegel and in ancient Greek usage movement was contradiction. This appears closely in the dramatic relationship that Marx sets up between the circulation system and the production system in *Capital.* The system of exchange of the market is the public face of capitalism. It is 'in fact a very Eden of the innate rights of Man' (p. 280). Exchanges are equal. To look for the source of inequality in the exchange system, like the Proudhonists, is to look in the wrong place. Yet, if exchanges are equal, how does capital accumulation take place? Equal exchange implies the principle of identity, of non-contradiction. It is, in Hegel's sense, the sphere of 'simple immediacy', the world as it first appears to the senses. It cannot move or develop, because it apparently contains no contradictory relations.

But this surface of things is not self-sufficient. It is 'the phenomenon of a process taking place behind it'. As a surface it is not nothing, but rather a boundary or limit. Contradiction and therefore movement is located in production. Here there is non-identity, the extraction of surplus labour disguised by the surface value form and its tendency to limitless expansion.

Thus, there are two processes, on the one hand that of the surface, that of immediate identity lacking the motive power of its own regeneration; on the other hand, that beneath the surface, a process of contradiction. Thus in Hegelian terms, the whole could then be defined as 'the identity of identity and non-identity'. In this whole, contradiction is the overriding moment, but the surface places increasingly formidable obstacles to its development, for instance, so-called 'realization' crises. Values can only be realized in an act of exchange and the medium of this exchange is money. But there is no guarantee that these exchanges must take place. The 'anarchy' of the market place is such that overproduction or disproportionality between sectors of production can only be seen after the event. Hence trade crises and slumps (see M. Nicolaus, Introduction to Marx 1857–8).

This is only one example of how Marx employed dialectical principles in his attempt to conceptualize the process or movement of a contradictory whole. Another would be the six books Marx originally planned to write in 1857–8, the original blueprint of *Capital.* Their order would have been: Capital, Wage Labour, Landed Property, State, World Market, Crises. This plan is reminiscent of Hegel's *Encyclopaedia.* It describes a circle in a Hegelian sense. The point of departure is not capital per se, but commercial exchange as appearance, then proceeding through the contradictory world of production and

eventually returning to commercial exchange again as the world market, but this time enriched by the whole of the preceding analysis.

There has been much controversy about the proximity or distance between the Hegelian and Marxian dialectics. Those who like Althusser (1965) argue for their radical dissimilarity, are on their strongest ground when arguing that in Marx the terms of the dialectic have been radically transformed. The contradiction between forces and relations of production cannot be reduced to the ultimate simplicity of that between Hegel's master and slave or of that between proletariat and bourgeoisie in the Hegelianized Marxist account of Lukacs. But it is far more difficult to establish as unambiguously the difference in the relationship between the terms in their respective dialectics. On the one hand, the relation between matter and form in Hegel is only one of apparent exteriority. Matter relates to form as other only because form is not yet posited within it. Once the terms are related, they are declared to be identical. Marx, on the other hand, insists upon the irreducible difference between matter and form, between the material and the social (even if he is not wholly successful in keeping them apart). Not only are matter and form different, but the one determines the other: value is determined in relation to the material production of use value; the opposite is not true. Relations of determination would seem to exclude identity, and this is confirmed by Marx's avoidance of the Hegelian notion of 'sublation' (*Aufhebung*), the higher moment of synthesis. The dialectical clash between forces and relations of production in the capitalist mode of production does not of itself produce a higher unity (socialism); rather what crises do, is to make manifest the otherwise hidden determination of value by use value, of form by matter. Against this, however, must be set one or two passages, including a famous peroration in *Capital* volume 1, where Marx does conceive the end of capitalism as a return to a higher but differentiated unity and does employ the notion of the negation of the negation (Marx 1873, p. 929), and, despite the best efforts of some modern commentators, it is difficult honestly to deny the strongly teleological imagination which underpins the whole enterprise of *Capital.*

Finally, in two important respects, Hegelian dialectic, however surreal, is less vulnerable than that of Marx. Firstly, Hegel's *Science of Logic* takes place outside spatio-temporal constraints. It is a purely logical progression of concepts, even if the principles on which one ontological category is derived from another 'have resisted analysis to this day' (Elster 1985, p. 37). Marx's effort to avoid giving any impression of the 'self-determination' of the concept, took the form of attempting to demonstrate that 'the ideal is nothing but the material world reflected in the mind of man and translated into forms of thought' (Marx 1873, p. 102). In practical terms this implied that there was some systematic relationship between the logical sequence of concepts in the exposition of the argument and the chronological order of their appearance in historical time. But this turned out to impose insurmountable difficulties in terms of presentation (and it is significant that, having begun with the product in the *Grundrisse*, he began with the commodity in *Capital*). Thus Marx both stated his position and violated it, bequeathing insoluble ambiguities surrounding his interpretation of value, of the meaning of 'reflection' and of the relationship between history and logic which have plagued even his closest followers ever since. Secondly, when it came to applying his dialectic to history, Hegel was categorical in refusing to project his theory into the future. The philosopher could explain the rationality of what had happened; it was only then that it could be grasped in thought. Marx, despite all his strictures against the voluntarism of other Young Hegelians and some of his fellow revolutionaries, was unable by the very nature of his project, fully to abide by the Hegelian restriction. Thus, while Hegel's owl of Minerva flew at dusk, the Marxian owl, unfortunately, took flight at high noon.

Bibliography

Althusser, L. 1965. *For Marx*. London: Allen Lane. 1969.
Bernstein, E. 1899. *Evolutionary socialism*. Stuttgart. English trans. E.C. Harvey. London: Independent Labour Party. 1909.
Bhaskar, R. 1983. *Dialectic, materialism and human emancipation*. London: New Left Books.

Cohen, G. 1978. *Karl Marx's theory of history: A defence*. London: Oxford University Press.
Della Volpe, G. 1950. *Logica come scienza positiva*. Messina: G. d'Anna.
Elster, J. 1985. *Making sense of Marx*. Cambridge: Cambridge University Press.
Hegel, G.W.F. 1812–1816. *The science of logic*. London: Allen & Unwin. 1961.
Kolakowski, L. 1978. *Main currents of Marxism Vol. I: The founders*. Oxford: Oxford University Press.
Lukács, G. 1923. *History and class consciousness*. London: Merlin. 1971.
Marx, K. 1844. Economic and philosophical manuscripts of 1844. In *Collected works*, ed. K. Marx and F. Engels, vol. 3. London: Lawrence & Wishart. 1975.
Marx, K. 1847. The poverty of philosophy. In: *Collected works*. vol. 6. London: Lawrence & Wishart. 1976a.
Marx, K. 1857–8. *Grundrisse*. Harmondsworth: Penguin. 1973.
Marx, K. 1873. *Capital*. Vol. I. 2nd ed. Harmondsworth: Penguin. 1976b.
Rosdolsky, R. 1968. *The making of Marx's capital*. Trans. P. Burgeis. London: Pluto Press. 1977.
Steedman, I. 1979. Marx on Ricardo. Discussion Paper No. 10, Department of Economics, University of Manchester.

Diamond, Peter (Born 1940)

Nicholas Barr

Abstract

Peter Diamond has made fundamental contributions to economic theory over a wide range of areas including search theory and its implications for unemployment (for which he was awarded the Nobel Memorial Prize), optimal taxation, which he pioneered with James Mirrlees, incomplete markets and their implications *inter alia* for the design of social insurance, and the role of government debt in fostering intergenerational efficiency. For most of his career, he has also been active in policy analysis, notably pensions.

Keywords

Behavioural economics; DMP model; Duality theory; Frictions; Incomplete markets; Matching; McFadden; Mirrlees; National debt; Optimal taxation; Overlapping generations; Pensions; Ramsey rule; Samuelson; Solow; Search theory; Social security; Use of models

JEL Classifications

B31

Personal and Career

Peter Arthur Diamond was born in 1940 in New York City and educated at public school, initially in the Bronx, but mostly in Long Island. He graduated *summa cum laude* from Yale in 1960 with a major in mathematics, including courses in economics, one of them a graduate course with Gerard Debreu. His PhD in economics at MIT, supervised by Robert Solow, was awarded in 1963. He was assistant professor, then acting associate professor at the University of California Berkeley from 1963 to 1966, before returning to MIT, where he spent the rest of his career, becoming Institute Professor in 1997.

Among many honours, he was a co-recipient of the Nobel Memorial Prize in 2010 and winner of the Mahalanobis Memorial Award (1980), the Nemmers Prize (1994), the Killian Award, MIT (2003–04), the Samuelson Award from TIAA-CREF (2003), the Jean-Jacques Laffont Prize (2005) and the Robert M. Ball Award (2008). He is a Fellow of the Econometric Society (1968) and the American Academy of Arts and Sciences (1978), and a Member of the National Academy of Sciences (1984).

Though this article is mainly about his professional contributions, no biography would be complete if it left out the personal side. He grew up in a very supportive family and met his lifelong partner, Priscilla (Kate) Myrick while at Berkeley. Their marriage and family life has been central. Over the years, the probability of his agreeing to travel was significantly higher if it took him to where one of their sons was living. A memorable conference in 2010 to celebrate his contributions to economics and to MIT showed not only his academic reach

but also the warmth of the friendships he inspired across all age groups and over many years. Alongside family, friends and economics, his other passion is the Boston Red Sox. Bringing all these together to celebrate his 70th birthday, Kate Diamond arranged for him to throw the first pitch at Fenway Park, with family, friends and MIT colleagues and students in the stands (a photo of the pitch, and one of the shirts worn by the graduate students at that game are in the Nobel Museum).

Research Contributions

Though Diamond and his co-recipients were awarded the 2010 Nobel Memorial Prize for their work on markets with search frictions, the citation (Economic Sciences Prize Committee of the Royal Swedish Academy of Sciences 2010) also makes clear Diamond's fundamental contributions to other areas, including optimal taxation, incomplete markets and intergenerational market inefficiencies. The discussion below is organised round those headings, though the list is far from complete, for example, omitting research on behavioural economics (Diamond et al. 1997; Diamond and Koszegi 2003; Diamond and Vartiainen 2007; Diamond 2008) and recent work on taxation (Banks and Diamond 2010; Diamond and Saez 2011; Diamond and Spinnewijn 2011).

Search Theory

Diamond's work on the theory of search frictions – impediments to trading (finding a job, buying a house) that arise from the need to discover price or to assess quality – was an early foray into a continuing interest on the impact of time on economic behaviour (Diamond 1994a).

First-best economic theory assumes that buyers and sellers are perfectly informed about both price and quality and hence find each other instantly and costlessly. In this setting it makes no difference that trade is with 'the market', not an identified trading partner. Diamond's (1971) starting point was to relax the assumption of perfect information about price by considering a situation where the only way a consumer can find the price of a homogeneous good is to go to the shop and ask. Once in the first shop, going to another shop to compare prices has a cost in terms of time. Thus the first shop has an element of market power and could in principle charge a slightly higher price than the second shop.

That much is obvious. What was surprising was that the market equilibrium occurs not at a price slightly above the competitive price, but where all shops charge the monopoly price, a result that holds however small the friction. The intuition is that each shop will want to exploit its element of monopoly power by raising its price slightly; however, each shop knows that all the other shops will do the same, and thus raises its price a bit more. The equilibrium is where all shops charge the monopoly price.

The broader conclusion is that

> Small search costs can have a large impact because price setters respond to the prices set by others, so there is a feedback process that greatly expands the impact of search costs – as each firm reacts both to the presence of the search costs of potential customers and to the responses of other suppliers to the same search costs. That a small amount of friction could create a large change. . . served as a marker of the importance of the study of equilibrium with frictions. (Diamond 2011a, p. 1048)

This result, known as the Diamond Paradox, attracted considerable attention and was a major initiator of research on markets with search costs, where participants in a market look for partners for mutually agreed trades. This may involve simple cases of a buyer and a seller of a homogeneous product, or more complex relations between heterogeneous employers and job seekers, or between firms and their suppliers. The models embrace variations in demand curves across buyers and costs across sellers, and different behaviours of sellers.

Though Diamond's 1971 paper considered one-sided search and prices set on a take-it-or-leave-it basis, it broke new ground as a fully worked out equilibrium model. To move from there to analysis of the labour market required a model with two-sided search, wage bargaining, and a process of matching heterogeneous workers to

heterogeneous jobs. Diamond's later work (1981, 1982a) was part of this development. So were papers by his fellow Nobel Laureates, Dale Mortensen and Christopher Pissarides. Mortensen's (2011) Nobel Prize lecture flags the key papers in the intellectual process leading to the Diamond–Mortensen–Pissarides (DMP) model, which has become the standard approach to analysing unemployment, job vacancies and wage determination, and in particular to how they are affected by macroeconomic policy, the design of unemployment benefits, and labour market regulation, such as rules about hiring and firing. Since improving the match of heterogeneous workers to heterogeneous jobs increases the efficiency of labour markets, an important conclusion is that the duration of unemployment should be optimised, not minimised, in order to allow the efficient amount of search activity.

Alongside the 1981 and 1982a papers already mentioned, Diamond's later work on search equilibrium included writing on macro aspects of the labour market (Diamond and Blanchard 1989, 1990a, b, 1992, 1994) and aggregate demand management (Diamond 1982b, 1984; Diamond and Fudenberg 1989). A key finding is the possibility that, even without any price or wage 'stickiness', markets with search costs may have multiple equilibria, for example, associated with different levels of employment. An implication is the role of government in seeking to move the economy towards the best outcome.

The applications of search theory by a range of other writers extend well beyond the labour market. The number of houses for sale varies over time and, connected, so does the time to find a buyer and to agree a price. Search theory has also been used to study aspects of monetary theory, public economics, financial economics, regional economics and family economics.

Optimal Taxation

A second area to which Diamond made a fundamental contribution is optimal taxation – the design of taxes to minimise inefficiency or jointly to optimise efficiency and distribution.

It is a standard proposition that lump-sum taxation is non-distortionary, since individuals cannot avoid it by changing their behaviour. Thus lump-sum taxes and transfers are compatible with the Second Fundamental Theorem of Welfare Economics, which shows that in a first-best economy any Pareto efficient allocation can be sustained as a competitive equilibrium by establishing a suitable set of initial endowments.

In practice, however, policy needs to provide poverty relief. Thus at least some taxes need to be related to income or consumption, and, by changing behaviour, will be distortionary in that they move outcomes away from the optimum. A tax on earnings is likely to alter labour supply by reducing the return to additional hours of work. A tax on consumption may also affect labour supply. And a tax on interest income may change savings behaviour by reducing the return to additional saving.

Optimal taxation considers how taxes should be designed so as to optimise a social welfare function. In the setting of a single representative household, this is equivalent to minimising the deadweight loss resulting from the need to raise a given revenue when lump-sum taxes are not an option. The 'Ramsey Rule' (Ramsey 1927) is that, provided goods are unrelated in consumption, deadweight losses are minimised where commodity tax rates are inversely proportional to the price elasticity of demand for each commodity.

Diamond's best-known contributions to the optimal taxation literature are the production efficiency result of Diamond and Mirrlees (1971a) and the extension of the Ramsey rule to include distributional concerns (Diamond and Mirrlees 1971b; Diamond 1975). These analyses were among the first to be framed in terms of duality, whose properties Daniel McFadden was exploring at Berkeley in the mid-1960s (Fuss and McFadden 1978), including work with Diamond (leading *inter alia* to Diamond and McFadden 1974). Not unusually for pathbreaking research, the 1971 Diamond–Mirrlees papers had a tortuous route from presentation as a single paper at the Econometric Society meetings in 1967 to eventual publication, in part because the editor of the *AER* insisted on splitting the paper into two to limit the extent of a single issue devoted to a single paper.

The 1971 papers consider an economy where firms face constant returns to scale and with no distortions apart from taxation (i.e. setting to one side problems like externalities). The papers established a series of results. First, they recast the Ramsey conclusion in terms of quantities: an optimum commodity tax requires an equi-proportionate reduction in the compensated demand for all commodities.

The second result, known as the production efficiency theorem, is that taxes on intermediate goods should be nondistortive of production, for example justifying a value-added tax regime. This finding revolutionised the analysis of commodity taxation. On the face of it, the result is surprising given the Lipsey and Lancaster (1956) conclusion that a distortion in one part of the economy should generally be offset by other distortions elsewhere. However, in a competitive economy where firms make zero profits, the availability of a complete set of taxes on transactions between households and firms dominates taxes on firms that move production inside the frontier, which latter are therefore never part of the optimum.

The simple Ramsey result assumed that all individuals are identical and hence excluded distributional concerns. Thus the inverse elasticity rule has the sole aim of minimising efficiency losses. Diamond and Mirrlees (1971b) and Diamond (1975) extend the Ramsey rule to include distributional issues. Suppose that there are two goods: a necessity, which absorbs a larger fraction of the income of the poor than the rich, and a luxury, which absorbs a larger fraction of the income of the rich than the poor. If the social welfare function gives greater weight to the marginal utility of the poor than the rich, then even if the demand for the necessity is price-inelastic, the social optimum might require a higher tax rate on the luxury. The deviation from the simple Ramsey rule is greater (a) the greater the concern with the utility of the poor (if we do not care about distribution, the simple rule applies) and (b) the greater the difference in the consumption patterns of rich and poor (if rich and poor have identical consumption patterns, there is no distributional gain from imposing a higher tax rate on the luxury).

Incomplete Markets

Search models address one form of imperfect information, for example about price or available jobs. Another line of inquiry relates to imperfect information in the form of uncertainty, which leads to missing markets, and in particular to missing insurance markets. In contrast, a complete set of markets would include insurance against all possible future contingencies. Diamond (1967) was a pioneering exploration of economies with some missing insurance markets, creating deviations from optimality which government intervention might be able to improve. A central, and highly influential, part of his argument is that governments also face limited abilities to address uncertainty. Thus the right comparison is not between state of the world A, with private allocations with missing markets, and state of the world B, where government completes the market structure. Instead, the comparison is between imperfect market outcomes and imperfect government intervention, since governments are also imperfectly informed. Following Diamond's 1967 paper, a large literature on incomplete markets developed, including his own later work (for example, Diamond and Mirrlees (1978) on social insurance) and a range of papers including Stiglitz (1972), Hart (1975) and Grossman and Hart (1979). The state of the art is the theorem in Geanakoplos and Polemarchakis (1986) and the text by Magill and Quinzii (1996).

Diamond's analysis has profound implications. Where insurance is absent or incomplete – for example against adverse labour market outcomes such as unemployment and low earnings – a second-best optimum will include both insurance and redistribution (for example, the redistributive tilt in the US social security pension), with distortionary effects on labour supply and saving. In a second-best economy, strict adherence to actuarial principles is suboptimal.

Government Debt and Intergenerational Efficiency

A fourth strand of Diamond's work analyses the influence of government debt on market outcomes

and consumer welfare. The incompleteness in this case arises because current generations and future generations cannot participate directly in the same market, referred to as incomplete participation – for example, future generations cannot influence the amount of debt built up by earlier generations. As a result, market outcomes may not be efficient.

Though not his first paper, Diamond (1965) was the one that first drew his work to wide attention in the profession. The paper develops an overlapping generations model, as in Samuelson (1958) (and, as rediscovered later, Allais (1947) – see Malinvaud 1987), with capital accumulation set in the Solow growth model to show that appropriate levels of government debt can be welfare-enhancing.

> His analysis of debt and its potentially welfare-improving consequences has had profound impact on the profession. His formulation of the overlapping-generations model, which combines the Solow–Swan growth model with an overlapping-generations population structure, still constitutes the benchmark model of government debt, social security, and intergenerational redistribution. (Economic Sciences Prize Committee of the Royal Swedish Academy of Sciences 2010, pp. 26–27)

Policy Analysis

Alongside basic research, Diamond has been involved in policy analysis from early in his career, his work on pensions being the best known (Diamond 2002, 2003; Barr and Diamond 2008). Some of that work was part of a continuing contribution to the US domestic debate about social security reform (Diamond et al. 1996; Diamond 1998, 1999a, 2004; Diamond and Orszag 2005). Other writing was sparked by international debate, including Chile (Diamond 1994b; Diamond and Valdes-Prieto 1994), China (Asher et al. 2005; Barr and Diamond 2010), and a range of other countries (Diamond 1999b, 2000, 2001, 2006).

The work on pensions was deeply rooted in his theoretical work and his social concerns. A central argument is

> . . .that any optimal program will and should generate distortions because, starting from laissez-faire, distortions generate second-order efficiency costs but first-order redistributive benefits. (review of Diamond (2003) by Emmanuel Saez, *Journal of Economic Literature*, June 2004, p. 530)

Thus,

> The desirability of insurance against adverse labor market outcomes, particularly toward the end of a career, calls for deviations from actuarial benefits to provide better insurance protection. That is, in the presence of asymmetric information, optimal insurance inevitably distorts choices. The information asymmetry comes from the fact that low labor market participation may be either voluntary (preference for more leisure) or involuntary (low pay or no work available), and only the worker knows which is the case. (Barr and Diamond 2008, p. 65)

The Wider Context

Broader Professional Contributions

Diamond was President of the Econometric Society and the American Economic Association, and a founding member and later President and Chair of the Board of the National Academy of Social Insurance. He has had editorial roles for the *Journal of Economic Theory*, the *Journal of Public Economics* and the *American Economic Review*, and was a member of the Committee on Science, Engineering, and Public Policy of the National Academy of Sciences.

His involvement in US public policy includes work for the US Congress, including the Senate Finance Committee (1974–75) and the Consultant Panel on Social Security of the Congressional Research Service (1975–76), and chairing the Panel on the Privatization of Social Security for the National Academy of Social Insurance.

In April 2010 President Obama nominated Diamond to be a Governor of the US Federal Reserve. The nomination became mired in political wrangling and, to the consternation of many in the USA and baffled incomprehension elsewhere, positions became entrenched after the Nobel Prize Committee's announcement later that year. To widespread regret, Diamond withdrew in June 2011 (Diamond 2011b).

Alongside these national roles were a range of public service activities in Massachusetts, and international roles advising governments on pension reform, including China in 2004 and 2009 and Sweden in 2010.

His connection with the economics department at MIT was central to both. Apart from sabbaticals and the standard MIT requirement to teach elsewhere for at least three years after completing his PhD, he arrived as a graduate student in 1960 and never left. The biographies of Paul Samuelson and Robert Solow in the *New Palgrave* make clear their central roles in establishing the modern department and in setting a tone of warmth, rigour and intellectual tolerance. Diamond fitted naturally into that environment and became a central figure in continuing the tradition. It was therefore fitting that he should have been the first holder of the Paul Samuelson Chair in the department from 1992 to 1997 before following Samuelson and Solow in becoming an Institute Professor, the highest award MIT can bestow.

Approach to Economics

Diamond's approach to economics should be seen as part of the wider person. The initial inspiration for some of his research grew out of teaching. His 1965 AER paper on national debt began life as a handout for the undergraduate course in public finance at Berkeley. Similarly, the starting point for his work on optimal taxation was in the classroom.

> So, I'm teaching how to measure dead weight burden using the expenditure function in the public finance class at MIT and I'm literally at the blackboard thinking: I could minimize this! So I finished class, went back to my office and reinvented Ramsey, because with a dual approach it is a piece of cake. (Moscarini and Wright 2007, p. 549)

Diamond concluded that 'While many are concerned about the tension between teaching and research, my experience is that they reinforce each other' (Nobel Prize autobiography: Diamond 2011c, p. 306). Given the MIT tradition of intellectual tolerance, it is not surprising that the influence of teaching on research has been eclectic.

Diamond's talent as a teacher is for providing a master class for the very best students rather than breaking the intellectual ice to open up topics for a more typical student. At root, he is a truth teller, not a story teller – giving the whole picture all the time, rather than simplifying to get across the gist of an idea, introducing the finer points only later. His talent as a teacher was illustrated beautifully by Amy Finkelstein at the 2010 conference mentioned earlier to celebrate Diamond's contributions to economics and to MIT. She reflected on going to see Diamond as a PhD student. He spoke and she took notes, without fully understanding. Three months later, the penny dropped. She said that she called such occasions her 'Peter moments', and judged her professional progress by the extent to which they came faster over time.

Alongside the interaction between teaching and research was an interaction between research and policy analysis. On the dust jacket of the *Economics of Welfare*, Pigou wrote

> When a man sets out upon any course of inquiry, the object of his search may be either light or fruit – either knowledge for its own sake or knowledge for the sake of good things to which it leads... [T]here will, I think, be general agreement that in sciences of human society... it is the promise of fruit and not of light that chiefly merits our regard. . .. (1920; quote from 4th edition, 1932)

For Diamond, light and fruit were intertwined. 'For me, policy analysis and basic research are mutually supportive. Policy discussions have alerted me to interesting research questions that had not received adequate analysis. And my policy analysis draws heavily on my understanding of economic theory and reading in the empirical literature' (Nobel Prize autobiography: Diamond 2011c, p. 309).

As one example, a series of papers with Mirrlees, starting with Diamond and Mirrlees (1978) on how the level of pensions should vary with the age at which benefits start, came from thinking about the question as a member of the Consultant Panel on Social Security of the Congressional Research Service, 1975–76. More generally, part of the motivation for Diamond's work on incomplete insurance markets grew out of an interest in the optimal design of a benefit system.

The interest in policy also led to reflections about methodology, and in particular about the importance of basing policy recommendations on multiple models, each of them shedding light on a particular aspect of the problem, and from other sources of insight, rather than stretching a single model beyond its useful range. Thus, for example, the design of pensions needs to take account of incomplete insurance markets, search frictions and lessons from behavioural economics (e.g. when comparing different pension policies offered by different providers).

Diamond's insistence that no model is complete is best expressed – as so often – in his own words.

> The complexity of the economy calls for the use of multiple models that address different aspects. . .. I am concerned that... too many economists take the findings of individual studies literally as a basis for policy thinking, rather than drawing inferences from an individual study, and combining them with inferences from other studies that consider other aspects of a policy question, as well as with intuitions about aspects of policy that have not been formally modeled. Assumptions that are satisfactory for basic research, for clarifying an issue by isolating it from other effects, should not play a central role in policy recommendations if those assumptions do not apply to the world. To me, taking a model literally is not taking a model seriously. It is worth remembering that models are incomplete – indeed, that is what it means to be a model. (Diamond 2011a, pp. 1045–6)

> Understanding of the economy, and policy recommendations and decisions, should reflect analysis through multiple models. And they should incorporate insights that seem right even though they have not yet been modeled. (*ibid*., p. 1070)

See Also

- Labour Market Search
- Mirrlees, James (Born 1936)
- Mortensen, Dale T. (Born 1939)
- Optimal Taxation
- Pensions
- Pissarides, Christopher (Born 1948)
- Ramsey Model
- Samuelson, Paul Anthony (1915–2009)
- Search Models of Unemployment
- Search Theory
- Search Theory (New Perspectives)
- Solow, Robert (Born 1924)

Selected Works

Asher, M., N. Barr, P. Diamond, E. Lim, and J. Mirrlees. 2005. Social security reform in China: Issues and options. *Policy Study of the China Economic Research and Advisory Programme* (Jan.); http://www.oup.com/us/pdf/social_security_study_2005 (in Chinese, http://www.oup.com/us/pdf/china_social_security_study).

Banks, J., and P.A. Diamond. 2010. The base for direct taxation. In *Dimensions of tax design: The Mirrlees review*, eds. J. Mirrlees, S. Adam, T. Besley, R. Blundell, S. Bond, R. Chote, M. Gammie, P. Johnson, G. Myles and J. Poterba, 548–674. Oxford: Oxford University Press.

Barr, N., and P.A. Diamond. 2008. *Reforming pensions: Principles and policy choices*. New York/Oxford: Oxford University Press. http://www.oxfordscholarship.com/oso/public/content/economicsfinance/9780195311303/toc. html

Barr, N., and P.A. Diamond. 2010. *Pension reform in China: Issues, options and recommendations*. China Economic Research and Advisory Programme, February, http://econ-www.mit.edu/files/6310

Diamond, P.A. 1965. National debt in a neoclassical growth model. *American Economic Review* 55: 1126–1150.

Diamond, P.A. 1967. The role of a stock market in a general equilibrium model with technological uncertainty. *American Economic Review* 57: 759–776.

Diamond, P.A. 1971. A model of price adjustment. *Journal of Economic Theory* 3(2): 156–168.

Diamond, P.A. 1975. A many-person Ramsey tax rule. *Journal of Public Economics* 4: 335–342.

Diamond, P.A. 1981. Mobility costs, frictional unemployment, and efficiency. *Journal of Political Economy* 89(4): 798–812.

Diamond, P.A. 1982a. Wage determination and efficiency in search equilibrium. *Review of Economic Studies* 49(2): 217–227.

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

Diamond, P.A. 1984. *A search-equilibrium approach to the micro foundations of macroeconomics*. Cambridge, MA: MIT Press.

Diamond, P.A. 1994a. *On time: Lectures on models of equilibrium*. Churchill Lectures. Cambridge: Cambridge University Press.

Diamond, P.A. 1994b. Privatization of social security: Lessons from Chile. *Revista de Análisis Económico* 9: 21–23; revised version in Diamond et al. (1996).

Diamond, P.A. 1998. The economics of social security reform. In *Framing the social security debate: Values, politics, and economics*, eds. R.D. Arnold, M.J. Graetz and A.H. Munnell, 38–64. National Academy of Social Insurance, Brookings Institution Press.

Diamond, P.A. (ed.) 1999a. *Issues in privatizing social security, report of an expert panel of the national academy of social insurance*, Cambridge, MA: MIT Press.

Diamond, P.A. 1999b. Social security reform with a focus on Italy. *Rivista Di Politica Economica*, December.

Diamond, P.A. 2000. Social security reform with a focus on Sweden. *Ekonomisk Debatt* 3, May.

Diamond, P.A. 2001. Social security reform with a focus on the Netherlands. *De Economist* 149(1): 81–114.

Diamond, P.A. 2002. *Social security reform*. The 1999 Lindahl Lectures. Oxford: Oxford University Press.

Diamond, P.A. 2003. *Taxation, incomplete markets and social security*. The 2000 Munich Lectures. Cambridge, MA: MIT Press.

Diamond, P.A. 2004. Social security. *American Economic Review* 94(1): 1–24.

Diamond, P.A. 2006. Reforming public pensions in the U.S. and the U.K. *Economic Journal* 116(509): F94–F118.

Diamond, P.A. 2008. Behavioral economics. *Journal of Public Economics* 92(8–9): 1858–1862.

Diamond, P.A. 2011a. Unemployment, vacancies, wages. *American Economic Review* 101: 1045–1072.

Diamond, P.A. 2011b. When a Nobel Prize isn't enough. *New York Times*, 5 June. http://www.nytimes.com/2011/06/06/opinion/06diamond.html?_r=2&hp=&pagewanted=print

Diamond, P.A. 2011c. Autobiography. http://www.nobelprize.org/nobel_prizes/economics/laureates/2010/diamond-autobio.pdf

Diamond, P.A., and O. Blanchard. 1989. The Beveridge curve. *Brookings Papers on Economic Activity* 1: 1–76.

Diamond, P.A., and O. Blanchard. 1990a. The aggregate matching function. In *Growth, productivity, unemployment: Essays to celebrate Bob Solow's birthday*, ed. P. Diamond. Cambridge: MIT Press.

Diamond, P.A., and O. Blanchard. 1990b. The cyclical behavior of the gross flows of U.S. workers. *Brookings Papers on Economic Activity* 2: 85–155.

Diamond, P.A., and O. Blanchard. 1992. The flow approach to labor markets *AER Papers and Proceedings* May, 354–359.

Diamond, P.A., and O. Blanchard. 1994. Ranking, unemployment duration and wages. *Review of Economic Studies* 60: 417–434.

Diamond, P.A., and D. Fudenberg. 1989. Rational expectations business cycles in search equilibrium. *Journal of Political Economy* XCVII: 606–619; correction *Journal of Political Economy* XCIX(1): 218–219; 1991.

Diamond, P.A., and B. Koszegi. 2003. Quasi-hyperbolic discounting and retirement (with Botond Koszegi), *Journal of Public Economics* 87: 1839–1872.

Diamond, P.A., and D. McFadden. 1974. Some uses of the expenditure function in public finance. *Journal of Public Economics* 3: 3–21.

Diamond, P.A., and J.A. Mirrlees. 1971a. Optimal taxation and public production I: Production efficiency. *American Economic Review* 61: 8–27.

Diamond, P.A., and J.A. Mirrlees. 1971b. Optimal taxation and public production II: Tax rules. *American Economic Review* 61: 261–278.

Diamond, P.A., and J.A. Mirrlees. 1978. A model of social insurance with variable retirement. *Journal of Public Economics* 10: 295–336.

Diamond, P.A., and P. Orszag. 2005. *Saving social security: A balanced approach* (revised edition). Washington, DC: Brookings Institution Press.

Diamond, P., and E. Saez. 2011. The case for a progressive tax: From basic research to policy recommendations. *Journal of Economic Perspectives* 25(4): 1–25.

Diamond, P., and J. Spinnewijn. 2011. Capital income taxes with heterogeneous discount rates. *American Economic Journal: Economic Policy* 3: 52–76.

Diamond, P.A., and S. Valdes-Prieto. 1994. Social security reforms. In *The Chilean economy: Policy lessons and challenges*, eds. B. Bosworth, R. Dornbusch, and R. Labán, 257–328. Washington DC: The Brookings Institution.

Diamond, P.A., and H. Vartiainen. (eds.) 2007. *Behavioral economics and its applications*. Princeton: Princeton University Press.

Diamond, P., D. Lindeman, and H. Young. (eds.) 1996. *Social security: What role for the future?* Washington, DC: The Brookings Institution.

Diamond, P.A., E. Shafir, and A. Tversky. 1997. Money illusion. *Quarterly Journal of Economics* 112(2): 341–374.

Bibliography

Allais, M. 1947. *Economie et Intérêt*. Paris: Imprimerie Nationale.

Economic Sciences Prize Committee of the Royal Swedish Academy of Sciences. 2010. Markets with search frictions, scientific background on the Sveriges Riksbank Prize in economic sciences in memory of Alfred Nobel 2010, Stockholm. http://www.nobelprize.org/nobel_prizes/economics/laureates/2010/ecoadv10.pdf

Fuss, M., and D. McFadden. 1978. Cost, revenue, and profit functions. In *Production economics: A dual approach to theory and applications*, vol. 1. Amsterdam: North-Holland.

Geanakoplos, J.D., and H.M. Polemarchakis. 1986. Existence, regularity and constrained suboptimality of competitive allocations when the asset structure is incomplete. In *Uncertainty, information and communication: Essays in honor of K. J. Arrow*, ed. W.P. Hell, R.M. Starr, and D.A. Starrett, vol. 3, 65–95. New York: Cambridge University Press.

Grossman, S., and O. Hart. 1979. A theory of competitive equilibrium in stock market economies. *Econometrica* 47: 293–329.

Hart, O. 1975. On the optimality of equilibrium when the market structure is incomplete. *Journal of Economic Theory* 11: 418–443.

Lipsey, R.G., and K. Lancaster. 1956. The general theory of second best. *Review of Economic Studies* 24: 11–32.

Magill, M., and M. Quinzii. 1996. *Theory of incomplete markets*. Vol. I. Cambridge, MA/London: MIT Press ; reprinted in Magill, M. and Quinzii, M. (eds.) 2008. *Incomplete markets*, Vol. 1: *Finite horizon economies*. International Library of critical writings in economics. Cheltenham, UK: Edward Elgar.

Malinvaud, E. 1987. The overlapping generations model in 1947. *Journal of Economic Literature* 25(1): 103–105.

Mortensen, D.T. 2011. Markets with search friction and the DMP model. *American Economic Review* 101: 1073–1091.

Moscarini, G., and R. Wright. 2007. An interview with Peter Diamond. *Macroeconomic Dynamics* 11: 543–565.

Pigou, A.C. 1920. *The economics of welfare*. 4th ed. London: Macmillan.

Ramsey, F.P. 1927. A contribution to the theory of taxation. *The Economic Journal* 37(145): 47–61.

Samuelson, P.A. 1958. An exact consumption-loan model of interest with or without the social contrivance of money. *Journal of Political Economy* 66(6): 467–482.

Stiglitz, J.A. 1972. On the optimality of the stock market allocation of investment. *Quarterly Journal of Economics* 86: 25–60.

Díaz-Alejandro, Carlos (1937–1985)

José Gabriel Palma

Abstract

Carlos Díaz-Alejandro was the most prominent Latin American economist of his generation. In his short professional life he gave us powerful insights into Latin America's trade and development, and its economic and financial history. In true Kindlebergian tradition, he was particularly fascinated by the region's many financial crises. His contributions were characterized by a rare capacity to weave together history and theory, abstract economic theory and complex Latin American sociopolitical life. In this way, he avoided the sterility of pure formalistic

theory that characterized so much of the economics of his own generation and the next.

Keywords

Absorption approach to the balance of payments; Capital account liberalization; Capital controls; Current account liberalization; Devaluation; Díaz-Alejandro, C; Economic history; Elasticity approach to the balance of payments; Export-led growth; Hirschman, A. O; Import substitution; Import-substituting industrialization; Inflation; International Monetary Fund; Kindleberger, C; Prebisch, R; Redistribution of income

JEL Classifications

B31

Carlos Díaz-Alejandro was born in Havana and died in New York one day short of his 48th birthday. At 32, he became Yale's youngest ever full professor of economics. In 1983 he moved to Columbia, and at the time of his sudden death he had just accepted a chair at Harvard.

During sabbatical leaves, he visited many Latin American and European universities. Among numerous other activities he was a (dissenting) member of the Kissinger Commission on Central America. He strongly criticized US support for the 'Contras' in Nicaragua (para-military groups associated with the Somoza dictatorship, opposed to the Sandinista government), and insisted that if the United States were serious about Central America it should tie economic assistance to human rights and allow Central American exports free access into its own market. Needless to say, such quixotic attempts to influence US foreign policy were never among his greatest successes!

From a personal point of view I admired his sense of humour and wit, his approachability and 'bridge-building' capacity, his aversion to positions of administrative power, his independence of mind and his common sense.

His work gave us powerful insights into Latin America's trade and development, and its economic and financial history. He was particularly fascinated by the region's many financial crises. His contributions were characterized by a rare capacity to weave together history and theory, abstract economic theory and complex Latin American socio-political life.

In his doctoral dissertation at MIT, Díaz-Alejandro revisited the controversy between the 'elasticity' and 'absorption' approaches to the balance of payments in the context of Argentina's experience of devaluation, concluding that on balance it supported the first approach (1965a). He further argued that one of the main mechanisms through which devaluation influences both the balance of payments and economic growth is through its effects on income distribution. The apparent paradox that many devaluations improve the trade balance but negatively affect the overall growth of output could be explained by the complex redistributive effects of devaluation. In fact, the effectiveness of a devaluation may depend more on the nature of its distributional outcome than on its capacity to change relative prices. Therefore, the exchange rate could be seen as yet another sphere in the struggle between different groups over their shares in national income.

Another peculiar feature of semi-industrialized economies is that '[i]n the long run, the success or failure of a stabilisation effort will depend more on the capacity of governments to obtain a national consensus over the objectives and policy instruments than on the approval or help that they could receive from foreign investors or governments and international agencies' (1985, p. 201).

Díaz-Alejandro also maintained a keen interest in Latin American economic history, writing first on Argentina (1970). Then, in an article on the 1930s crisis, he identified the causes of the dissimilar performances of Latin American economies in the fact that some countries pursued an 'active' approach to fighting recession, while others stuck to conventional 'passive' adjustment mechanisms (1982). The 'active' countries were mainly the large ones, but also included Chile and Uruguay. They performed much better by abandoning the gold standard and by adopting flexible monetary and fiscal policies, real devaluations, moratoria on their foreign debt, and spending massively on public works. This heterodox response of some countries was in part a reaction

to the emergence after the 1929 crash of a protectionist, interventionist and nationalistic Centre.

Díaz-Alejandro's articles on trade and development also discussed the high import intensity of import substitution (1965b), and the transition from import-substituting industrialization to export-led growth (1974). Diaz-Alejandro was particularly sceptical about the idea that this transition would help achieve both faster and more equitable growth. He strongly supported export orientation, but did not believe that it could be achieved simply by 'getting the prices right'; he also feared that it could contribute to 'stop-go' macroeconomics. Moreover, he thought that most of the advice given to Third World countries for their trade policies '. . .suggest evangelical fervour rather than scientific analysis' (1980, p. 332). Díaz-Alejandro reexamined all these issues in his book on Colombia (1975).

He was also a critic of the intervention of the International Monetary Fund (IMF) in markets which were not within its competence:

> It is the business of the IMF to insist on balance of payments targets . . . It is not the business of the IMF to make loans conditional on . . . food subsidies, utility rates, or controls over foreign corporations. . . It was a brilliant administrative stroke for the IMF staff to develop the 'monetary approach to the balance of payments' during the 1950s, allowing the translation of balance of payments targets into those involving domestic credit, but for many LDCs [less developed countries] the assumptions needed to validate such translation, such as a stable demand for money, have become less and less convincing. (1984, p. 169)

He also strongly criticized the IMF intervention in the debt crisis of the 1980s: 'Since August 1982 the world has lived with . . . a peculiar semi-cartelization shakily managed by central banks and the IMF [which] imposes on countries like Brazil the costs of monopoly (for example, larger spreads and fees) without some of its benefits (the ability to plan ahead)' (1983, p. 32).

The economic reforms of the late 1970s and 1980s provided another major intellectual challenge. Not since the 1930s had Latin America witnessed such dramatic economic and political experiments. The new military regimes of the Southern Cone applied their Chicago-oriented policies with a degree of ferocity that rivalled their treatment of political dissent. As Velasco said, Díaz-Alejandro's wisdom was twice as useful because it was delivered in a timely fashion (1988, p. 5). His papers of the late 1970s contain the basic ideas which later became accepted wisdom regarding the policy mistakes of the pro-Chicago governments in Latin America and the irrational behaviour of borrowers and lenders in (highly liquid) national and international financial markets. He particularly questioned the feasibility of simultaneous current and capital account liberalization, the lack of capital controls on speculative inflows, and the use of exchange rate policy to fight inflation.

Among his many articles from this period, his 'Southern Cone Stabilisation Plans' (1981) stands out. Appearing just before the Mexican moratorium which triggered the debt crisis, his argument ran completely against the tide of dominant opinion. Finally, a detailed analysis of the dynamics of the 1982 crisis was the last – and probably best known – of Díaz-Alejandro's contributions (see Palma 2003).

Díaz-Alejandro began his studies at MIT at the time when Fidel Castro landed clandestinely in Cuba in 1956, and graduated at the time of the Bay of Pigs invasion (an unsuccessful CIA-planned and funded invasion by Cuban exiles in south-west Cuba in 1961). He felt that the complexity of the situation was such that he opted for the Miltonian hope that 'they also serve who only stand and wait'.

He had a fascination with Latin American economics. His approach was firmly grounded in the real world, and his work on economic history was rooted in the idea that all history is always the history of the present. As Gustav Ranis remarked, he always 'respected history, used data carefully, and theory selectively' (1989, p. xiv). Like his mentors Hirschman, Kindleberger, Lewis and Prebisch he basically belonged to the 'markets are good servants but bad masters' Keynesian school of economic thought, and always studied economic problems in their historical context, thus avoiding the sterility of pure formalistic theory that characterized so much of the economics of his own generation and the next.

See Also

- ▶ Elasticities Approach to the Balance of Payments
- ▶ Furtado, Celso (1920–2004)
- ▶ Kindleberger, Charles P. (1910–2003)
- ▶ Lewis, W. Arthur (1915–1991)
- ▶ Prebisch, Raúl (1901–1986)
- ▶ Structuralism
- ▶ Terms of Trade
- ▶ Third World Debt

Selected Works

1965a. *Exchange rate devaluation in a semi-industrialized country: Argentina 1955–1961*. Cambridge, MA: MIT Press.

1965b. On the import intensity of import substitution. Repr. in Velasco (1988).

1970. *Essays on the economic history of Argentine*. New Haven, CT: Yale University Press.

1974. Some characteristics of recent export expansion in Latin America. Repr. in Velasco (1988).

1975. *Foreign trade regimes and economic development: Colombia*. New York: NBER.

1980. Discussions. *American Economic Review* 70: 330–335.

1981. Southern cone stabilization plans. Repr. in Velasco (1988).

1982. Latin America in the 1930's. Repr. in Velasco (1988).

1983. Some aspects of the 1982–83 Brazilian payments crisis. *Brookings papers in economic activity* 1983(2): 515–522.

1984. Some economic lessons of the early 1980s. Repr. in Velasco (1988).

Bibliography

Palma, J.G. 2003. The three routes to financial crises. In *Rethinking development economics*, ed. H.-J. Chang. London: Anthem.

Ranis, G. 1989. Carlos Díaz-Alejandro: An appreciation. In *Debt, stabilization and development: Essays in memory of Carlos Díaz-Alejandro*, ed. G. Calvo et al. Oxford: Blackwell.

Velasco, A. (ed.). 1988. *Trade, development and the world economy: Selected essays of Carlos Díaz-Alejandro*. Oxford: Blackwell.

Dickinson, Henry Douglas (1899–1969)

David Collard

Keywords

CES production function; Institutional rents; Market socialism; Neoclassical growth theory

JEL Classifications

B31

Dickinson went from the King's School, Wimbledon, to Emmanuel College, Cambridge, where he took the Part II Tripos in both Economics and History. He carried out research at the London School of Economics under Cannan, then went to teaching posts at Leeds and Bristol, where he held the chair of economics from 1951 to 1964. Although his *Institutional Revenue* (1932) is of interest for generalizing the concept of institutional rents, he is deservedly known for a series of writings which attempted to reconcile choice and individual freedom with socialist planning, in the tradition of market socialism. Together with Taylor, Lange and Lerner he provided a rebuttal (based on actual markets) of von Mises's view that rational allocation under socialism was impossible. He saw 'the beautiful systems of economic equilibrium' not as 'descriptions of society as it is but prophetic visions of a socialist economy of the future' (1933, p. 247). During the 1930s his writings were well known to intellectuals of the Left, including Cole, Dalton, Durbin and Laski. The best-known of his works is the *Economics of Socialism* (1939). His technical prowess was later exhibited in a *Review of Economic Studies* article of 1954–5 in which he formulated a constant elasticity of substitution production function (CES) for the first time and anticipated some of the neoclassical growth results of Solow and Swan. 'Dick', as he was universally known, was a much loved, unworldly, eccentric figure with a keen sense of fun and a most astute mind.

Selected Works

1932. *Institutional revenue: A study of the influence of social institutions on the distribution of wealth*. London: Williams & Norgate.
1933. Price formation in a socialist community. *Economic Journal* 43: 237–250.
1939. *Economics of socialism.* London: Oxford University Press.
1955. A note on dynamic economics. *Review of Economics Studies* 22(3): 169–179.

Diderot, Denis (1713–1784)

Peter Groenewegen

Philosophe and editor of the *Encyclopédie raisonée* (1751–72). Born at Langres he was educated locally by the Jesuits and moved to Paris in 1728 to complete his education at the University of Paris and earn his living as a writer and translator. Diderot is ensured immortal fame for his role in commencing, editing and publishing the famous *Encyclopédie,* initially with D'Alembert but, after final government prohibition in May 1757, by himself. This task took close to half his lifetime; it was first mooted in 1746 and in 1772 the last volumes of engravings were published. The first volume of text appeared in 1751, the last in 1765; those containing the important economic contributions by Quesnay (1756, 1757) and Turgot (1757) appearing just before its official proscription by the censor. The completion of this task allowed Diderot time for a seven-month visit to Catherine the Great, whom he advised on various matters including economic policy.

In 1774 homesickness induced his departure from St Petersburg for his beloved Paris, where he died in 1784. His departure from St Petersburg has also been ascribed to a practical joke executed by Euler but inspired by the Czarina herself because of Diderot's boorish behaviour at Court. This involved an algebraic proof of the existence of God. Euler is said to have confronted Diderot with the following statement, spoken in a tone of perfect conviction.

$$'\text{Sir, } \frac{a+b^n}{n} = x, \text{ hence God exists, reply!}$$

Not at all skilled in mathematics Diderot had no answer, and humiliated by the unrestrained laughter which greeted his silence he asked, and received, Catherine's permission to return immediately to France (Bell 1937, pp. 159–60).

As Bauer (1894, p. 577) noted, 'there is hardly a single branch of science which does not owe some debt of gratitude to the universal genius of this very able and characteristic French writer.' Because Bauer (1894) also gives a detailed summary of Diderot's economic contributions particularly for the *Encyclopédie,* other aspects of Diderot's importance for economics are mentioned here, particularly the role he played in disseminating French economics through its pages. This included work by Quesnay and Turgot and from lesser economic lights like Forbonnais, Leroy and Morellet. Beccaria and Verri learnt of Physiocracy from this source; likewise Sir James Steuart (1767, I p. 110) who cited Quesnay (1757), while Adam Smith (1756, p. 246) in his second published article praised the *Encyclopédie* as a work promising 'to be the most compleat of this kind which has ever been published or attempted in any language'. The fact that Diderot enabled Quesnay to publish his first two economic essays in this work is particularly noteworthy, when it is realized these provide both the analytical foundations and the actual impetus for the creation of a Physiocratic school. Groenewegen (1983, pp. xii–xiii) has surmised that the reasons for this may have been Quesnay's considerable influence with Mme de Pompadour and its potential usefulness to Diderot for solving the continuing censorship crises the *Encyclopédie* was facing. Sadly, the censor ultimately prevented Quesnay from publishing two further contributions and consequently these remained virtually unknown for nearly 150 years after they were written. Likewise, official proscription made Turgot decline Diderot's invitation to write articles on

topics including the rate of interest and taxation. Diderot appears also to have had an initial enthusiasm for Physiocratic thought, sufficiently strong to praise Mercier (1767) but in 1770 he actively supported Galiani's (1770) anti-Physiocratic grain trade position by first seeing the *Dialogues* through the press and then defending them against Morellet's criticisms (Mason 1982, pp. 324–6).

Diderot's influence on some prominent nineteenth century thinkers may also be briefly noted. Diderot (1761) is distinguished by being the only modern work cited in Hegel's *Phenomenology of Spirit.* Marx singled it out as a 'masterpiece from beginning to end', describing Diderot as one of his favourite authors (Mason 1982, pp. 184, 367). Marx's views were probably inspired by both Diderot's vigorous materialism and his wit, but may also derive from his views on the philosopher's role, so similar to Marx's own: 'What use is philosophy if it is silent? You must either speak out, or renounce the title of instructor of the human race. You will be persecuted, that is your destiny . . .' (Diderot 1778, p. 365). Only such sentiments could have sustained the 15 years lonely and arduous labour of completing his task of 'collecting the scattered knowledge of the world, revealing its overall structure and passing it to future generations' (Diderot 1755, p. 174) as he himself defined his work on the *Encyclopédie.*

Selected Works

1755. Encyclopédie. In *Encyclopédie ou dictionnaire raisonné des sciences, des arts et des métiers*, Paris, vol. 5.
1761. *Rameau's nephew.* Trans. L.W. Tencock. Harmondsworth: Penguin Books, 1966.
1778. *Essay on the Reigns of Claudius and Nero*. Extracts translated in Mason (1982).

References

Bauer, S. 1894. Denis Diderot. In *Dictionary of political economy*, vol. I, ed. R.H.I. Palgrave, 577–9. London: Macmillan.
Bell, E.T. 1937. *Men of mathematics*, 1953. Harmondsworth: Pelican Books.
Galiani, F. 1770. *Dialogues sur le commerce des bléds*. Paris.
Groenewegen, P.D. 1983. *Introduction to Quesnay, Farmers*. Sydney: Department of Economics/Sydney University.
Mason, J.H. 1982. *The irresistible Diderot*. London: Quartet.
Mercier de la Rivière, P.P. 1767. *L'ordre naturel et essential des sociétés politiques*. Paris.
Smith, A. 1756. A letter to the authors of the Edinburgh review. In *Essays on philosophical subjects*, ed. W.P.D. Wightman, J.C. Bryce, and I.S. Ross. Oxford: Clarendon, 1980, 242–54.
Steuart, Sir James. 1767. *Principles of political oeconomy.* London

Dietzel, Carl August (1829–1884)

Alan Peacock

German writer on public finance; Privatdozent of Heidelberg and later Professor of Public Finance, Marburg. Dietzel was struck by the contrast between British views on public debt such as those of Hume – 'if the Nation does not destroy Credit, Credit will destroy the Nation' – and the fact that the notable growth in British public debt during the 19th century had not been accompanied by the ruin of the British economy. Writing in 1855 he attacked the orthodox view that state borrowing required a Sinking Fund, arguing that government investment financed by renewable loans was a necessary condition for the growth in national production. His views were endorsed by several prominent German writers, notably Adolph Wagner, and were recalled during the post-1936 debate in support of Keynesian views on public debt policy.

Selected Works

1855. *Das System der Staatsanleihen im Zusammenhang der Volkswirtschaft betrachtet*. Heidelberg.

References

Stettner, W. 1948. Carl Dietzel, public expenditures and the public debt. In *Income, employment and public policy: Essays in honor of Alvin H. Hansen*, ed. L.A. Metzler. New York: Norton.

Dietzel, Heinrich (1857–1935)

H. C. Recktenwald

Keywords

Dietzel, H.; Eucken, W.; Freiburg School

JEL Classifications

B31

Born in Leipzig, Dietzel was appointed to a chair at the University of Dorpat in 1885 after studies in economics and law in Heidelberg, Göttingen and Berlin. In 1890 he accepted a chair in the philosophy faculty in Bonn. There he died in 1935.

Dietzel was a respected figure in circles of 19th-century German economists (such as Rau, von Thünen, von Hermann, von Mangoldt and Wagner) who were endeavouring to defend, pursue and modify classical methods and principles. He kept a sceptical distance from both the younger Historical School and the Austrian School, and was sharply opposed to popular Marxism. Nevertheless his excellent biography of Rodbertus and his writings on the early socialists are proof of his academic openness and liberal fairness. Enthusiastically though not successfully engaged in propagating free trade, Dietzel (in contrast to Manchester liberalism) was not dogmatic concerning the functions of the state in a concrete mixed economy.

His most important contribution to theory, the *Theoretische Sozialökonomie* (1895), unfortunately remained a torso. It is a pioneering analysis of the two main orders of an economy, namely, the individualistic system of competitive markets and the collective system of compulsion of the state. This concept of the two (centralized and decentralized) elementary forms replaced the unscientific notions of capitalism and socialism, with their ideological bias. It opened the way to the foundation of an order theory that his disciple in Bonn, Walter Eucken, and the Freiburg School further developed and later on applied in Germany.

Though Dietzel dealt with self-interest, methodological theory (1911) and value theory, he and his followers (as Smithians) did not attempt to unify Smith's three systems of ethics, economics and politics to an integrated order theory via reconstructing and developing his 'obvious and simple system of natural liberty'. They also failed to produce an analysis of state and collective failures while they originally stressed the state's responsibility for ensuring sufficient market competition.

Nevertheless they made a number of contributions to the field and pointed to the right road to be taken in the future.

Selected Works

1882. *Über das Verhältnis der Volkswirtschaftslehre zur Sozialwirtschaftslehre*. Berlin: Puttkammer und Mühlbrecht.

1886–8. *Karl Rodbertus: Darstellung seines Lebens und seiner Lehre*. 2 vols. Jena: G. Fischer.

1895. *Theoretische Sozialökonomie, I*. Leipzig: Winter.

1911. Selbstinteresse und Methodenstreit in der Wirtschaftstheorie. In *Handwörterbuch der Staatswissenschaften,* vol. 7. Jena: G. Fischer.

1921. *Vom Lehrwert der Wertlehre und vom Grundfehler der Marxschen Verteilungslehre*. Leipzig-Erlangen: Scholl.

1922. *Technischer Fortschritt und Freiheit der Wissenschaft*. Bonn–Leipzig: Schroeder.

Bibliography

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

Recktenwald, H.C., and P.A. Samuelson. 1986. Über Thünen's 'Der isolierte Staat'. *Wirtschaft und Finanzen*, Darmstadt-Düsseldorf.

Difference-in-Difference Estimators

Alberto Abadie

Abstract

This article discusses difference-in-differences (DID) estimators, which are commonly applied in evaluation research. In particular, the discussion focuses on (*a*) motivation, definition and interpretation of DID estimators, (*b*) conditions under which DID estimators are valid, (*c*) data requirements to compute DID estimators, (*d*) representative applications of DID estimators in the empirical economics literature, (*e*) extensions of DID estimators, and (*f*) a simple indirect test to assess the validity of these estimators.

Keywords

Difference-in-differences estimators; Evaluation studies; Fixed effects; Minimum wages

JEL Classifications

C13

Motivation and Definition

Difference-in-differences (DID) estimators are often used in empirical research in economics to evaluate the effects of public interventions and other treatments of interest in the absence of purely experimental data.

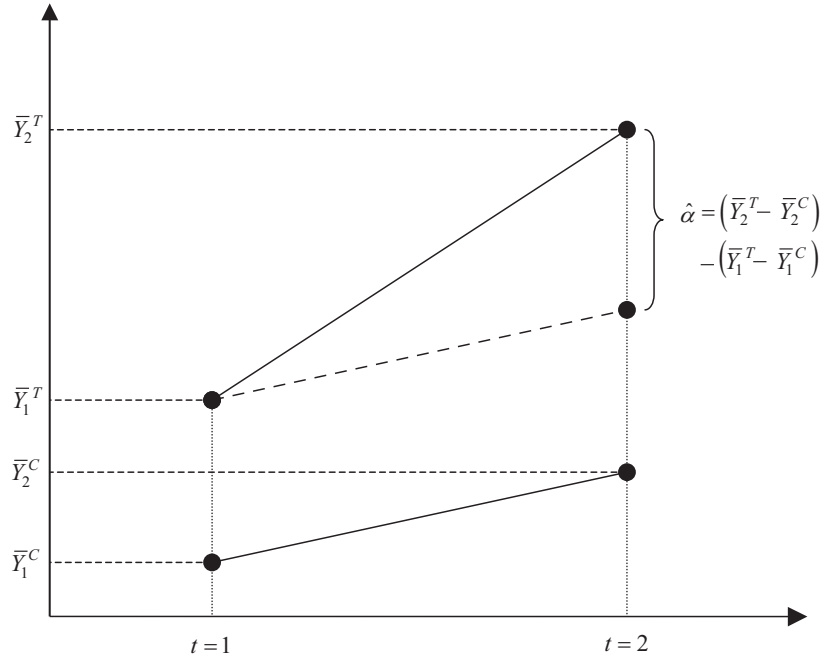
The usual goal of evaluation studies is to estimate the average effect of a treatment (for example, participation in a vocational training programme) on some outcome variable of interest (for example, earnings or employment). Often researchers concentrate on estimating the average effect of the treatment on the treated, that is, on those individuals exposed to the treatment or intervention (for example, the trainees). In the typical setting of an evaluation study, we observe an outcome variable, Y_i, for a sample of treated individuals and also for a sample of untreated individuals. The main challenge in evaluation research is to find an appropriate comparison group among the untreated individuals, in the sense that the distribution of the outcome variable for the untreated comparison group can be taken as an approximation to the counterfactual distribution that the outcome variable, Y_i, would have followed for the treated in the absence of the treatment.

Sometimes the sample of untreated individuals may not provide an appropriate comparison group, and therefore differences in the distribution of the outcome variable between treated and untreated reflect not only the effect of the treatment but also intrinsic differences between the two groups. To address this problem, the DID estimator uses the assumption that in the absence of the treatment the average difference in the outcome variable, Y_i, between treated and untreated would have stayed roughly constant. Then, the average difference in the outcome variable between treated and untreated before the treatment can be used to approximate the part of the difference in average outcomes after the treatment that is created by intrinsic differences between the two groups and not by the effect of the treatment.

Let $\overline{Y}_1^T$ and $\overline{Y}_1^C$ be the average outcomes in period t (t =1, 2) in the treated and untreated samples, respectively. Period $t = 1$ takes place before the treatment and period $t = 2$ takes place after the treatment. The difference in average outcomes between treated and untreated after the treatment is $\overline{Y}_2^T - \overline{Y}_2^C$. The same difference for the pre-treatment period is $\overline{Y}_1^T - \overline{Y}_1^C$. Then, the DID estimator is defined as follows:

$$\hat{\alpha} = \left(\overline{Y}_2^T - \overline{Y}_2^C\right) - \left(\overline{Y}_1^T - \overline{Y}_1^C\right) \qquad (1)$$

Figure 1 provides a graphical interpretation of the DID estimator. The solid lines represent the evolution in average outcomes for the treated and the untreated comparison group between the pre-treatment period ($t = 1$) and the post-treatment period ($t = 2$). The dashed line approximates the counterfactual evolution that the average outcome would have experienced for the treated in the absence of the treatment. This line is constructed under the DID



Difference-in-Difference Estimators, Fig. 1

assumption that, in the absence of the treatment, the difference in average outcomes between treated and untreated would have stayed roughly constant in the two periods. As reflected in Fig. 1, an equivalent formulation of the DID assumption is that, in the absence of the treatment, average outcomes for treated and untreated would have followed a common trend. As a result, the untreated comparison group can be used to infer the counterfactual evolution of the average outcome for the treated in the absence of the treatment.

Difference in differences estimators have been applied to the study of a variety of issues in economics. Card and Krueger (1994) evaluate the employment effects of an increase in the minimum wage in New Jersey using a contiguous state (Pennsylvania), which did not increase the minimum wage, to approximate how employment would have evolved in New Jersey in the absence of the raise. Card (1990) applies DID estimators to evaluate the employment effects of the massive flow of Cuban immigrants to Miami during the 1980 Mariel boatlift. To estimate the effects of the boatlift, Card uses a group of four comparison cities to approximate how employment would have evolved in Miami in the absence of the 1980 immigration shock. Other applications of the DID estimator include studies of the effects of disability benefits on time out of work (Meyer et al. 1995), the effect of anti-takeover laws on firms' leverage (Garvey and Hanka 1999), and the effect of tax subsidies for health insurance on health insurance purchases (Gruber and Poterba 1994).

The DID estimator has a simple regression representation. Let Y_{it} be the outcome of interest (for example, earnings) for individual i at time t, with $i = 1, \ldots, N$ and $t = 1, 2$. Let D_i be an indicator of membership to the treatment group, so $D_i = 1$ for the treated and $D_i = 0$ for the untreated. Finally, let $\Delta Y_i = Y_{i2} - Y_{i1}$ be the change in the outcome variable between the pre-treatment and the post-treatment period for individual i. The regression representation of the DID estimator is:

$$\Delta Y_i = \mu + \alpha D_i + u_i, \tag{2}$$

where u_i is a regression error, which is mean independent of D_i (that is, $E[u_i]D_i = 1] = E[u_i]$

$D_i = 0$]). It can be easily seen that the ordinary least squares estimator of α in Eq. (2) is numerical identical to the DID estimator, $\hat{\alpha}$, in Eq. (1). Regression standard errors along with the point estimate, $\hat{\alpha}$, can be used to construct confidence intervals for α and perform statistical hypothesis tests. As reflected in Eq. (2) and emphasized in Blundell and MaCurdy (1999), the DID estimator is a particular case of fixed effects estimators for panel data, with only two time periods and a fraction of the sample exposed to the treatment in the second time period.

Extensions

In some instances, the common trend assumption adopted for DID is not plausible because treated and untreated differ in the distribution of some variables, X_i, that are thought to affect the trend of the outcome variable. In this situation, treated and untreated may exhibit different trends in the average of the outcome variable between $t = 1$ and $t = 2$, even if the treatment does not have any impact on the outcome of interest. The regression formulation of the DID estimator is useful to compute a conditional version of the DID estimator that corrects for the effect of X_i on the trend of Y_i:

$$\Delta Y_i = \mu + \alpha D_i + X_i'\beta + u_i.$$

Abadie (2005) and Heckman et al. (1997) develop semiparametric and nonparametric versions of the conditional DID estimator.

Panel data are not always necessary to apply the DID estimator. A simple inspection of Eq. (1) indicates that $\hat{\alpha}$ can be estimated from repeated cross sections, using a cross-section at time $t = 2$ to estimate $\overline{Y}_2^T - \overline{Y}_2^C$ and a cross section at time $t = 1$ to estimate $\overline{Y}_1^T - \overline{Y}_1^C$. A regression formulation of the DID estimator is also available for repeated cross sections (see, for example, Meyer 1995; Abadie 2005). When the DID estimator is constructed using repeated cross sections, it is important to check whether there exist compositional changes in the sample between the two periods. Compositional changes may constitute a threat to the assumption that the difference in the average outcome between treated and untreated would have stayed constant in the absence of the treatment.

In general, the DID assumption cannot be tested directly with data from $t = 1$ and $t = 2$ only. However, if the common trend assumption extends to more than one pretreatment period for which data are available, pre-existing differences in the trends of the outcome variable between treated and untreated can be detected by applying the DID estimator to pretreatment data. This is done by constructing ΔY_i as the difference in the outcome variable for individual i between two pretreatment periods. Then, a test of the hypothesis $\alpha = 0$ in Eq. (2) is a test of the common trend assumption. In addition, the DID assumption can sometimes be rejected when the dependent variable has bounded support (for example, when Y_i is a binary variable). If the dependent variable has bounded support the DID assumption may imply that, in the absence of the treatment, the average outcome for the treated would have lain outside the support of the dependent variable (see Athey and Imbens 2006).

For a more detailed explanation of the theory behind DID estimators, see Abadie (2005), Angrist and Krueger (1999), Ashenfelter and Card (1985), Blundell and MaCurdy (1999), Heckman et al. (1997), and Meyer (1995).

See Also

- ▶ Fixed Effects and Random Effects
- ▶ Treatment Effect

Bibliography

Abadie, A. 2005. Semiparametric difference-in-differences estimators. *Review of Economic Studies* 72: 1–19.

Angrist, J.D., and A.B.. Krueger. 1999. Empirical strategies in labor economics. In *Handbook of labor economics*, ed. O. Ashenfelter and D. Card, Vol. 3A. Amsterdam: North-Holland.

Ashenfelter, O., and D. Card. 1985. Using the longitudinal structure of earnings to estimate the effects of training programs. *Review of Economics and Statistics* 67: 648–660.

Athey, S.C., and G.W. Imbens. 2006. Identification and inference in nonlinear difference-in-difference models. *Econometrica* 74: 431–498.

Blundell, R., and T. MaCurdy. 1999. Labor supply: A review of alternative approaches. In *Handbook of labor economics*, ed. O. Ashenfelter and D. Card, Vol. 3A. Amsterdam: North-Holland.

Card, D. 1990. The impact of the Mariel Boatlift on the Miami labor market. *Industrial and Labor Relations Review* 44: 245–257.

Card, D., and A.B.. Krueger. 1994. Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review* 84: 772–793.

Garvey, G.T., and G. Hanka. 1999. Capital structure and corporate control: The effect of antitakeover statutes on firm leverage. *Journal of Finance* 54: 519–546.

Gruber, J., and J. Poterba. 1994. Tax incentives and the decision to purchase health insurance: Evidence from the self-employed. *Quarterly Journal of Economics* 109: 701–733.

Heckman, J.J., H. Ichimura, and P.E. Todd. 1997. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies* 64: 605–654.

Meyer, B.D. 1995. Natural and quasi-experiments in economics. *Journal of Business & Economic Statistics* 13: 151–161.

Meyer, B.D., W.K. Viscusi, and D.L. Durbin. 1995. Workers' compensation and injury duration: Evidence from a natural experiment. *American Economic Review* 85: 322–340.

Differential Games

Simone Clemhout and Henry Y. Wan Jr.

A differential game studies system dynamics determined by the interactions of agents with divergent purposes. As a limit form of multi-stage games, its non-cooperative solution is subgame perfect; thus it may facilitate the study of credible threats and repeated play. Reducing each stage to a single point in continuous time, differential game applies control theoretic tools (including phase diagrams) to yield results more general and more detailed than other methods. Its applications range from common-property resource utilization to macroeconomic stabilization.

Model

A differential game has four components: (a) a *state space, X,* where x in X embodies all relevant data at a particular stage, (b) a *time horizon, T:* a closed interval with a final instant equal to infinity or decided by some termination rule, (c) *a set of players,* $\overline{N} = \{1, \ldots, i, \ldots, N\}$, with each player distinguished by four aspects: (1) a *space for possible moves* (or 'controls'), K_i; (2) a *point-to-set correspondence for allowable moves,* C_i: $X \times T \Rightarrow K_i$, $(x, t) \mapsto C_i(x, t)$ which vary with (x, t); (3) a *space for admissible 'strategies'* (or 'policies'), $R_i = \{r_i r_i : X \times T \to \mathrm{U}_{X \times T} C_i(x, t)$ and r_i satisfies *additional conditions*$\}$, where each r_i assigns an allowable move at every (x, t); and (4) the *instantaneous payoff* u_i: $X \times T \times \prod_j K_j \to R$, $((x, t), (c_j)) \mapsto u_i((x, t), (c_j))$. The additional conditions include (i) any restrictions on the information used for decisions, (ii) regularity conditions (e.g., being step-wise continuous) needed for a well defined model (d) a *state equation* for state transition, $F : X \times T \times \Pi_j K_j \to X$, $\big((x, t), (c_j)\big) \mapsto .x, X$ and K_1 ,. . ., K_n are all subsets of Euclidean spaces.

Players select *strategies* at the outset, not piecemeal *moves.* Strategies are defined here as state-and-time dependent or 'feedback' strategies including the subclass which are 'open-loop' or time-dependent (only).

An Example with Two Variations

Two users share a natural resource, which may be a petroleum reserve or fishery, under common-property tenure. The state space X is the set of all non-negative resource levels and the time horizon is $T = [0, t_f]$ where $t_f = +\infty$ or the instant when all the resource is used up. For all (x, t), $i = 1, 2$, the 'allowable moves' form a set $C_i(x, t) = K_i = R_+$, the set of all non-negative rates of use. Specimens of strategies include $r_i = kx$ for some $k \geq 0$, or $r_i = g(t)$ for some non-negative-valued function. The instantaneous payoffs of both players are assumed to be: $u_i = \exp(-\mathrm{at}) \log c_i$ for some $a > 0$. The 'state equation' is: $\mathrm{d}x/\mathrm{d}t = f(x) \propto - c_1 - c_2$ where: $f(x) = 0$ for the case of petroleum reserves, and $f(x) = x(b - \log x)$ for the fishery. The latter form agrees with the Gompertz recruitment function.

Solution Concepts

How players choose strategies under various scenarios is summarized as three *solution concepts,* e.g. (1) *The noncooperative equilibrium* (Cournot–Nash): each player's choice is his 'best reply' to the choices of all other players. This choice must be 'best' for all initial (x, t); (2) *The cooperative equilibrium* (Pareto): all players make choices such that no modification can benefit any player without harming another. This property holds for all initial (x, t); (3) *The hierarchical equilibrium* (Stackelberg): the 'leader' selects a committed choice to elicit the followers' 'best replies' so that the leader's payoff is maximized.

An equilibrium is a vector of strategies, one for each player, which is not liable to change. In differential games, players may change strategies in midgame, unless prevented by prior commitment (as in (3)), or by requiring the choices to be appropriate, once *and* forever (as in (1) and (2)). Significantly, the Cournot–Nash solution is thus subgame perfect à la Selten.

The Cournot–Nash solution is most frequently used. In particular, it depicts externalities under laissez-faire. For any problem it can be compared to Pareto solutions which assess any extra gains resulting from cooperation.

If an acknowledged leader (e.g. the *government* in a macroeconomy) can offer credible commitments, he prefers to play Stackelberg (with a higher payoff for himself) rather than Cournot–Nash, since all followers' best replies are now under his influence rather than given independently.

The differential game sheds light on two additional features: (i) the *credibility* of the leader's professed strategy which is at issue, since he has both (a) the opportunity to renege on promises made and honoured at different times, and (b) the incentive to renege (his choice is subgame imperfect); (ii) 'Reputation' (rather than 'enforcement') is often the reason why commitments are kept, and may be modelled as a state variable suggested in Clemhout and Wan (1979). Hence credibility is established from a balance of the gains of reneging with the damage from reputation lost. This suggests a synergistic approach between Cournot–Nash and Stackelberg.

Alternative Formulations of Cournot–Nash Models Over Time

To characterize the Cournot–Nash differential game in feedback strategy, one contrasts it with alternative versions of differing assumptions, 'what information players use' and 'the modelling of time'. Examples show that:

(a) To explain *reality* and provide *policy relevance,* 'feedback' strategies are preferable to 'open-loop' strategies for two reasons; (1) in non-cooperative games, the subgame-perfect equilibrium is the image of reality, and (2) in models of common-property resources, policy relevance hinges on identifying the source of inefficiency. Our petroleum example (cf. Clemhout and Wan 1985a) is a non-cooperative model of common-property utilization, and thus should have an *inefficient* but *subgame-perfect* equilibrium. This is the case when strategies are 'feedbacks'. The opposite is true if strategies are modelled as 'openloop' in which all equilibria are then *efficient* and *subgame-imperfect* and each is compatible only with one initial resource stock.

(b) For *reasonableness* and *convenience,* 'history-dependent state variables' are preferable to 'history-dependent strategies'. While history matters in contexts such as performance contracts, history-dependent strategies tend to require an infinite amount of information at every move. The use of history-dependent state variables (Smale 1980, cf. Clemhout and Wan 1979) is a reasonable alternative for players with bounded rationality. It also conforms to the finite-dimensional state space in differential games.

(c) For *game-theoretic* and *analytic* reasons 'continuous time' is preferable to 'discrete time'. Our fishery example (Clemhout and Wan 1985a) illustrates two points. First only in continuous time is the model a game, according to Ichiishi (1983). To ensure non-negativity of the resource level, discrete-time models require the allowability of one player's move to depend upon the moves of all others at the same time, thus they become 'pseudo-games' by losing

playability. The second point is that only in continuous time are the dual variables (which are analytically important) derivable from the conditions necessary for optimality whether the recruitment function is concave or not. This is because the adjoint system, in differential equation form, involves the slope of the recruitment function alone and not its curvature.

Strengths of Differential Games

Differential games can obtain precise results either independently of particular functional forms, or by using empirically validated formulations. In our fishery example these include characterization of the resource level: (a) does it reach a sustained level? (b) does it approach extinction asymptotically, if so, how rapidly? (c) is it heading for extinction in finite time? (d) what difference do risks of random perturbation or extinction make? (e) what if several harvested species form pre-predator chains? and (f) do tax-incentives improve allocation efficiency? (Clemhout and Wan 1985a, b, c).

In contrast with differential games, intuitive reasoning or simple examples (in two or three periods) can suggest certain outcomes, but cannot rule out the opposite outcome occurring in plausible situations. Simulation models can start from any assumptions but cannot assure equilibrium.

In macro-economics, the linear-quadratic-Gaussian differential game can further analyse quantitatively real-life data. The estimation and interpretation of the parameters in such models is still subject to ongoing research. The same model also yields deep economic insights in their micro-economic applications.

Concluding Remarks

Pioneered by Isaacs and generalized by Case, the theory of differential games is now covered by excellent texts (e.g., Basar and Olsder 1982), with reference to contributions by Blaquiere, Berkovitz, Cruz, Fleming, Friedman, Haurie, Ho and Leitmann, among others. Further progress in its economic applications now hinges on the development of 'techniques of analysis', akin to phase diagrams in control theory. Using these techniques one can deduce implications crucial to economists working with particular classes of models. This is often accomplished by utilizing structural properties common to entire families of models. The explicit solutions are neither required nor derived. Such feats are clearly attainable for differential games, as they have been for control models: the phase diagram itself has been recently applied to some models (Clemhout and Wan 1985b) and contraction mappings in others (Stokey 1985). Given the state of the art in this field, additional advances in theory (e.g., generalizing the model, proposing new solution concepts, etc.) are certainly most welcome, but no longer crucial for economic applications.

See Also

- ▶ Game Theory
- ▶ Non-cooperative Games
- ▶ Optimal Control and Economic Dynamics
- ▶ Repeated Games

Bibliography

Basar, T., and G.J. Olsder. 1982. *Dynamic noncooperative game theory*. New York: Academic.

Clemhout, S., and H. Wan Jr. 1979. Interactive economic dynamics and differential games. *Journal of Optimization Theory and Applications* 27(1): 7–30.

Clemhout, S., and H. Wan Jr. 1985a. Resource exploitation and ecological degradations as differential games. *Journal of Optimization Theory and Applications* 19: 471–481.

Clemhout, S., and H. Wan Jr. 1985b. Cartelization conserves endangered species? In *Optimal control theory and economic analysis*, vol. 2, ed. G. Feichtinger. Amsterdam: North-Holland.

Clemhout, S., and H. Wan Jr. 1985c. Common-property exploitations under risks of resource extinctions. In *Dynamic games and applications in economics*, ed. T. Basar. New York: Springer.

Ichiishi, T. 1983. *Game theory for economic analysis*. New York: Academic.

Smale, S. 1980. The Prisoner's Dilemma and dynamical systems associated to non-cooperative games. *Econometrica* 48(7): 1917–1934.

Stokey, N. 1985. The dynamics of industry-wide learning. In *Essays in honour of Kenneth J. Arrow*, ed. W.P. Heller, R.M. Starr, and D.A. Starrett. Cambridge: Cambridge University Press.

Difficulty of Attainment

F. Y. Edgeworth

A phrase used by De Quincey, Mill, and others, to denote a condition which must be superadded to utility in order that there should exist value in exchange.

> Any article whatever, to obtain that artificial sort of value which is meant by exchange value, must begin by offering itself as a means to some desirable purpose; and secondly, even though possessing incontestably this preliminary advantage, it will never ascend to an exchange value in cases where it can be obtained gratuitously and without effort (De Quincey, *Logic of Political Economy,* p. 13; quoted by Mill, *Political Economy,* book iii, ch. ii, § 1).

The difficulty of attainment here indicated is primarily that which is experienced by the purchaser. But it is usual to extend the term to the difficulty experienced by the producer. Thus De Quincey continues:

> Walk into almost any possible shop, buy the first article you see; what will determine its price? In the ninety-nine cases out of a hundred simply . . . difficulty of attainment. . . . If the difficulty of producing it be only worth one guinea, one guinea is the price which it will bear.

So Mill, of what he considers the general case, 'the obstacle to attainment consists only in the labour and expense necessary to produce the commodity' (ibid., § 2). And by others difficulty of attainment is used as equivalent to cost of production. Thus Walker (First Lesson in Political Economy, Art. 67), 'Cost of production is only another name for difficulty of attainment.' This transition from the sense in which the difficulty, like the other factor utility, is experienced by the individual purchaser is legitimate, where there exists such perfect 'industrial' competition that it is free to any one to enter any occupation. In that case the sacrifice made to attain a commodity by purchase tends to be equivalent to the efforts and sacrifices made in attaining it by production. If the value in exchange were higher, the commodity would not be purchased; if lower, it would not be produced.

The wider conception is particularly appropriate to the case which Mill, dividing the different kinds of difficulty, places second; where, 'without a certain labour and expense it [the commodity] cannot be had; but, when any one is willing to incur this, there needs be no limit to the multiplication of the product' . . . up to a point which there is no need, for practical purposes, to contemplate (Political Economy, book iii, ch. ii, § 1). In this case difficulty of production has a certain pre-eminence over the co-factor utility, both as (a) a cause, and (b) a measure of value. (a) The cause of a phenomenon being usually a somewhat arbitrarily selected portion of its total antecedent (Mill, Logic, book iii, ch. v, § 3; Venn, Empirical Logic, p. 57 et seq.), it is not paradoxical that sometimes utility, sometimes cost, should be regarded as the cause of value. Utility indeed is invariably an antecedent. But the scale of utility may, in the case supposed, be varied without any variation of value. 'If the demand for hats should be doubled, the price would immediately rise; but that rise would be only temporary, unless the cost of production of hats . . . were raised' (Ricardo, Political Economy, ch. xxx). Whereas, if the cost of production of an article is varied, its value varies concomitantly. 'Diminish the cost of production of hats, and their price will ultimately fall to their new natural price, although the demand should be doubled, trebled, or quadrupled' (Ricardo, ibid.). Prediction, the prerogative of causation, is attached to cost rather than utility. (b) Accordingly, in the case supposed, the comparative difficulty of producing two commodities affords a simple measure of their relative value. It is true also that value is proportioned to final utility. But this measure cannot be read until the measurement is already given. We cannot tell what the final utilities will be till we know the values. In some cases indeed (see below (4) and (5)) it is conceivable that, given the dispositions, the Demand-Curves of all the dealers in a market, we could deduce the rate of exchange which will be set up. The calculation is indicated by Professor Walras in his Éléments d'économie politique

pure, Art. 50. Still difficulty of production, in the case most favourable to its operation, measures value directly, as a clock measures time; whereas utility at best is a measure like the shadow cast by the sun, which can only be interpreted by a difficult calculation.

This theory is subject to several reservations and exceptions.

(1) The pre-eminence of difficulty of production as a regulator of value depends largely on the assumption that labour is perfectly homogeneous. If all labour consisted of raising weights in precisely similar circumstances, the theory might be literally true. 'If . . . it usually cost twice the labour to kill a beaver which it does to kill a deer, one beaver should naturally exchange for or be worth two deer' (Adam Smith, quoted by Ricardo), there being only one mode of labour, work being as homogeneous as, say, gold. But suppose, besides effort of exertion, the sacrifice of waiting is required. Then, as between commodities involving these elements in different proportions (cf. Ricardo, ch. i, § 4), it would no longer be possible to assign the rate of exchange between the commodities without being given the comparative remuneration for the two kinds of sacrifice. But this datum could not in general be obtained a priori, but only as a result of the higgling of the market. (This reservation holds even upon the imaginary supposition that there existed a competition so perfect that it is free to any one to choose whether he will labour or abstain, a fortiori, when, as in reality the abstainers form a 'non-competing group'; and so fall under head (3).) Now, in fact, there are not only two, but many, kinds of sacrifice. The general principle is that the 'net advantages' (Marshall, Principles of Economics, vol. i, 2nd edn, p. 136) in occupations between which there is 'industrial competition' (Cairnes), tend to be equal. Accordingly the statement that the 'quantity of labour realised in commodities' (Ricardo) regulates their exchangeable value, can be true only on an average with wide deviations. Take the case put by De Quincey of a pearl-diver who sometimes obtains, along with 'ordinary', superior pearls. The true principle is that the net advantages of pearl-diving are the same as those of any other occupations between which there is industrial competition. How much truth is there in the proposition that the value of any pearl is proportioned to the 'quantity of labour realised' in it? The instance taken is a mild case of plural occupations, or joint production. The application of the general principle of net advantages here affords little light as to the value of particular articles (cf. Sidgwick, Political Economy, book ii, ch. ii, § 10).

(2) The pre-eminence of difficulty over utility, as a regulator of value, disappears altogether when we pass from Mill's second case to a category comprising both Mill's third case (Political Economy, book iii, ch. ii, § 2), in which the cost of production increases with the quantity produced, according to the law of Diminishing Returns, and the converse case, in which the cost of production diminishes with the quantity produced according to the law of Increasing Returns. In this case the two factors, utility and value, become coordinate. As Professor Marshall says (Economics of Industry, 1st edn, p. 148),

> the amount produced and its normal value are to be regarded as determined simultaneously under the action of economic laws. It is then incorrect to say as Ricardo did, that cost of production alone determines values; but it is no less incorrect to make utility alone, as others have done, the basis of value.

With reference to what Jevons calls the 'mechanics of industry' it seems trifling to inquire whether the force or the resistance contributes more to the determination of equilibrium. The simultaneousness of the two conditions is indicated by Jevons in his discussion of cost of production (Theory, ch. v). Jevons there entertains the unreal conception that it is free to the producer to apply his efforts in 'doses' to different kinds of production. This at most is true of the mere inventor as distinguished from the entrepreneur and operative. Still the conception may be usefully employed as symbolical of the actual working of competition in a regime of division of labour

(Pantaleoni, Principii. Theorema di Ricardo ed Marshall). The simultaneousness of the two conditions may best be shown by imagining the disutility, as well as the utility, to be of the sort called 'final'.

(3) The coordinateness of difficulty of production with utility disappears when industrial competition is no longer supposed. In this case the assumed equation between the purchaser's and the producer's difficulty of attainment fails. The typical instance is international trade. There is no correspondence between the efforts of the Chinese producer of tea and the sacrifices which the English purchaser incurs to obtain it. It is pointed out by Cairnes that the principle of international trade governs domestic industry where 'non-competing groups' exist. With reference to this case, as well as the preceding, Dr. Sidgwick justly says: 'It is not merely inconsistent with facts but with other parts of Mill's teaching, to say broadly that 'the value of things which can be increased at pleasure does not depend . . . upon demand' (Political Economy, book ii, ch. ii, § 9). In this case the value of an article is proportioned to its final utility for the purchaser in the same sense as in the preceding cases. But it is not proportioned to the difficulty of attainment in the same sense.

(4) The coordinateness of difficulty of production with utility is not even supposable, when we pass to another category, Mill's first: 'things of which it is physically impossible to increase the quantity beyond certain narrow limits;' such as 'ancient sculptures' . . . 'rare books or coins' . . . 'houses and building-ground in a town of definite extent,' and 'potentially all land whatever' (Political Economy, book iii, ch. ii, § 2).

(5) With Mill's first class go those commodities which are temporarily 'unsusceptible of increase of supply' (ibid. § 5); in short all cases of Market as distinguished from Normal Value.

(6) Lastly, all cases of monopoly must be excepted from the sphere within which the difficulty of attainment experienced by the purchaser is equateable with the difficulty production. Outside this sphere the difficulty experienced by the purchaser is due to the niggardliness of his fellow-man, rather than the stubbornness of nature; and is measured only by his own reluctance to part with some useful commodity, and not also by his (potential) effort in producing the article purchased.

D

It is easier to refine upon these logical distinctions than to prove what is the relative extent and importance of the categories defined; which conception, if any, may be taken as typical of the facts. This is a matter of judgment rather than demonstration; about which there is much disagreement between economists of the last and the present generation. The case which one treats as the general rule, another treats as exceptional or non-existent. Mill speaks of his second category as 'embracing the majority of all things that are bought and sold' (Political Economy, book iii, ch. ii, § 2). To the same effect Ricardo on the very first page of his Principles. The reservations which are here indicated under heading (1) are waived by Ricardo. Of the effect of the rate of profits on value he says, 'the reader however should remark that this cause of the variation of commodities is comparatively slight in its effects' (ibid., ch. i, § iv.) The difficulties caused by the difference in the qualities of labour he dismisses in a few sentences (ch. i). The extreme recoil from Ricardo's position is marked by the Austrian School, who emphasize utility as the determining principle of value, and assign quite a secondary place to Cost. See especially Professor Wieser, Ueber den Ursprung . . . des wirthschaftlichen Werths; and Dr. Böhm-Bawerk, Kapital und Kapitalzins, interpreted by Mr James Bonar in the Quarterly Journal of Economics, October 1888, January 1889. In this attitude they had been anticipated by Jevons. But Jevons, as has been shown, admitted cost of production as a simultaneous factor. The simultaneousness of the two conditions in a regime of industrial competition has been defended by the present writer in the Revue d'Économie Politique for October 1890. In fine there are those who regard all abstract theory as futile. Cliffe Leslie, Adolf von Held, Brentano and others, harp on the unreality of the Ricardian

assumptions. Neumann's article on prices in Schönberg's Handbuch teems with cases which it is difficult to reconcile with any theory of the relation between value and difficulty of attainment.

Bibliography

von Böhm-Bawerk, E. 1884–9. *Kapital und Kapitalzins.* Innsbruck: Wagner.

Marshall, A. (With Mary Paley.) 1871. *The economics of industry.* London: Macmillan.

Marshall, A. 1891. *Principles of economics*, 2nd ed. London: Macmillan.

Mill, J.S. 1843. *A system of logic*. London: J.W. Parker.

Mill, J.S. 1848. *Principles of political economy.* London: J.W. Parker.

Ricardo, D. 1817. *Principles of political economy and taxation*. London: J. Murray.

Sidgwick, H. 1883. *Principles of political economy.* London: Macmillan.

Venn, J. 1884. *The principles of empirical or inductive logic*. London/New York: Macmillan.

Walras, L. 1874–7. *Eléments d'économie politique pure.* Lausanne: Corbaz.

Wieser, F.F.B. 1884. Ueber den Ursprung ... des Wirthschaftlichen Werthes. Vienna: Holder.

Difficulty or Facility of Production

John Eatwell

The materialist view of the world characteristic of the writings of the classical economists was manifest not only in their concern with production and accumulation, but also in their theories of value. Petty and Cantillon, for example, both argued that the value of commodities is determined by the *amounts* of land and labour used in their production. Smith dropped land from the calculation, and argued that the value of commodities is determined by the quantity of labour used to bring them to market (at least in the early and rude state of society), though this material approach was somewhat blurred by reference to the 'toil and trouble' involved.

In the writings of Ricardo, however, the link between the material conditions of production and the value of commodities is both clear and prominent. In the Essay on Profits, prior to the formulation of his theory of value, Ricardo argued that

> wherever competition can have its full effect, and the production of that commodity be not limited by nature, as in the case with some wines, the difficulty or facility of their production will ultimately regulate their exchangeable value. (1815, p. 60)

To this proposition he appended a footnote:

> Though the price of all commodities is ultimately regulated by, and is always tending to, the cost of their production, including the general profits of stock, they are all subject, and perhaps corn more than most others, to an accidental price, proceeding from temporary causes. (1815, p. 60n)

The equation between 'difficulty or facility of production' and the 'cost of production', or value, of commodities remained a dominant theme in Ricardo's treatment of value in the three editions of the Principles of Political Economy and Taxation (1817, 1819, 1821) and in the papers on Absolute Value and Exchangeable Value, written in 1823 in the last few months of his life. In the draft version of these papers he declared, 'to me it appears a contradiction to say a thing has increased in natural value while it continues to be produced under precisely the same conditions as before' (1823, p. 375).

For Ricardo, difficulty of production referred only to the produced means of production and the labour required to produce a commodity. Non-produced means of production, such as the services of land, are not included. The limited availability of fertile land will be manifest in the extent to which more commodities and/or labour may be required to produce a further unit of output; i.e. the extent to which the difficulty of production will be increased.

But whilst the notion of difficulty of production is intuitively clear, it is not at all obvious how it may be represented as a single magnitude and so related to the exchangeable value of commodities. It was this latter relationship which was to be the source of the considerable difficulties which Ricardo encountered in the formalization of his theory of value and distribution, in particular once the influence of changes in distribution on exchange value was taken into account.

The representation of difficulty of production as a single magnitude is possible only if that magnitude is the quantity of labour embodied directly and indirectly in the production of the commodity. Changes in this quantity can derive only from changes in the technology – where by technology is meant the produced means of production and the labour used in total in the production of a commodity, i.e. the means of production and labour of the integrated sub-system which would (hypothetically) have as its net product one unit of the commodity in question (Sraffa 1960, Appendix A). So Ricardo's adoption of the labour theory of value was a natural outcome of his materialist view of economic relations, allowing him to move freely from material conditions of production to rates of exchange, and from material net product to the general rate of profit.

Yet it was exactly Ricardo's material conception of cost which exacerbated the contradictions which emerge once the influence of changes in distribution upon exchangeable value is considered. In 1823 he commented sadly that 'the increased or diminished facility of producing them' was 'by far the greatest cause' of variation in the exchangeable value of commodities, though 'it is not strictly the only one' (1823, p. 367). The focus on variation in the value of commodities, rather than the difference between labour values and natural prices, precipitated the fruitless search for an invariable standard of value as a means of tying variations in prices to variations in the difficulty or facility of production alone.

These difficulties notwithstanding (and the story of their resolution may be followed in Garegnani 1984) Ricardo's persistent use of the idea of difficulty of production is indicative of his materialist conception of political economy. The variables from which his theory of value and distribution is constructed are objective (in the sense that they are all, in principle, directly observable and measurable), being the empirical description of the process of production and the real wage determined by the concrete institutional characteristics of economic society. And Ricardo's link between conditions of production and value is neither misplaced nor archaic – the dominant explanation of changes in relative values in modern economies is, surely, the differential rates of technological progress as between the production processes of different commodities, i.e. changes in the difficulty or facility of production.

See Also

- ▶ British Classical Economics
- ▶ Cost of Production

Bibliography

Garegnani, P. 1984. Value and distribution in the classical economists and in Marx. *Oxford Economic Papers* 36(2): 291–325.

Ricardo, D. 1815. An essay on the influence of a low price of corn on the profits of stock. Reprinted in *The works and correspondence of David Ricardo,* ed. P. Sraffa, vol. IV. Cambridge: Cambridge University Press, 1951.

Ricardo, D. 1823. Notes on 'absolute value and exchangeable value'. Reprinted in *The works and correspondence of David Ricardo,* ed. P. Sraffa, vol. IV. Cambridge: Cambridge University Press, 1951.

Sraffa, P. 1960. *Production of commodities by means of commodities*. Cambridge: Cambridge University Press.

Diffusion of Agricultural Technology

David Zilberman

Abstract

A high rate of technological change is a major feature of modern agriculture. New technologies are introduced gradually; diffusion is the process through which technologies spread throughout the farm sector over time. While adoption is the decision by an individual producer to use a new technology at a given moment, diffusion is the aggregate measure of adoption decisions. Early studies of diffusion were conducted by sociologists. Rogers (1962) measured technology usage as a fraction of farmers that had adopted a certain technology at a given point in time. Other studies

measured diffusion by the fraction of land employed with the new technology. Rogers noticed that diffusion rates of hybrid corn in the United States fit very well as an S-shaped function of time:

Keywords

Diffusion of agricultural technology; Diffusion of technology; Discrete-choice estimation; Green Revolution; Imitation model of technology diffusion; Technical change; Technology treadmill; Threshold model of technology diffusion

JEL Classifications

Q1

A high rate of technological change is a major feature of modern agriculture. New technologies are introduced gradually; diffusion is the process through which technologies spread throughout the farm sector over time. While adoption is the decision by an individual producer to use a new technology at a given moment, diffusion is the aggregate measure of adoption decisions. Early studies of diffusion were conducted by sociologists. Rogers (1962) measured technology usage as a fraction of farmers that had adopted a certain technology at a given point in time. Other studies measured diffusion by the fraction of land employed with the new technology. Rogers noticed that diffusion rates of hybrid corn in the United States fit very well as an S-shaped function of time:

$$S_t = \frac{K}{1 + \exp^{-(a+bt)}},$$

where S_t is the level of diffusion at time t, K is the diffusion level at the limit and $K \leq 1$, a is a measure of initial diffusion, and b is a measure of the speed of diffusion. Rogers modeled diffusion as a process of imitation. In the early and late stages of diffusion, the level of diffusion is low because either the potential population of adopters or the population of users of the new technology to be imitated is small. During the middle period the diffusion rate takes off as there is a sufficient number of potential adopters, as well as a large population of established users to imitate. Rogers (1962) emphasized the role of distance from urban centres in explaining diffusion, finding that villages closer to urban centres had higher coefficients of diffusion.

Griliches (1957) argued that diffusion is an economic phenomenon and showed, using the diffusion data for hybrid corn in Iowa, that the three parameters K, a, and b are affected by profit. Other studies also found that the rate of diffusion tends to increase with farm size and the education of the farmer. However, as the review of Feder et al. (1985) suggests, the imitation model lacks a microeconomic foundation. An alternative model, the threshold model, suggests that the population of potential adopters is heterogeneous, and at every moment there is a critical variable that distinguishes between them. At every moment there is a threshold level of this variable that separates adopters from non-adopters.

Threshold models have three components: microeconomic behaviour, sources of heterogeneity, and a dynamic factor that drives the threshold level up or down. For example, adoption of mechanical innovation reflects the maximization of discounted net benefit. Farms vary in size, and at each moment there is a farm size threshold that distinguishes adopters from non-adopters. Over time, the cost of machinery may go down due to learning by doing, or the gain from adoption may go up because of learning by using, and that will reduce the adoption threshold. Empirical models, based on cross-sections of adopters, use discrete-choice estimation techniques to identify the key sources of heterogeneity. They found in many cases that size increases adoption of mechanical innovation, education explains adoption of more complex crops, and modern irrigation technologies that actually augment land quality are adopted earlier on lower-quality lands.

Much of the research has attempted to explain the diffusion of new 'Green Revolution' varieties in developing countries. In those cases, adoption was often partial (meaning farmers switched only

a portion of their crops to the new technologies), and adoption rates were sometimes low, even given the significantly higher yields of Green-Revolution varieties. These facts emphasize the importance of risk considerations in explaining diffusion processes. Land allocation choices of riskaverse farmers were modelled as a portfolio, leading farmers to consider partial adoption of modern varieties because of their increased vulnerability to variable weather conditions. In addition to risk, wealth, human capital, and physical conditions, institutional forces have been identified as major determinants of diffusion rates. For example, renters are less likely to adopt new innovations than owners, especially when the rental contract is short. Lack of availability of credit is another deterrent to adoption. On the other hand, government policies, in the forms of output price subsidies and extension services that reduce the fixed costs of adoption, as well as technology and credit subsidies, can enhance the diffusion of modern agricultural technologies. For irrigation technologies, subsidies of water combined with restrictive trading regulations slow the diffusion of improved irrigation practices; water conservation can be enhanced by reducing constraints on water trading.

When demand for agricultural products is inelastic, the main beneficiary of the diffusion of more efficient technology is the consumer, while farmers are stuck on a 'technology treadmill'. Early adopters also benefit from the introduction of the new technology, but followers, who make up the majority of the farm population, may adopt only to stay competitive, while sometimes the laggards may go out of business. When the demand for agricultural products is elastic, then the gain from adoption of modern technologies contributes to enhanced land values, but the individual farm operators may not gain significantly because of the technology treadmill effect.

See Also

- ▶ Agricultural Research
- ▶ Diffusion of Technology

Bibliography

Feder, G., R.E. Just, and D. Zilberman. 1985. Adoption of agricultural innovations in developing countries: A survey. *Economic Development and Cultural Change* 32: 255–298.

Griliches, Z. 1957. Hybrid corn: An exploration in the economics of technological change. *Econometrica* 25: 501–522.

Rogers, E.M. 1962. *Diffusion of innovation*. New York: Free Press.

Diffusion of Technology

Wolfgang Keller

Abstract

The diffusion of technology has a major impact on per-capita income. Moreover, international convergence turns on whether technology diffusion is local or global. This article characterizes the creation of technological knowledge and discusses the primary determinants of diffusion. It is shown that even today technology diffusion is to an important degree local, allowing for many technological knowledge levels in the world to coexist. This article focuses on the data and empirical methods employed in the estimation of diffusion patterns.

Keywords

Asymmetric information; Control functions; Diffusion of technology; Economic growth; Human capital; Instrumental variables; Intermediate goods; Knowledge spillovers; Monopoly; Neoclassical growth theory; Patents; Product market competition; Research and development; Tacit knowledge; Technical change; Technology spillovers; Total factor productivity

JEL Classifications

O3

The technology of a firm or country determines the efficiency with which inputs are mapped into outputs. Technological change may result in the ability to produce entirely new products, or it may allow an existing product to be produced with fewer inputs. This process has long been viewed as central to economic growth. The question of whether or not there is convergence across firms and countries raises issues related not only to the process of technical change but also to the diffusion of technology. Beginning in the late 1950s, economists have formalized their thinking as to how such technological knowledge diffuses from one economic entity to another. The early efforts were primarily directed to understanding firms' technology adoption decisions that often yield an S-shaped diffusion pattern over time. Since the 1990s, a vibrant literature has emerged in which the issues addressed are considerably broader, and where much more emphasis is placed on seeking high-quality empirical evidence.

A firm's technology and its productivity are closely related, and the two are identical if technology is identified with total factor productivity, an approach frequently adopted since the 1950s. The development of models of endogenous technical change in the early 1990s represented a step forward in that the R&D resources devoted to innovation were separated from the new technological knowledge itself. For example, consider the technology production function

$$\dot{N} = \eta N^{\lambda} H_N, \tag{1}$$

where η and λ are parameters, $\eta, \lambda > 0$. The term H_N denotes the skilled-labour resources devoted the R&D, which according to Eq. 1 lead to a flow of new technological knowledge of $\dot{N}$. A higher level of R&D produces a higher level of technology, N, and that in turn can be shown to result in higher productivity.

According to Eq. 1, a higher stock of existing technological knowledge facilitates innovation. This stock of technological knowledge will rarely be entirely self-produced, so that (1) typically involves the diffusion of technology – diffusion between different persons, firms or countries. Technology is sometimes purchased or licensed in a market transaction, but, due to asymmetric information and other problems in the market for technology, non-market transactions in the form of technological externalities, called knowledge spillovers, are much more important.

What are the nature and the size of these knowledge spillovers? Since technological knowledge is non-rival, such externalities can in principle benefit many economic agents.

A useful benchmark is the complete diffusion of technology, which describes the case where technological knowledge created anywhere in the world is available worldwide immediately. This could underlie the assumptions of common-to-all and free technological knowledge of neoclassical growth theory. Clearly, this is not true in reality, where the diffusion of technology is gradual and uneven.

Why? First of all, acquiring technology involves making complementary investments, and the equilibrium choice for such investments often implies that not all technology diffuses. For instance, in Keller's (1996) model, international trade enables domestic producers to raise productivity by importing specialized foreign intermediate goods. Since these goods embody foreign R&D investments, this means the diffusion of technology from one country to another. For this imported technological knowledge to trigger domestic innovation, however, additional investments are necessary. According to Keller (1996), these investments mean additional training of workers so that they have the skills to manufacture products according to new blueprints. In addition, domestic innovators may have to invest resources in reverse engineering the foreign intermediate goods in order to fully comprehend the underlying foreign technological knowledge.

Second, another major determinant of the firm's decision to acquire the existing technology and innovate is the degree of product market competition. For example, in early Schumpeterian endogenous growth models, a higher degree of product market competition leads to lower monopoly profits and thus to a lower rate of innovation. More recent work by Aghion et al. (2001), for example, shows that, if technological laggards must first catch up with the leading-edge

technology before battling for technological leadership in the future, the overall effect of more product market competition may be positive. The reason for this is that, even though more competition means lower monopoly profits, technological leaders now also have an incentive to innovate to avoid competition with technological laggards, and, if the latter effect is strong enough, product competition has a positive effect on technology diffusion and growth.

Third, there is no complete diffusion because it is simply not in the interest of the original creator of the technology, since his market for the technology would shrink if there were additional suppliers. In some cases, innovators obtain a patent that provides government-sanctioned protection of economic interests for a limited period of time in exchange for release of the technological information. Another strategy on the part of the original innovator is to use a varying amount of resources to keep the technological knowledge secret. At the same time, studies show that it often is no more than two years until new technology becomes publicly available.

Another, probably the most important, reason why knowledge spillovers are limited is that only the broad outlines of technology are codified – the remainder is the 'tacit' part of the knowledge. A person who is engaged in a problem-solving activity can often not fully define (and hence prescribe) what exactly he or she is doing. Along these lines, technology is only partially codified because it is impossible or at least very costly to fully codify it. For technology diffusion to occur completely, it may be necessary that the person who learns about the new technology can observe another person in the process of applying the technology. Even if this can be dispensed with, person-to-person contacts will generally be beneficial to the diffusion of technology.

Research has now turned to the essential task of assessing the importance of these processes empirically. As an intangible, technology is intrinsically difficult to measure, and economic data is hard to come by. This is even more the case for the non-market effects caused by technological knowledge. The main approach for quantifying technical change has been to study the relationship between R&D investments and productivity (Griliches 1979). For example, Keller (2002a) estimates

$$tf \quad p_{it} = \beta s_{it} + X'\gamma + \varepsilon_{it}, i = 1, \ldots I, \text{and}$$
$$t = 1, \ldots, T, \tag{2}$$

where tfp_{it} is log total factor productivity in industry i at time t, s_{it} are industry i's cumulative R&D investments (in logs) in period t, X is a vector of other observed determinants of productivity, and the error ε_{it} picks up unobserved effects. The parameter β, estimated in Keller (2002a) at $\beta = 0.15$, measures how R&D investments translate into higher productivity, thereby implicitly capturing the rate of technical change.

This approach is attractive since R&D spending is the main cause of technical change, and data on R&D expenditures is relatively easy to collect and compare across units (firms, industries and countries). A drawback is that measuring technical change this way requires an estimate of β. This can be complicated if productivity is badly measured, R&D is endogenous, or unobserved determinants on productivity are important, as in practice is often the case. Applications of instrumental-variable and control-function approaches have shown much promise in addressing the major estimation concerns (see Gong and Keller 2003). Patents are an alternative measure of technology, with the advantage that patent data is available for a broader set of countries and a longer time horizon than is data on R&D (Jaffe and Trajtenberg 2002). While patent counts are an imperfect measure of technology because the distribution of patent values is extremely skewed, recent work using citations-weighted patent data has addressed this point since citations of a particular patent are a plausible indicator of its value. At the same time, patents cannot capture more than the codified part of technological knowledge, apart from the fact that across industries and firms the prevalence of patenting varies strongly for reasons that are difficult to fully ascertain.

Technology spillovers, as the major form of technology diffusion, are mainly analysed by

extending Eq. 2 above to estimate as well the effects of R&D investments conducted elsewhere. For example, in addition to the effects of own-industry R&D, Keller (2002a) estimates the effects of R&D in other domestic industries $\left(S_{it}^{do}\right)$, as well as those of R&D in the same and other foreign industries $\left(S_{it}^{f}\right)$ and $\left(S_{it}^{fo}\right)$, respectively):

$$tf\ \ p_{it} = \beta_{1S_{it}} + \beta_2 S_{it}^{do} + \beta_3 S_{it}^{f} + \beta_4 S_{it}^{fo} + \tilde{X}'\gamma + \varepsilon_{it}. \quad (3)$$

In this framework, the estimates of β_1, β_2, β_3, and β_4 determine the relative strength of intra- and inter-industry, and of domestic and international technology diffusion. For his sample of eight large Organisation for Economic Co-operation and Development (OECD) countries, Keller (2002a) finds that intra-industry effects dominate inter-industry spillovers, and that about 25 per cent of the total effect is due to international technology diffusion.

Other interesting approaches have employed multi-country extensions of recent models of endogenous technical change that include international technology diffusion (Eaton and Kortum 1999). Because here the economic environment is fully specified, it is straightforward to simulate a model and perform interesting policy experiments. At the same time, typically there is little data on technology diffusion employed in the econometric estimation of these models. Consequently, the model's structure has a great influence on the results, while the implications for the diffusion of technology are not clear.

One major finding has been that the diffusion of technology is geographically localized, both domestically and internationally. For example, Keller (2002b) studies international technology diffusion between the G-5 countries (the United States, Japan, Germany, France, and England) and nine smaller OECD countries by estimating

$$tf\ \ p_{it} = \beta\left[s_{it} + \sum_{j \in G5} \exp\left(-\delta Dist_{ij}\right) s_{jt}\right] + X'\gamma + \varepsilon_{it}, i = 1, \ldots, I, \text{ and } t = 1, \ldots, T. \quad (4)$$

Here, $Dist_{ij}$ is the geographic distance between country i and G-5 country j. The parameter δ determines the extent of geographic localization: the higher is δ, the stronger is the degree of the localization of technological knowledge, while if $\delta = 0$, international technology diffusion is complete in the sense that geography has no impact whatsoever. The geographic reach of technology spillovers is a critical determinant of the cross-country income distribution, since global spillovers favour income convergence while local spillovers lead to income divergence. Keller's (2002b) results for the years 1970 to 1995 strongly reject the null hypothesis of complete diffusion. Instead, he estimates that with every additional 1,200 kilometres there is a 50 per cent drop in technology diffusion. The results imply that the benefits of being located next to major technology producers are substantial, highlighting the danger for isolated areas of being left behind.

While distance still shapes technology diffusion in a major way, there is also evidence that between 1970 and 1995 geography's grip on technology diffusion has weakened. Keller (2002b) estimates that the size of the δ parameter in Eq. 4 fell substantially from the late 1970s to the 1990s, consistent with the idea that innovations in information and communication technologies have led to a major improvement in technology diffusion.

Such improvements in countries' abilities to draw on international innovations also imply that increasingly the ultimate sources of domestic productivity growth lie abroad. This is especially true for medium-sized and small countries, where the contribution of foreign technology to domestic productivity growth often exceeds 90 per cent. At the same time, because successful technology diffusion requires complementary investments in terms of adaptive R&D and/or human capital, domestic activities have a significant impact on the ease of technology diffusion.

See Also

► Transfer of Technology

Bibliography

Aghion, P., J. Harris, P. Howitt, and J. Vickers. 2001. Competition, imitation and growth with step-by-step innovation. *Review of Economic Studies* 68: 467–492.

Eaton, J., and S. Kortum. 1999. International technology diffusion: Theory and measurement. *International Economic Review* 40: 537–570.

Gong, G., and W. Keller. 2003. Convergence and polarization in global income levels: A review of recent results on the role of international technology diffusion. *Research Policy* 32: 1055–1079.

Griliches, Z. 1979. Issues in assessing the contribution of research and development to productivity growth. *Bell Journal of Economics* 10: 92–116.

Jaffe, A., and M. Trajtenberg. 2002. *Patents, citations, and innovations: A window in the knowledge economy.* Cambridge, MA: MIT Press.

Keller, W. 1996. Absorptive capacity: On the creation and acquisition of technology in development. *Journal of Development Economics* 49: 199–227.

Keller, W. 2002a. Trade and the transmission of technology. *Journal of Economic Growth* 7: 5–24.

Keller, W. 2002b. Geographic localization of international technology diffusion. *American Economic Review* 92: 120–142.

Digital Marketplaces

Andrey Fradkin
MIT Sloan School of Management, Cambridge, MA, USA

Abstract

Digital marketplaces represent a new and important organizational form in the economy. Since their emergence in the mid 1990's, they have transformed many industries including accommodation, retail, and transportation. I start this entry by outlining the ways in which digital marketplaces differ from traditional firms. I then discuss three research areas relating to digital marketplaces. The first research area concerns the determinants of marketplace diffusion and its effects. The second concerns the economics of digital market design, with an emphasis on search and matching, pricing, and trust and safety mechanisms. The last research area is about policy issues prompted by digital marketplaces. I conclude by discussing new research topics relating to emerging technologies and continued marketplace growth.

Keywords

Industrial organization; Digital marketplaces; Peer-to-peer markets; Search and matching; Market design; Marketplace design; Digitization; Structural change; Sharing economy; Reputation systems

JEL Classification

J2; J4; L1; L15; L2; L5; L86; M13

Introduction

Digital marketplaces represent a new and important organizational form in the economy. They have enabled new types of transactions such as online auctions, ride-sharing, and home-sharing, have grown to dominate verticals such as travel and books, and have continued spreading through other industries. The growth and increasing importance of these marketplaces has prompted a new and growing body of research. In this entry, I summarize this research, divided into three areas: effects on the economy, market design, and policy implications.

There is no universally accepted definition of an online marketplace. Delineating between online marketplaces and traditional firms is becoming harder as more firms are embracing technology. One way to do this delineation is to make a list of companies that could be considered digital marketplaces and consider what they have in common. A non-comprehensive list would include Airbnb, Alibaba, Amazon Marketplace, Craigslist, eBay, Expedia, Uber, and Upwork. Although these firms serve different verticals and use a variety of market designs, they share certain characteristics. I list these below both to highlight their salient features and to limit the scope of this entry.

- **Digital Matching:** The process of search and matching between buyer and seller occurs digitally via a browser, app, or text interface. Digital interfaces allow for a precise tracking of the actions of users, which enables algorithmic matching. This is in contrast to older, relatively information poor shopping interfaces such as physical stores or mail-order catalogs.
- **Low Entry Costs:** A variety of sellers are allowed to participate in the platform and the entry costs are typically low. That means that non-professional sellers such as hobbyist collectors on eBay can compete with large firms such as Target.
- **Ex-post Screening:** A significant share of the screening is conducted ex-post, through explicit or implicit feedback given by users regarding transaction quality. Ex-post screening usually involves online reviews but can also include data on user engagement and customer service complaints.
- **Non-exclusive and Short-Run Contracts:** Sellers are not obliged to exclusively use a particular platform, do not engage in long-term employment relationships, and retain at least some control rights over their product.
- **Direct Transactions:** The money paid by the buyer is transferred at least partially to the seller. This excludes other digital intermediaries such as streaming platforms (Netflix and Spotify), dating sites (Tinder and Okcupid), and advertising platforms (Google and Facebook) which have similarities to digital marketplaces.

These characteristics allow the concept of a digital marketplace to encompass and subsume a variety of terms commonly used in research and public discourse such as 'peer-to-peer', 'the sharing economy', and 'the on-demand economy'. What distinguishes the marketplace from a re-seller such as Macy's or Zappos.com is that at least some of the control rights regarding pricing, advertising, customer service, and order fulfilment remain with the seller (Hagiu and Wright 2014). This means that the marketplace serves as an aggregator and matchmaker of heterogeneous and autonomous buyers and sellers, even if in some cases the marketplace does participate as a buyer or seller in its own market.

Digital marketplaces pose new research questions and challenges for economists. First, the growth of these firms has affected consumers, workers, and firms. These effects have prompted new debates about the proper role of these firms in society. Much of this debate has happened without rigorous analysis and the economics profession is just catching up.

Second, the design of these marketplaces poses new challenges. Technology has enabled market mechanisms including reputation systems, search engines, algorithmic recommendations, and signalling mechanisms. The choice of the proper design is important because of the scale of these companies. A single design change in a major marketplace can affect hundreds of millions of consumers and millions of sellers. Due to the complexity of these markets, design decisions often have unintended consequences such as altering the distribution of income on the platform or making it easier for users to discriminate.

Lastly, these marketplaces have also drawn attention from regulators. The diverse policy issues relating to these companies include antitrust, licensing, labour practices, data sharing and privacy, and discrimination. These topics will become even more salient as more of the economy becomes intermediated by these firms.

The rest of the entry discusses these three topics in detail.

The Causes and Consequences of Digital Marketplace Diffusion

The share of digital transactions varies greatly across industries, locations, and over time. In the retail sector, for example, books and magazines had a 44% digital market share in 2014. In contrast, digital purchases of clothing, accessories, and footwear had a 15% market share and drugs, health, and beauty had just 4.7% of market share (Hortaçsu and Syverson 2015). Even within a given marketplace, growth varies over time and across cities as shown by Farronato and Fradkin (2017) for Airbnb, Hall and Krueger (2015) for

Uber, and Cullen and Farronato (2016) for Taskrabbit, a marketplace for local services. A theory of digital marketplaces must explain why a transaction occurs digitally rather than at a physical location, why the transaction occurs on a marketplace rather than directly with a seller or pure reseller, and why the consumer chooses a particular marketplace (e.g. eBay vs. Amazon or Airbnb vs Booking.com).

The Causes of Digital Marketplace Diffusion

The consumer's choice between a digital and a physical transaction is governed by the benefits of examining a good in person, the relative hassle costs between search online and offline, the benefit of instant product availability offline, differences in assortment, and regulation. In a world where digital devices are ubiquitous, transactions with no in-person component, such as flight or hotel purchases, will naturally take place digitally. On the other hand, purchases like furniture, where examining a good in person is valuable, face a large hurdle to occurring online. Digital distribution also affects the cost side, most obviously due to firms no longer needing a physical retail presence.

Importantly, the attractiveness of digital transactions is endogenous and dynamic because firms can invest in services, market designs, and technologies such as same-day delivery, insurance, matching mechanisms, and customer service to enable new types of digital transactions. To the extent that there are returns to scale in these activities, firms which intermediate large volumes of transactions in equilibrium have the greatest incentive to make these investments and to conduct research. This also means that the growth in digital transactions can be driven by innovation spurred by competition between marketplaces. These innovations are discussed in the next section of this entry.

The growth of digital marketplaces directly affects consumers, firm owners, and workers. The magnitude and sign of these effects will vary depending on whether agents are associated with traditional firms (e.g. hotels), re-sellers (e.g. Barnes & Noble), new entrants (e.g. Airbnb hosts or eBay sellers), or the intermediaries themselves. Furthermore, there may be spillovers to seemingly unrelated markets (e.g. the housing market) or externalities (e.g. traffic congestion or noise) due to these new transactions.

The Effects of Digital Marketplace Diffusion

Contributions to the literature on the effects of marketplaces can typically be divided into theoretical papers, empirical work which tests predictions, and structural estimation. As an illustrative example, I discuss Farronato and Fradkin (2017), which combines all three of these approaches to theoretically and empirically study the direct effects of the growth of Airbnb on the accommodations industry in the United States. In their theoretical model, a market consists of a day and city. Consumers enter the market and can choose between Airbnb hosts (peers), traditional hotels, and an outside option. The role of the marketplace in this model is to lower the entry and marginal costs of peers and to increase demand for these peers. Consequently, the marketplace increases the competitiveness of peer sellers and increases the assortment of options available to consumers.

A consumer's choice between options is determined both by the extent to which peer hosts offer a differentiated product and by the price of peer supply relative to traditional hotels. On the supply side, traditional hotels have relatively high fixed costs due to the fact that it takes time and money to build a new hotel. Traditional hotels also have low marginal costs because cleaning costs and other services tend to be cheap. In contrast, peer hosts have low entry costs, which consist of signing up on the Airbnb on website. On the other hand, their marginal costs can be high due to opportunity costs (hosts typically have a traditional job) and due to risk from the fact that strangers may damage the property or cause other problems.

The above framework predicts that peer transactions will be more likely to occur in places where hotel fixed costs are higher, when peer marginal costs are lower, and when demand is higher. Farronato and Fradkin (2017) show that these predictions are borne out across major US cities between 2011 and 2014, where hotel fixed costs are proxied by undevelopable land area and building regulations; peer marginal costs are

proxied by share of unmarried adults, who have lower risks from hosting; and demand is proxied by incoming flights and Google searches for accommodations.

The implication of this framework is that the entry of the marketplace will have direct effects on three constituencies: consumers, peer firms, and traditional firms. Most research studies each of these effects separately. For example, Cohen et al. (2016) use discontinuities in Uber's pricing algorithm to estimate its consumer surplus. They find that UberX, the most commonly used service option on Uber, generated $2.9 billion in consumer surplus in four US cities in 2015. This large surplus is driven both by technology and by the fact that taxis, the traditional firms in this industry, were heavily constrained in their supply and pricing by regulations. Markets with a 'long-tail' of niche products generate benefits through a similar mechanism. Quan and Williams (2016) use transactional data to measure the size of this gain for apparel and footwear. A related mechanism is that, by allowing for increased entry and experimentation, digital markets help uncover unexpectedly high quality products and services. This mechanism is evident on the crowd-funding platform, Kickstarter, and its importance has been documented for music by Aguiar and Waldfogel (2016).

Turning to producer outcomes, e-commerce has generally been found to reduce equilibrium prices and price dispersion (see Lieber and Syverson (2012) for a summary of the literature and Ellison and Ellison (2014) for an exception in the case of niche books). Goldmanis et al. (2010) study the effect of e-commerce on physical retail sales. In their framework, the primary characteristic of digital transactions is lower search costs. Their empirical results corroborate the model predictions and show that employment falls the most at small firms with a physical presence. Cramer (2016) uses cross-city variation to study the effects of Uber's growth on the labour supply and earnings of traditional drivers and finds no effect as of 2015. This is due to the fact that Uber increases the total demand for rides and taxi drivers can also earn money on Uber. Owners of taxi medallions have been hurt due to the falling prices of taxi medallions. In contrast early investors in successful digital marketplaces have benefited given the multibillion dollar valuations of these companies.

Farronato and Fradkin (2017) estimate a structural model of equilibrium in the accommodations industry to jointly quantify the effects on consumers, peer producers (Airbnb hosts), and traditional firms. They find that consumer surplus increased due to both the fact that that Airbnb offers a differentiated product and the fact that hotels face more competitive pressure, especially in high demand periods where they would otherwise have market power. Second, traditional firms lose revenue, and this revenue loss will be driven by price adjustment in high demand periods in cities with high hotel entry costs. This prediction is also corroborated in the case of Airbnb by Zervas et al. (2015).

In the long-run, the availability of peer-to-peer accommodations should reduce the equilibrium number of traditional firms, but we do not study this effect as it is out of sample. Lastly, we find substantial dispersion in the marginal costs of Airbnb hosts and that most hosts are close to the margin of hosting. Consequently, the typical host's listing is usually not booked and hosting generates much larger benefits in high demand periods. Hall et al. (2016) find similar results for Uber drivers, who typically drive part-time and are highly responsive to price and expected utilization changes on the margin.

The difference between peers and professionals has generated a vigorous debate in the media and amongst regulators. The worry on the part of critics and regulators is that purported peers are full-time sellers who avoid regulation by using a marketplace. More generally, the decision to own or rent an asset such as an apartment is endogenous. Digital marketplaces enable assets to be utilized a higher share of the time by making renting easier for buyers and owners.

Horton and Zeckhauser (2016) study the implications of the ability to rent out assets on equilibrium asset ownership and prices. In their model, a fall in the cost of bringing an asset to market

causes owners with a relatively low expected utilization or valuation to switch to renting. On the other hand, non-owners may now rent due to the existence of a rental market. The long-run effects on total asset ownership and prices depend on the model parameters. Fraiberger and Sundararajan (2015) calibrate a model of car ownership with a peer-to-peer market and find that equilibrium asset ownership should fall.

The above discussion has treated digital marketplaces as technologies that statically affect the attractiveness of certain types of trades. In practice, marketplaces attempt to manage their growth, both in order to harness network effects and in order to pre-empt competition. This has been the stated justification for billion dollar financing rounds for companies like Uber (Sorkin 2016). White and Weyl (2016) present a theory of this decision, where the firm's expansion strategy is a function of network effects and their heterogeneity across users.

Marketplace Design

The role of a digital marketplace is to maximize its profit by facilitating matches between buyers and sellers. The value of these matches, including the cost of using the marketplace, must exceed the value of the outside option. The marketplace fulfils its role through its market design, defined broadly to include both policies and technologies. Marketplace design varies across industries, over time within an industry, and within a given marketplace. Most research suggests that design is an important factor in marketplace growth and competition.

It is useful to divide marketplace design choices into three categories. First, the marketplace chooses the process by which buyers and sellers match with each other. Second, it chooses the manner in which prices, inclusive of fees, are set. Third, it chooses mechanisms which ensure that goods or services are delivered reliably and with minimal risk. Although these areas interact with each other, I follow the literature in describing them separately.

Search and Matching

Buyers and sellers find each other in a variety of ways, including directed search, auctions, and centralized matching. The choice of mechanism often involves trade-offs between three factors: the quality of a match, the hassle costs of finding a match, and the overall balance of matches in the market.

These trade-offs are well illustrated by the differences between Uber's and Airbnb's matching mechanisms, also discussed in Einav et al. (2016). In Uber's app, consumers are algorithmically assigned a car and cannot choose specific makes and drivers. In contrast, Airbnb's search engine allows consumers to choose between all options which are not explicitly marked as unavailable. The primary reason for this difference is the relative difficulty of expressing preferences across these two markets. Conditional on pickup and drop-off location, Uber riders mostly care about wait times, which are predicted by Uber, and prices, which are set by Uber. In contrast, Airbnb guests to a given city may have different preferences over location, room characteristics, and price. It is difficult to predict the option that a guest will find most appealing and search rankings, while helpful, do not eliminate the need for extensive consumer search (Fradkin 2017).

The most common mechanism used for matching is the search engine, where searchers form a consideration set through textual search and filtering. The results shown on each page are determined according to an algorithm, which may be as simple as a reverse chronological ordering or as complex as a personalized ranking determined by a neural network. The market design for the search engine consists of the algorithm, the information presented about each option, the interface for search (including filters), and the manner in which that information is presented.

Numerous papers in economics, marketing, and computer science have studied search ranking. A full summary of this literature is beyond the scope of this entry but several lessons stand out. On the theoretical side, the structure of the search process affects equilibrium outcomes such as price dispersion (e.g. Baye and Morgan 2001),

and intermediaries may have an incentive to divert search away from the social optimum (e.g. Hagiu and Jullien 2011). On the empirical side, much of the literature has found that ranking matters (e.g. Ursu 2016 and references), that changes in algorithms can improve match rates in these settings (e.g. Fradkin 2017), and that there is substantial heterogeneity in the effects of ranking (Goldman and Rao 2014). There is also an entire field of computer science focused on designing recommender systems (see Jannach et al. 2016 for a recent overview).

From the perspective of the marketplace, the key choices regarding algorithms are which objective function to maximize and which information to use.

One may naively think that rankings should be determined according to a prediction of the consumer's expected utility. However, this ignores several complicating factors. First, in two-sided markets, the other side of the market may also have preferences. For example, Fradkin (2017) and Horton (2016) show that rejections of searchers occur on both Airbnb and Upwork, a business services marketplace created through the merger of Odesk and Elance, and that these rejections cause searchers to leave the marketplace. Therefore, the ability of the search engine to filter out bad matches is critical for the marketplace to compete with the outside option. Second, rankings may have equilibrium effects on congestion, available options, and other outcomes, which the marketplace should try to account for. Third, alternative objective functions may be desired if there is uncertainty about user preferences, if rankings serve as incentive mechanisms for sellers, or if rankings help the marketplace learn. Lastly, much effort by ranking algorithm engineers goes into generating 'signals' to input into the algorithm, but incorporating certain signals may be costly from an engineering perspective and may raise privacy concerns.

Information regarding users provides a complementary role to the ranking algorithm. Lewis (2011) studies information and disclosure costs for car auctions on eBay. He shows that the information displayed in photos and text affects equilibrium prices and that reductions to disclosure costs increase the information provided in the market and equilibrium prices. Tadelis and Zettelmeyer (2015) use a field experiment to show how the provision of information in the market can increase prices even for low quality goods, which see an increase in demand due to the reduction in quality uncertainty. Data on historical transaction volume and online reviews is also ubiquitous in digital marketplaces and will be covered later in the entry.

The design of the filtering and sorting interface in a marketplace also affects market outcomes. The managers of digital marketplaces consider design important and employ well compensated user experience designers to create these interfaces. Much of their work involves devising visual cues to users that make the interface easy to understand and convenient to use. Other design decisions involve the dimensions on which users are allowed to search. Fradkin (2017) notes that Craigslist's search engine in 2005 did not let users filter for short-term rentals based on trip dates, that there were no standardized prices, and that the geography filter was inaccurate. In contrast, Airbnb searchers in 2014 used trip date filters, price filters, and map filters over 50% of the time. I estimate a model of choice amongst a set of search results and show that with a random ranking rather than the actually seen consideration set, searchers would be 68% less likely to find a suitable option. Relatedly, Chen and Yao (2016) estimate a model of search on a travel site and use it to show that filters (called 'refinements' by the authors) increase the utility of products by 17%.

Both Uber and Airbnb are two-sided markets, where both buyers and sellers have heterogeneous preferences over potential transactions. A simple form of preference heterogeneity in many markets occurs due to the limited capacity of firms. Uber drivers and Airbnb hosts can only service one trip at a time. Consequently, there needs to be a mechanism that allows the seller to signal preferences, which include availability. Otherwise, searching users will be rejected from seemingly good matches.

Both Uber and Airbnb solve at least part of the availability problem by operating a payments platform, which gives them data on bookings as

they happen. In contrast, Homeaway, traditionally a marketplace for vacation rentals, has historically operated based on a pay to list model and was consequently unable to track bookings in real time. Furthermore, even on Airbnb, peer hosts do not always signal to the platform when they are unavailable. There are other reasons why sellers may reject. For example, an Uber driver may not like the destination of a trip or an Airbnb host may not want guests with no reviews. Users may also discriminate against certain ethnicities or nationalities (e.g. Doleac and Stein 2013; Ge et al. 2016; Edelman et al. 2016).

Sellers who reject buyers create an externality for the platform because buyers do not like being rejected. Romanyuk (2017) theoretically shows how the platform can coarsen the information set of sellers in order to increase matching probabilities and welfare. This justifies the movement towards 'instant booking' and away from communication in successful digital marketplaces. Under 'instant book' systems, sellers pre-commit to a coarse set of conditions under which they will accept a buyer. This allows the marketplace to display only options which are guaranteed to accept a buyer's proposal. Other mechanisms that can alleviate these problems include capacity signalling (Horton 2016) and platform rules which punish users who reject frequently.

Lastly, marketplaces such as Amazon, eBay, and Taobao, the major Chinese retail marketplace, have developed search advertising platforms that allow sellers to bid for paid placement next to 'organic' results determined by an algorithm. Paid advertising has potentially interesting effects on market outcomes. First, and most directly, it offers another way for the marketplace to earn revenue. Second, it allows sellers with private information about the returns to high placement to signal that information in a credible manner. Third, it potentially reduces the overall quality of a user's experience. Lastly, it gives sellers and products a way to be discovered (Zhang 2017).

Pricing

From eBay's auction mechanism to Uber's surge pricing, digital technology has enabled a variety of innovative pricing mechanisms. The market design decisions regarding pricing mechanisms can be divided into three components. First, who has the right to set prices and what mechanism should be used? Second, what price should be set or recommended to the seller, conditional on a mechanism? Third, how should the marketplace generate revenue?

Moving first to the question of control rights and mechanism, the literature has identified several factors that affect who sets the price and how. The first is the relative importance of price discovery versus the hassle costs of price discovery (Einav et al. Forthcoming). A second factor determining the price mechanism is the relative informational advantage of the marketplace and the seller. If individual sellers receive more informative private signals regarding demand conditions or costs than the marketplace, then they should set prices. Third, the presence of moral hazard or spillovers can shift the optimal price setting decision (Hagiu and Wright 2016).

The auction mechanism is best in situations when demand, and consequently a good price, is uncertain. Einav et al. (Forthcoming) use eBay data to show that sellers use auctions for used goods, idiosyncratic products, and when they have less experience. They also show that demand for auctions relative to fixed price has fallen over time. This is likely to be driven by the availability of an outside option (Amazon) for consumers where prices are fixed and have the reputation for being low. Given that auctions take cognitive effort and time, consumers prefer fixed price mechanisms, all else equal. There is also a recent literature (Backus and Lewis 2016; Bodoh-Creed et al. 2016; Coey et al. 2016) examining the efficiency of various auction formats on eBay.

Auctions have proven to be a successful mechanism in other marketplaces such as Upwork, for business services, and Thumbtack, for local services. In both settings, buyers demand the fulfilment of an idiosyncratic task (e.g. interior painting or programming) and face search costs. An auction mechanism where sellers bid reduces the search costs for the buyer and allows for price and quality discovery. Furthermore, because each task is idiosyncratic, there is typically no low friction outside option for the buyer. While this

format is advantageous for the buyer, it may be unattractive to the seller. Consequently, online marketplaces have experimented with features such as reserve prices and limits on the number of bids in order to make seller participation more attractive.

Another common arrangement in marketplaces, seen on Airbnb and Etsy, leaves the pricing up to the seller. In both of these marketplaces, sellers offer idiosyncratic products and services and have significant cost heterogeneity, which may vary over time. Consequently, both marketplaces make it easy for sellers to change prices and set prices based on specific conditions (e.g. weekend vs. weekday). On the other hand, neither marketplace forces the sellers to accept pre-determined prices. One drawback of seller pricing is that sellers may choose to obfuscate relevant product prices and characteristics from consumers (Ellison and Ellison 2009).

The costs and benefits of platform mediated pricing can change over time. Advances in data collection and machine learning may make it more attractive for marketplaces to set prices. For example, Airbnb has implemented 'Pricing Tips', which suggest prices to hosts, and 'Smart Pricing', which automatically sets prices if sellers opt-in. There is an interesting incentive problem in these mechanisms because marketplaces generally have a different objective than sellers.

In other cases, as on Uber and in many lending marketplaces, the marketplace determines the price. Centralized price setting is efficient when marketplaces are better able to observe aggregate demand conditions than individual sellers, can group sellers into well-defined categories, and benefit from internalizing externalities arising from pricing decisions. For example, because Uber observes both real-time and historical user behaviour and can experiment, it can predict the demand and supply responses to changes in price at a detailed geographic and temporal level (Hall et al. 2016). Furthermore, because consumers are relatively indifferent between drivers and car makes conditional on a minimal quality threshold, Uber can set the same price for all cars in each category and location. This allows Uber to set prices in order to maximize a marketplace-wide objective function.

A final consideration is the fee structure in a marketplace. Marketplaces use a variety of fees including platform entry fees, listing fees, bidding fees, and transaction fees, which may be fixed or a percentage of the sale price. Furthermore, marketplaces also choose how a fee is spread across buyers and sellers and whether there are additional surcharges for value added services (e.g. international site visibility on eBay). There has been little theoretical or empirical work on this topic, although there are clear parallels between optimal marketplace fees and the literatures on pass-through (Weyl and Fabinger 2013), platform design (Weyl 2010), and platform competition (Rochet and Tirole 2003). Hagiu and Wright (2016) provide an analysis of optimal revenue sharing between a principal and agent where there is two-sided moral hazard. They find that the side that gets control rights over the non-contractible and transferable action, such as pricing or equipment maintenance, is typically the one that receives a larger percentage of the sale revenue. Platform fees are often obfuscated and may differ in their salience relative to the prices set by sellers. These factors can shift the optimal fee structure for behavioural reasons.

Some settings, notably local services marketplaces, face the threat of disintermediation, where buyers and sellers meet on the platform but transact off of the platform. Generating revenue while avoiding disintermediation is challenging and is a hypothesized reason for the failure of Homejoy, an 'Uber for cleaning' start-up. Other local services marketplaces such as House, Porch, and Thumbtack have avoided disintermediation by relying on bidding or ad placement fees rather than the transaction fees.

Although the economics of optimal fees is complex, an interesting fact is that many marketplaces avoid experimenting with fee structures. For example, Upwork, both in its current iteration and previous one as Odesk, has consistently kept a 20% transaction fee on contracts. This may be the result of a brand commitment to a 'fair price' or due to the difficulty of measuring the equilibrium effects of platform fees.

Reputation Systems and Other Mechanisms for Trust and Safety

A final component of marketplace design concerns ensuring that transactions are safe and reliable and convincing users that this is the case. Both buyers and sellers face risks in anonymous transactions. Sellers risk not being paid, having their assets damaged, or having to deal with an overly demanding or unpleasant buyer. Buyers face the risk of not getting the good or service that they expected to get. The typical solution to the problem of trust has been a combination of firms developing reputable brands and governments requiring that sellers comply with regulations.

Digital technology offers new mechanisms to make transactions safe and lowers the costs of existing mechanisms. A non-comprehensive list of these mechanisms includes digital reputation systems, escrow services, insurance, fraud detection algorithms, identity and credential verification, dispute resolute procedures, and customer service. I begin by describing reputation systems, which have been the most salient of the above to both users and researchers.

Reputation systems work by tracking the transactions of an agent and allowing the counterparty to rate or review the transaction after it has been completed. Much of the work regarding reputation systems has focused on determining whether reviews affect consumer demand and seller behaviour. The overwhelming consensus is that reviews do affect demand and that they reduce moral hazard on behalf of sellers (e.g. Dellarocas 2003; Cabral and Hortaçsu 2010; Luca 2013; Pallais 2014). Furthermore, the existence of marketplaces such as eBay or Airbnb seems impossible without reputation systems, suggesting that reputation systems 'work'.

That said, just because reputation systems have effects, does not mean that they are appropriately designed. One fundamental problem for any marketplace is that informative reviews are a public good because writing reviews takes effort and has the potential to trigger retaliation. A second problem concerns the best manner in which to use review information throughout the platform. Importantly, these two choices are related because the incentives of reviewers depend on how the marketplace uses those reviews.

The empirical literature on reputation system design has studied review informativeness as a sufficient statistics for its design quality. Fradkin et al. (2017) use the setting of Airbnb to study the extent to which submitted reviews accurately represent the experiences of guests and hosts. They find that approximately 70% of users submit reviews after a transaction and that public reviews typically conform with more objective metrics of transaction quality including private and anonymous ratings only seen by the platform, customer service complaints, and return rates to the platform. This suggests that even without financial incentives, reviews are informative.

That said, the reviews are not fully informative. The authors use two large-scale field experiments in Airbnb's reputation system to study sources of information loss in the review system. The first experiment studies a simultaneous reveal system proposed initially in Bolton et al. (2012). The idea behind this policy is that, in a two-sided review system, there is the potential that a negative review results in retaliatory negative review by the counterparty. A simultaneous reveal system removes this possibility by ensuring that reviews are not revealed until both parties have submitted or the submission period has expired.

Fradkin et al. (2017) evaluate such a system and show that while it does work as predicted, the overall effects are relatively small.

The other Airbnb experimental policy that they study incentivizes reviews through coupons. They show that the coupon induced reviews have lower ratings and that the explanation for this is that those with worse experiences are less likely to review. This corroborates findings by Dellarocas and Wood (2007) and Nosko and Tadelis (2015) for eBay. Cabral and Li (2014) study a similar experiment in which the seller provides a rebate for a review and show that this policy induces reviews but that these reviews are biased upward by reciprocity on behalf of buyers. Fradkin et al. (2017) also document that social reciprocity generated by communication between buyers and sellers results in upwardly biased ratings.

One potential solution to the problem of partially informative reviews is to augment or aggregate these reviews in an appropriate manner. Nosko and Tadelis (2015) show that if non-reviewers have worse experiences, then the review rate is also informative about seller quality. They demonstrate how a search algorithm can use this additional data to steer buyers towards better sellers. Other papers have studied alternative methods for eliciting, displaying, and aggregating reviews (Horton 2014; Aperjis and Johari 2010; Dai et al. 2012). Design choices also include the review prompt, whether reviews should be associated with reviewer identifies, and the types of reviews that are included in an aggregate score.

Reputation systems also face the threat of manipulation by interested parties. For example, Mayzlin et al. (2014) use differences in reputation system design across Expedia and Tripadvisor to document that hotels leave promotional reviews for themselves and fake negative reviews for competitors. One way to reduce the threat of fake reviews is to require that reviewers have a valid transaction prior to a review.

Lastly, there are a variety of other less studied trust and safety mechanisms used by marketplaces. For example, some marketplaces such as Airbnb and Uber conduct identity verification through both government issued documents and social media (e.g. ensuring a legitimate Facebook account). Other platforms such as Lyft and Thumbtack conduct formal background checks and verify professional certifications and licenses. New companies have arisen with the goal of reducing the costs of these activities. For example, Checkr offers an API for conducting verification, and Sift Science offers a service for identifying fake accounts, malicious content, and credit card fraud.

Customer service and dispute resolution are also roles undertaken by marketplaces. In the case of a bad transaction, the marketplace may compensate the buyer or seller or find them a better match for free. A reputation for having a reliable customer service operation can be an important competitive advantage. Sometimes marketplaces also offer explicit insurance contracts. For example, both Airbnb and Uber provide insurance for sellers for any property damage occurring during a transaction. Determining the importance of these mechanisms is a topic for future research.

Policy Relating to Digital Marketplaces

Do laws regarding offline transactions apply to related digital transactions and who bears the responsibility for enforcement? These dual questions unite a seemingly disparate set of policy questions about marketplaces including taxation, licensing, zoning, and discrimination. Intermediaries generally argue that they are not responsible for enforcing government regulations regarding the transactions of independent buyers and sellers. Marketplaces view enforcement as costly because assuming regulatory responsibility creates legal risk and complexity, especially when laws vary across jurisdictions. In contrast, governments often argue that intermediaries are best situated to enforce regulations because they have a comparative advantage in enforcement and because they generate value from these transactions. The observed balance between these positions depends on the economics of each regulation, the importance of each marketplace, and on idiosyncrasies in political environments.

One of the first policy issues with this flavour concerned the collection of taxes by Amazon and eBay. Sales taxes in the United States are collected at the local level. However, jurisdictions often do not have the power to collect taxes from externally located sellers. Instead, consumers are legally required to calculate and pay the appropriate tax. However, due to the lack of enforcement, many do not pay. Research by Goolsbee (2000) and Einav et al. (2014) shows that the lack of effective sales tax on online purchases provides a competitive advantage for online marketplaces relative to traditional retailers. States have, with varying degrees of success, tried to pass laws to compel major online marketplaces to collect appropriate taxes. One, as of yet unresolved question, is whether this regulatory burden constitutes a significant entry barrier for new companies.

Taxation issues are also relevant for vertical specific taxes. For example, Airbnb has traditionally not collected hotel taxes on its transactions. The argument for not collecting taxes has an additional layer of complexity in the case of Airbnb, who has argued that individual hosts who occasionally rent out a room do not necessarily engage in transactions covered by hotel taxes. Airbnb's strategy has been to offer the possibility of collecting taxes as a carrot to cities in exchange for legitimizing the Airbnb-style transaction with explicit regulation.

Another issue, especially important in services marketplaces, is whether sellers must comply with existing licensing regulations. For example, taxi drivers in many major cities must obtain a medallion and a license to drive. In contrast, Uber and Lyft have their own vetting mechanisms which involve fewer upfront costs but more ex-post monitoring through reputation systems. If there is no conceptual difference between an Uber ride and a taxi ride, then this creates a disparate regulatory burden on traditional taxi drivers. Proponents of ride-sharing make two related arguments. The first is that the ride-sharing transaction is different from a traditional taxi transaction and therefore does not fall under the same regulatory framework. The second argument is that traditional taxi regulation is a form of regulatory capture to exclude competition.

The success of ride-sharing suggests that consumers do not value traditional taxi licenses enough to continue using taxis. Similarly, consumers are willing to book on Airbnb even though most hosts do not go out of their way to follow hotel safety regulations. Other marketplaces, such as Thumbtack, verify licenses on behalf of sellers but do not require that sellers be licensed to bid for a job. They leave it up to the consumer to determine whether the service provider has the ability to do the job.

Employment regulation poses another legal grey area for marketplaces. Peer-to-peer marketplaces typically treat their sellers as independent contractors and do not provide them with benefits such as health insurance, retirement plans, or vacation. However, some share of sellers on these platforms work full-time hours (Hall and Krueger 2015). This has raised a vigorous regulatory debate regarding whether these workers are misclassified and, if not, whether new employment regulations are needed to account for gig-work (Harris and Krueger 2015). A longer run and more speculative concern is that new technology may shift the economy wide mix of jobs to alternate models, with fewer protections and benefits. Equity issues also arise in other contexts. For example, ride-sharing companies might decrease public support for public transport, which would hurt those who rely on public transport the most.

Other areas of debate include the scope of zoning laws and externalities from transactions. For example, critics of Airbnb claim that the presence of tourists hurts a neighbourhood, especially if tourists are loud or disruptive. These critics also allege that properties are being converted from long-term rentals to short-term rentals, even though zoning excludes hotels from particular city areas. However, there is still no academic research regarding the validity of these claims and whether Airbnb increases housing prices and results in evictions.

In response to this debate, some cities and Airbnb have agreed on regulatory frameworks which often cap the number of nights a listing can be rented. This type of regulation ostensibly reserves property for long-term rentals but allows individuals to make extra money by renting the place to tourists on occasion.

Another regulatory issue is digital discrimination and equity. Companies cannot compel two parties to transact with each other. At the same time, the Civil Rights Act makes it illegal for hotels and motels to discriminate based on race, colour, religion, or national origin. This raises the question of whether marketplaces are responsible for reporting and banning discriminatory sellers.

Relatedly, marketplaces can try to reduce discrimination by removing race related information, but there is a potential for such measures to backfire. For example, removing real names and user pictures may reduce overall trust in the platform.

Marketplaces also possess a variety of data that is useful in city planning and enforcing regulations. For example, if cities had data on Airbnb transactions, then they could find and leverage

fines for any violations by hosts. Data on outcomes could also be used to evaluate the effectiveness of existing regulations in ensuring service quality. However, data sharing also raises privacy concerns because both governments and third-parties could potentially abuse this data.

There are already active secondary markets for data, and there may be reasons to regulate the manner in which marketplace data can be sold. These issues are just beginning to gain policy relevance.

Conclusion

The digital marketplace represents a novel and increasingly important form of economic activity. I have discussed three aspects of the economics of these marketplaces. First, what is the effect of marketplaces on economic outcomes? Second, how should these markets be designed? Lastly, what is the appropriate regulation? By necessity, this entry only skims this complicated topic.

Digital marketplaces also have a role to play as laboratories to study economic behaviour. Detailed data on behaviour allows researchers to observe behaviour such as search, communication, pricing, and labour supply decisions with unprecedented granularity. It is also much easier to conduct experiments online (Horton et al. 2010). This creates several advantages for researchers. First, they can use prior experiments conducted by the platform in clever ways to isolate casual mechanisms. Second, researchers can help companies design experiments with both an academic and business relevance. Lastly, because digital marketplaces have low entry costs, it is relatively easy to conduct experiments on the platform even without the platform's cooperation.

In conclusion, I will briefly mention several speculative topics that may have relevance in the future. First, new technologies such as voice interfaces and the Blockchain may further affect the structure of digital businesses. In particular, the Blockchain may reduce the costs of entry and the structure of reputation systems (e.g. Catallini and Gans 2016). Second, as digital transactions become ubiquitous, companies such as Uber may be able to implement Pigouvian taxation in order to reduce congestion externalities. This could result in a more efficient transportation system. Lastly, many digital marketplaces are already large players in their respective industries. If there are substantial network effects and returns to scale, then these companies may be subject to anti-trust enforcement. These topics are sure to generate exciting research for many years to come.

See Also

- ▶ Online Platforms, Economics of
- ▶ Matching and Market Design
- ▶ Pricing on the Internet

Bibliography

Aguiar, Luis, and Joel Waldfogel. 2016. *Quality predictability and the welfare benefits from new products: Evidence from the digitization of recorded music*. National Bureau of Economic Research working paper 22675.

Aperjis, Christina, and Ramesh Johari. 2010. Optimal windows for aggregating ratings in electronic marketplaces. *Management Science* 56 (5): 864–880.

Backus, Matthew, and Gregory Lewis. 2016. *Dynamic demand estimation in auction markets*. Cambridge: National Bureau of Economic Research.

Baye, Michael R., and John Morgan. 2001. Information gatekeepers on the Internet and the competitiveness of homogeneous product markets. *American Economic Review* 91: 454–474.

Bodoh-Creed, Aaron, Joern Boehnke, and Brent Richard Hickman. 2016. How efficient are decentralized auction platforms? Manuscript.

Bolton, Gary, Ben Greiner, and Axel Ockenfels. 2012. Engineering trust: Reciprocity in the production of reputation information. *Management Science* 59 (2): 265–285.

Cabral, Luís, and Ali Hortaçsu. 2010. The dynamics of seller reputation: Evidence from eBay*. *The Journal of Industrial Economics* 58 (1): 54–78.

Cabral, Luis M.B., and Lingfang (Ivy) Li. 2014. *A dollar for your thoughts: Feedback-conditional rebates on eBay*. Social Science Research Network SSRN scholarly paper ID 2133812. Rochester.

Catalini, Christian and Gans, Joshua S., Some Simple Economics of the Blockchain (November 23, 2016). Rotman School of Management Working Paper No. 2874598; MIT Sloan Research Paper No. 5191–16. Available at SSRN: https://ssrn.com/abstract=2874598 or http://dx.doi.org/10.2139/ssrn.2874598

Chen, Yuxin, and Song Yao. 2016. Sequential search with refinement: Model and application with click-stream data. *Management Science* ePub ahead of print September 28, http://dx.doi.org/10.1287/mnsc.2016.2557

Coey, Dominic, Bradley Larsen, and Brennan Platt. 2016. *A theory of bidding dynamics and deadlines in online retail*. Cambridge: National Bureau of Economic Research.

Cohen, Peter, Robert Hahn, Jonathan Hall, Steven Levitt, and Robert Metcalfe. 2016. *Using big data to estimate consumer surplus: The case of Uber*. National Bureau of Economic Research working paper 22627.

Cramer, Judd. 2016. *Disruptive change in the taxi business: The case of Uber*. Cambridge: National Bureau of Economic Research.

Cullen, Zoe, and Chiara Farronato. 2016. Outsourcing tasks online: Matching supply and demand on peer-to-peer Internet platforms. Manuscript.

Dai, Weijia, Ginger Z. Jin, Jungmin Lee, and Michael Luca. 2012. *Optimal aggregation of consumer ratings: An application to Yelp.com*. Stanford: National Bureau of Economic Research.

Dellarocas, Chrysanthos. 2003. The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management Science* 49 (10): 1407–1424.

Dellarocas, Chrysanthos, and Charles A. Wood. 2007. The sound of silence in online feedback: Estimating trading risks in the presence of reporting bias. *Management Science* 54 (3): 460–476.

Doleac, Jennifer L., and Luke C.D. Stein. 2013. The visible hand: Race and online market outcomes. *The Economic Journal* 123 (572): F469–F492.

Edelman, Benjamin G., Michael Luca, and Dan Svirsky. 2016. *Racial discrimination in the sharing economy: Evidence from a field experiment*. Social Science Research Network SSRN scholarly paper ID 2701902. Rochester.

Einav, Liran, Dan Knoepfle, Jonathan Levin, and Neel Sundaresan. 2014. Sales taxes and Internet commerce. *American Economic Review* 104 (1): 1–26.

Einav, Liran, Chiara Farronato, and Jonathan Levin. 2016. Peer-to-peer markets. *Annual Review of Economics* 8: 615–635.

Einav, Liran, Chiara Farronato, Jonathan Levin, and Neel Sundaresan. Forthcoming. Auctions versus posted prices in online markets. *Journal of Political Economy*.

Ellison, Glenn, and Sara Fisher Ellison. 2009. Search, obfuscation, and price elasticities on the Internet. *Econometrica* 77 (2): 427–452.

Ellison, Glenn, and Sara Fisher Ellison. 2014. *Match quality, search, and the Internet market for used books*. Cambridge: Massachusetts Institute of Technology.

Farronato, Chiara, and Andrey Fradkin. 2017. *The Welfare Effects of Peer Entry in the Accommodations Market: The Case of Airbnb*. Manuscript.

Fradkin, Andrey. 2017. Search, matching, and the role of digital marketplace design in enabling trade: Evidence from Airbnb. Manuscript.

Fradkin, Andrey, Elena Grewal, David Holtz, and Matthew Pearson. 2017. The determinants of online review informativeness: Evidence from field experiments on Airbnb. Manuscript.

Fraiberger, Samuel P., and Arun Sundararajan. 2015. *Peer-to-peer rental markets in the sharing economy*. Social Science Research Network SSRN scholarly paper ID 2574337. Rochester.

Ge, Yanbo and Knittel, Christopher R. and MacKenzie, Don and Zoepf, Stephen, Racial and Gender Discrimination in Transportation Network Companies (October 2016). *NBER Working Paper No*. w22776.

Goldman, Mathew, and Justin M. Rao. 2014. Experiments as instruments: Heterogeneous position effects in sponsored search auctions. Available at SSRN 2524688.

Goldmanis, Maris, Ali Hortaçsu, Chad Syverson, and Önsel Emre. 2010. E-Commerce and the market structure of retail industries*. *The Economic Journal* 120 (545): 651–682.

Goolsbee, Austan. 2000. In a world without borders: The impact of taxes on Internet commerce. *The Quarterly Journal of Economics* 115 (2): 561–576.

Hagiu, Andrei, and Bruno Jullien. 2011. Why do intermediaries divert search? *The Rand Journal of Economics* 42 (2): 337–362.

Hagiu, Andrei, and Julian Wright. 2014. Marketplace or reseller? *Management Science* 61 (1): 184–203.

Hagiu, Andrei, and Julian Wright. 2016. *Enabling versus controlling*. Boston: Harvard Business School.

Hall, Jonathan, and Alan B. Krueger. 2015. *An analysis of the labor market for Uber's driver-partners in the United States*. Cambridge: National Bureau of Economic Research. Manuscript.

Hall, Jonathan, Cory Kendrick, and Chris Nosko. 2016. The effects of Uber's surge pricing: A case study. Manuscript.

Harris, Seth B., and Alan B. Krueger. 2015. *A proposal for modernizing labor laws for 21st century work: The "independent worker"*. Brookings Institution. Washington D.C.

Hortaçsu, Ali, and Chad Syverson. 2015. The ongoing evolution of US retail: A format tug-of-war. *The Journal of Economic Perspectives* 29 (4): 89–111.

Horton, John J. 2014. Reputation inflation in online markets. Manuscript.

Horton, John J. 2016. *Buyer Uncertainty about Seller Capacity: Causes, Consequences, and a Partial Solution*. Working paper.

Horton, John J., and Richard J. Zeckhauser. 2016. *Owning, using and renting: Some simple economics of the "sharing economy"*. Cambridge: National Bureau of Economic Research.

Horton, John, David G. Rand, and Richard J. Zeckhauser. 2010. The online laboratory: Conducting experiments in a real labor market. *Experimental Economics* 14 (3): 399–425.

Jannach, Dietmar, Paul Resnick, Alexander Tuzhilin, and Markus Zanker. 2016. Recommender systems –

Beyond matrix completion. *Communications of the ACM* 59 (11): 94–102.
Lewis, Gregory. 2011. Asymmetric information, adverse selection and online disclosure: The case of eBay motors. *The American Economic Review* 101 (4): 1535–1546.
Lieber, Ethan, and Chad Syverson. 2012. Online versus offline competition. In *Oxford handbook of the digital economy*, ed. M. Peitz and J. Waldfogel, 189–223. Oxford: Oxford University Press.
Luca, Michael. 2013. *Reviews, reputation, and revenue: The case of Yelp.com*. HBS working knowledge.
Mayzlin, Dina, Yaniv Dover, and Judith Chevalier. 2014. Promotional reviews: An empirical investigation of online review manipulation. *American Economic Review* 104 (8): 2421–2455.
Nosko, Chris, and Steven Tadelis. 2015. *The limits of reputation in platform markets: An empirical analysis and field experiment*. Cambridge: National Bureau of Economic Research.
Pallais, Amanda. 2014. Inefficient hiring in entry-level labor markets. *American Economic Review* 104 (11): 3565–3599.
Quan, Thomas W., and Kevin R. Williams. 2016. *Product variety, across-market demand heterogeneity, and the value of online retail*. New Haven: Connecticut Cowles Foundation for Research in Economics, Yale University.
Rochet, Jean-Charles, and Jean Tirole. 2003. Platform competition in two-sided markets. *Journal of the European Economic Association* 1 (4): 990–1029.
Romanyuk, Gleb. 2017. Ignorance is strength: Improving the performance of matching markets by limiting information. Manuscript.
Sorkin, Andrew Ross. 2016. Why Uber keeps raising billions. *The New York Times*.
Tadelis, Steven, and Florian Zettelmeyer. 2015. Information disclosure as a matching mechanism: Theory and evidence from a field experiment. *The American Economic Review* 105 (2): 886–905.
Ursu, Raluca Mihaela. 2016. *The power of rankings: Quantifying the effect of rankings on online consumer search and purchase decisions*. Social Science Research Network SSRN scholarly paper ID 2729325. Rochester.
Weyl, E. Glen. 2010. A price theory of multi-sided platforms. *American Economic Review* 100 (4): 1642–1672.
Weyl, E. Glen, and Michal Fabinger. 2013. Pass-through as an economic tool: Principles of incidence under imperfect competition. *Journal of Political Economy* 121 (3): 528–583.
White, Alexander, and E. Glen Weyl. 2016. *Insulated platform competition*. Social Science Research Network SSRN scholarly paper ID 1694317.
Zervas, Georgios, Davide Proserpio, and John Byers. 2015. *The rise of the sharing economy: estimating the impact of Airbnb on the hotel industry*. Social Science Research Network SSRN scholarly paper ID 2366898. Rochester.
Zhang, Hongkai. 2017. Accelerated quality discovery through sponsored search advertising in online marketplaces. Manuscript.

Dimensions of Economic Quantities

P. H. Wicksteed

A unit is a concrete magnitude selected as a standard by reference to which other magnitudes of the same kind may be compared. A derived unit is a unit determined with reference to some other unit. Thus the unit of area may be derived from the unit of length by being defined as the area of the square, erected on the unit of length. The unit of speed may be derived from the unit of length and the unit of time, by being defined as that speed at which the unit of length is traversed in the unit of time. In relation to the derived units of area and speed, the units of length and time would then be fundamental— 'fundamental' being a term correlative to 'derived'.

The theory of dimensions is concerned with 'the laws according to which derived units vary when fundamental units are changed' (Everett). A fundamental unit, together with the magnitudes of like kind referred to it, is regarded as having one dimension. Thus a length had the dimension L. The unit of length enters twice into the unit of area, first determining the base and then the altitude of the unit rectangle, and therefore the dimensions of an area are LL, usually written L2. If we alter the unit of length, say from a foot to an inch (1:12) the unit of area will be reduced in the same ratio twice successively (1:144 in all). The variations of the unit of area, therefore, are directly as the squares of the variations in the unit of length. The units of length and of time enter once each into the unit of speed, but they do not enter on the same footing. If the unit of time be the minute, and the unit of length the foot, the unit of speed will be a foot per minute. This unit will become smaller if we make the unit of length smaller, since an inch per minute is a smaller speed than a foot per minute; but it will become larger if we make the unit of time smaller, a foot a second being a greater speed than a foot a minute. This is expressed by saying that the dimensions of time T enters negatively into speed. The dimensions of

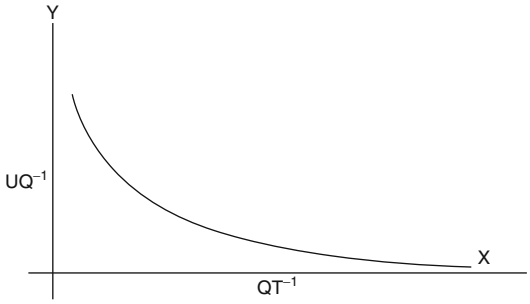
speed, then, are expressed as LT–1. A unit into which a dimension enters negatively is always a unit of rate, and measures amount of x per unit of y, −y being the quantity the dimension of which enters negatively.

We have now examined simple cases of the variations of derived units, but it is obvious that the numerical values of concrete magnitudes vary inversely as the units by reference to which they are estimated. The smaller the unit the greater the numerical value of any given magnitude. The numerical value of a magnitude, therefore, will vary inversely as the unit whose dimension enters into it positively, and directly as the unit whose dimensions enters into it negatively. Thus, let the unit of speed (dimensions LT–1) be a foot per minute, and let the numerical value of a certain concrete speed be 10, i.e. let the speed be ten per minute. Then change the unit of length to an inch (1:12) and the unit of time to a second (1:60); the derived unit will now be an inch per second, and its relation to the former derived unit is obtained by altering directly in the ratio of 1:12 (dividing by 12) and inversely in the ratio of 1:60 (multiplying by 60), so that the new unit is five times as great as the old one, an inch per second being five times as great a speed as a foot per minute; but the numerical value of the concrete speed we had to express must be altered inversely as 1:12 and directly as 1:60, and is now only 2 – i.e. the speed is two inches per second – or one-fifth of what it was before.

If we are measuring such a magnitude as feet of vertical motion per foot of horizontal motion in the path of a projectile, the dimensions will be LL–1 and will cancel each other. No change in the unit of length, then, will in any way affect the numerical value of this magnitude, and as no other dimension enters into it at all, it may be said to have no dimensions. Angular magnitudes, defined as ratios between arcs and radii, trigonometrical functions, and ratios generally are of this nature. They have no selected units, and their numerical values are absolute.

When the elements of the theory of dimensions have been thoroughly grasped it will be easy to apply it to economic questions; and it will be found an invaluable check in the more intricate problems of co-ordination and analysis. Thus, if the unit of value-in-use or utility be taken as fundamental, and regarded as having the dimension U, and if the commodity we are considering be taken as having the dimension Q, then degree of utility of the commodity, being the rate at which satisfaction is secured per unit of commodity consumed, will have dimensions UQ–1, and, will be readily distinguished from rate of enjoyment, accruing to the consumer, per unit of time, with dimensions UT–1. Price, determined by marginal, or final, degree of utility, will have dimensions UQ–1 or P; and hire, being price per unit of time, will obviously have dimensions PT–1 or UQ–1T–1. When the thing hired is money and is used commercially, the utility derived from it is a commodity of like nature with itself. The dimension U then becomes Q, and the dimensions of interest (as a rate) are QQ–1T–1 or T–1, which will be found on reflection and experiment to be correct.

The theory of dimensions should be applied to economics in close connection with the diagrammatic method. But of course the connection between dimensions, as now explained, and the geometrical dimensions of the diagrams is purely arbitrary. The physicist may, according to his convenience, represent the height of a projectile – a magnitude of one dimension – by a line, or by an area, and speed by a line of an inclination. So the economist may represent a magnitude measured by a complicated derived unit by a line, or a magnitude measured by a fundamental unit by an area or a solid; and if he keeps the theory of dimensions well before him he may vary his methods indefinitely without any danger of confusion. In all cases, however, the dimensions of those quantities represented by areas or solids will be compounded of the dimensions of those represented by the lines which determine them. Again, those who have any acquaintance with the elements of the calculus will see that if the equation of a curve be differentiated to x then the area of the derived curve will have the same dimensions as the ordinate of the fundamental curve; the ordinate of the derived curve will have the dimensions of the ordinates of the fundamental curve positively, and those of its abscissae negatively;



Dimensions of Economic Quantities, Fig. 1

and the abscissae of the two curves will have the same dimensions. In other words, differentiation introduces the dimensions of the variable to which we differentiate negatively, and integration introduces the dimensions of the variable to which we integrate positively (Fig. 1).

By way of illustration take a figure, on the ordinate of which intensity of desire, or degree of utility, is represented, while supply of commodity per unit of time is measured on the abscissae. Now imagine a third axis (of Z) perpendicular to the page, along which time is measured. Such a figure will enable us to represent all the quantities we have to deal with in an ordinary problem of consumption. Rate of supply is represented on axis of X, dimensions QT–1; degree of utility on axis of Y, dimensions UQ–1; time on axis of Z, dimension T; rate of enjoyment on areas parallel to plane of axes of X and Y, dimensions UQ–1QT–1 or UT–1; total enjoyment on solid figure, dimensions UQ–1QT–1T, or U; total supply on areas parallel to plane of axes of X and Z, dimensions QT–1T, or Q, and in like manner price, hire, total sum paid, etc., may be read, and their dimensional relations seen at a glance.

[The theory of dimensions was (according to Jevons, Principles of Science, 1887, p. 325) first clearly stated by Joseph Fourier. He expounded it with great lucidity in his Théorie Analytique de la Chaleur, 1882, §§ 159–62. An excellent popular statement of the theory, as it has since been elaborated, will be found in the beginning of J. D. Everett's C.G.S. System of Units, 1891. Jevons was the first to suggest the application of the theory to economics (Theory of Political Economy, 1888, pp. 232–52), but he unfortunately fell into some apparent errors and confusions which made the suggestion barren in his hands. A criticism of his treatment of the subject and an independent working-out of his suggestion, by the writer of the present article will be found in the American Quarterly Journal of Economics for April 1889, pp. 297–314.]

Bibliography

Everett, J.D. 1875. *Illustrations of the centimetre-gramme-second system of units, with tables of physical constants*. London.
Fourier, J. 1822. *Theorie analytique de chaleur*. Paris.
Jevons, W.S. 1871. *The theory of political economy*. London: Macmillan.
Jevons, W.S. 1887. *Principles of science*. London: Macmillan.
Wicksteed, P.H. 1888. *The alphabet of economic science*. London: Macmillan.
Wicksteed, P.H. 1889. On certain passages in Jevons's *Theory of Political Economy*. *Quarterly Journal of Economics* 3: 293–314.
Wicksteed, P.H. 1895. In *A symposium on value*, ed. J.H. Levy. London: Macmillan.
Wicksteed, P.H. 1910. *The common sense of political economy*. London: Macmillan.

Direct Taxes

John Kay

The distinction between direct taxes and indirect taxes traditionally rests on a view of the incidence of the two kinds of tax. The incidence of a tax identifies who suffers loss of income or welfare as a result of the imposition of the tax. This may differ from the location of the legal liability for payment of the tax if the payer is able to shift part or all of this liability to some other agent. The capacity to shift the tax burden in this way depends on the elasticities of demand and supply of the taxed factor or commodity. Direct taxes are those for which the legal liability and the incidence are identical: indirect taxes are those where the tax is shifted, most usually to final consumers.

Thus income taxes are generally regarded as direct taxes and commodity taxes as indirect. This supposes that factor supplies are completely inelastic and commodity supplies perfectly elastic, an empirical observation which may hold in particular cases but which cannot be seen as a universal truth. In reality, all taxes are shifted to some extent and none completely. A more recent reformulation of the direct/indirect distinction (Atkinson and Stiglitz 1980) describes indirect taxes as those differentiated by the nature of the transaction and direct taxes are those differentiated by the identity of the transactors; but this too breaks down on closer examination and the classification is one with no particular economic significance. The most important direct taxes are progressive wealth, income and expenditure taxes levied on individuals, and tax imposed on the income of corporations.

Before the emergence of modern systems of public finance, ad hoc wealth taxes were a primary source of revenue. But this was possible only when wealth mostly took the form of real property and revenue requirements were relatively minor. Many countries still have a tax on wealth, but there is none in which it makes an important contribution to revenue. Taxes are often levied on transfers of capital, on death or sometimes when substantial gifts are made. Such a tax may be donor or donee based. An inheritance tax levied on the donor is a progressive tax based on the total of gifts made by the tax payer. An accession tax on the donee is one in which the rate of tax is based on the cumulative total of gifts received.

Although there is extensive academic discussion of the potential of a direct tax on expenditure, no major country has adopted one. Direct taxes are primarily income taxes, on the incomes of individuals and of corporations. We begin by looking at the base of the tax and then consider the criteria which should determine the rates at which income – personal or corporate – should be taxed.

It seems trite to observe that in order to tax income it is necessary to define it, but in fact the taxing statutes of most states do not attempt to do so. Income is exemplified rather than defined. For economists, the classic definition of income is that of Hicks (1939) – 'income is the maximum value which a man can consume during a week and still be as well off at the end of the week as he was at the beginning'. By the same principle, corporate income might be defined as the maximum which a company can distribute, and still be as well off at the end of the accounting period as at the beginning. But these are not operational concepts for a tax inspector. How is he to determine what a man expects? And what is he to do if these expectations are unreasonable?

Thus attention has instead been devoted to the concept of 'comprehensive income', or Haig–Simons income, so-called after its principal advocates (Simons 1938). As Hicksian income looks forward, so Haig-Simons looks back, and measures not what a man could have expected to consume but what he could in fact have consumed. If expectations are always fulfilled, then the two concepts are identical: but windfall gains, excluded from the Hicksian concept of income, fall within the Haig–Simons one. It follows that Haig–Simons income requires that all accruing capital gains should be included within the tax base and taxed as income. In fact no country has gone as far as this; some tax certain capital gains as income; almost all tax most kinds of capital gain more lightly, if at all.

The application of either a Hicksian or a Haig–Simons measure of real income implies indexation of the tax base. This means not only that capital gains should be adjusted for inflation, but that investment income – paid or received – should be adjusted also.

Such inflation adjustment should relate to individual income, to capital gains, and to the income of corporations. Inflation adjustment to the income of individuals is very rare, although several countries now provide for indexation in calculating capital gains. Most attention to the effects of inflation on the measurement of income has been given in the corporate sector inflation. Accounting profit becomes a misleading indicator of the returns earned by a company under inflation because depreciation is generally based on the historic cost rather than the current cost of equivalent assets; because the rise in the price of goods held in stock (stock depreciation) is included in

profits; and because interest paid or received is expressed in nominal rather than real terms. All countries with recent experience of high rates of inflation have considered changes to accounting standards to remove these distortions but agreement on appropriate adjustments has proved elusive.

In the absence of accepted accounting principles, tax systems have responded to inflation in ad hoc ways. The inadequacy of historic cost depreciation allowances has been partly compensated for by acceleration of the rate at which such allowances may be taken. Relief for the effect of inflation on stock values has been given, either by accepting accounting practices such as LIFO (last in, first out) which automatically give relief at current prices, or by particular measures of stock relief. Tax authorities have been much more reluctant to make allowance for the effect of inflation in eroding the value both of the monetary assets of companies and of their debts.

A tax system which was fully indexed in this way would be neutral with respect to the rate of inflation, but it would not equalize pre-and post-tax rates of return because the real return earned by the company would continue to be subject to tax. Full neutrality could be achieved by means of a cash flow tax, which allows immediate deductibility of all capital expenditure – either in stocks or on fixed assets – but denies any relief for financing costs, whether interest or otherwise. Such a tax was proposed by the Meade Committee (1978) and it uses as its base the flow of funds from the real operations of the company to those who finance it.

Once income has been defined, at what rate should it be taxed? Differentiation between types of personal income was a principal issue when income tax was introduced in the nineteenth century. The argument rested on the precariousness of income from employment relative to property income, and this, is was suggested, provided a reason for taxing investment income more heavily. These arguments read rather oddly in a twentieth-century context, in which inflation and economic fluctuations have generally made property income appear more precarious than earnings, and this argument has largely vanished from discussion and its consequences from tax schedules.

A tax schedule is progressive if the average rate of tax increases with income. This does not require that marginal tax rates are increasing and indeed a linear tax schedule is progressive if its intercept is positive. There is no unambiguous measure of progressivity, and the same term is sometimes used to cover both the extent to which the schedule deviates from proportionality and the redistributive effect of tax structure. It will be apparent that a tax which departs substantially from proportionality but generates little revenue will have less redistributive effect than a more nearly proportional but heavier tax.

Nineteenth-century utilitarian arguments suggested alternative rate structures. The principle of equal sacrifice, for example, demanded a schedule which imposed equal utility losses on all the taxpayers. This implied payments from those with higher incomes, but not necessarily proportionately larger payments, the outcome depending on the elasticity of the marginal utility of income. Utility maximization subject to a revenue constraint requires equal marginal sacrifices, with similar implications.

However, these analyses take no account of the effects of tax schedules on labour supply. Like indirect taxes, income taxes impose a deadweight loss or excess burden in addition to the revenue which they raise. The magnitude of these losses depends on marginal tax rates and the wage elasticity of labour supply. It follows that there is a direct conflict between the progressivity of a tax schedule – which implies high marginal rates of tax – and its efficiency properties – which require low marginal rates.

Mirrlees (1971) was the first to examine this trade-off explicitly and although a substantial literature on optimal income tax structures has developed since, relatively few results of general application have emerged. There is some indication that marginal tax rates should be lower at the extremes of the distribution than in the middle of it. The disincentive effects of high marginal tax rates depend on the numbers of individuals in the relevant range, whereas their redistributive function depends on the number of individuals above

that range. As we move up the income distribution, this redistributive effect steadily diminishes, while the disincentive effect remains; and thus the balance between the two factors changes in a direction which points to lower marginal rates of tax. Similar arguments can be applied at the lower end of the distribution.

While the welfare effects of income taxation are principally the product of marginal rates, the overall effect on labour supply is determined by both income and substitution effects, and is therefore influenced by the average as well as the marginal tax rate at any point in the distribution. For this reason, while the efficiency costs of increasing taxation are unambiguous the labour supply effect may be positive or negative in sign. Labour supply is presumably zero at tax rates of 100 per cent, however, and if an interior maximum exists (which is by no means certain) then there will be some rate below this which yields maximum revenue. This observation yields what has become known as the Laffer curve.

The structural issues which influence redistribution across the income distribution are concerned with vertical equity in taxation. Horizontal equity reflects its concern with the relative tax burdens at the same point in the income distribution. Horizontal equity implies that individuals in the same circumstances should be treated similarly and would exclude, for example, random taxation (even though this might, under certain circumstances, be efficient). However the principle of horizontal equity has limited application because of the difficulty of agreeing an objective definition of 'similar circumstances'. The most important issues of horizontal equity in practice concerns the tax treatment of the family, an area of taxation in which there is direct conflict between two conflicting principles – the desire to respect the right of individuals to individual treatment, which points to an individual basis for taxation, and the desire to relate liability to the whole of an individual's circumstances, which necessarily includes the circumstances of those with whom he or she lives. Most tax systems incorporate elements of both individual and unit bases.

In fixing the rate of corporate income tax, it is necessary to begin by asking why we tax corporate income at all. Although corporations have distinct legal personalities, they have no economic personality and ultimately generate no command over resources other than those of the individuals who work for them, manage them, buy their products, or own their shares. It is these individuals who pay corporation tax. The economic rationale for corporate income taxes therefore requires justification.

One such argument is that they are there: the phenomenon of tax capitalization implies that if particular assets, such as the equity of corporations, are subject to discriminatory taxation then these taxes will be reflected in the prices of the assets concerned. To remove such a tax would effect no current efficiency gain, and would confer windfall gains on current shareholders; this is the rationalization of the traditional maxim that 'an old tax is a good tax'. Corporation tax may also enable countries to derive revenue from the assets of non-residents; this is a powerful argument for such a tax in many countries.

An important point is that in the absence of corporate income tax, individuals would avoid the personal income tax through incorporation. This suggests that the income of corporations should be attributed to its owners and taxed as their income. Although the possibility of full integration of corporate and personal income taxation has been discussed, and was recommended for Canada by the Carter Commission (1966), no country has yet adopted it. The classical system of corporation tax is one in which the income of corporations is taxed at a flat rate entirely separate from the income of shareholders. This is the system used in the United States; most European countries, however, now employ an imputation system in which the shareholder receives some credit against his own income tax bill on dividends for corporation tax paid by the company from which he receives them. This relieves the element of double taxation implicit in the classical system, but still tends to tax income accruing through corporations more heavily than other kinds of income.

Corporation tax has usually been seen as a tax on capital employed in the corporate sector. It follows that this purpose is discriminated against

relative to other uses of capital in the domestic economy, such as agriculture or property. This is the approach adopted in Harberger's classic (theoretical) analysis of the incidence of corporation tax (Harberger 1962), which traced its effects on returns to capital in different sectors of the economy. It is also implicit in most empirical studies of the impact of corporation tax, such as those of Musgrave and Krzyzaniak (1964), which have considered the question of the extent to which a tax on the capital employed by corporations can be shifted forward into the prices of goods produced by the corporate sector. Their work suggested that the extent of such shifting might be substantial.

More recent analysis has challenged this approach to the incidence of company taxation (Stiglitz 1976). The argument is that corporation tax cannot appropriately be represented as a tax on capital employed. Most corporate taxes allow extensive deductions for capital costs, such as interest and depreciation. If capital costs are fully deductible, then the corporate tax system is neutral. Such neutrality can be achieved either if all investment costs can be expensed, or if depreciation allowances correspond to true economic depreciation and financing costs are fully deductible, through tax relief on interest paid and imputation for company dividends. If the tax regime provides – as is common in many countries – both for deductions for the costs of finance and for accelerated depreciation, then the corporation tax may actually act as a subsidy to corporate capital rather than a tax. The post-tax rate of return may exceed the pre-tax rate. Such a tax may still yield revenue, since it will still fall on pure profits, i.e. returns earned by the firm which are not directly attributable to its capital employed.

Pure profits are generated by entrepreneurship, a word which may describe the classic entrepreneurial function of bringing different factors of production together; the exploitation or establishment of monopoly rents; or the generation of new means of organization or invention. Thus the new view of corporation tax sees it as a levy on those items, combined with a rather arbitrary array of taxes and subsidies to different types of investment. The rates of these taxes and subsidies depend on the degree to which a given activity can be financed by debt rather than equity and the relationship between true economic depreciation and what is permitted for tax purposes.

Direct taxation can be adjusted sensitively to bold social and economic objectives, and as modern states have developed and their revenue requirements have grown so reliance on them has tended to increase. More recently, however, dependence on personal income tax has been seen to imply excessive rates. The result has been some moves back towards broadly based indirect taxes, particularly the value added tax, which has been introduced throughout the European community and in about thirty other states.

Similar pressures have been evident in the corporate sector. Taxing corporate income is therefore not the only means of taxing corporations and, given the difficulties involved in identifying the country within which income arises, measuring income in a period of inflation, and taxing declining real profitability, taxes on corporate income have tended to diminish in importance. The average share of total OECD tax receipts derived from corporation tax fell from 9.2 per cent to 7.4 per cent between 1965 and 1983. At the same time, however, other taxes on business, particularly payroll and social security taxes, have tended to increase: implying an overall shift in relative tax rates on capital and labour as factors of production.

See Also

- ▶ Taxation of Income
- ▶ Taxation of Wealth

Bibliography

Atkinson, A.B.., and J.E. Stiglitz. 1980. *Lectures on public economics*. New York: McGraw Hill.

Carter Commission. 1966. Report of the Royal Commission on Taxation.

Harberger, A.C. 1962. The incidence of the corporation income tax. *Journal of Political Economy* 70: 215–240.

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

Meade, J.E. (chairman). 1978. *The structure and reform of direct taxation.* London: Allen & Unwin.

Mirrlees, J.A. 1971. An exploration in the theory of optimal income taxation. *Review of Economic Studies* 38: 175–208.
Musgrave, R.A., and M. Krzyzaniak. 1964. *The shifting of the corporation income tax*. Baltimore: Johns Hopkins Press.
Simons, H.C. 1938. *Personal income taxation*. Chicago: Chicago University Press.
Stiglitz, J.E. 1976. The corporation tax. *Journal of Public Economics* 5(3–4): 303–311.

Directly Unproductive Profit-Seeking (DUP) Activities

Jagdish N. Bhagwati

Keywords

Chicago School; Directly unproductive profit-seeking (DUP) activities; Endogeneous tariffs; Immiserizing growth; Lobbying; Optimal tariffs; Predation; Production subsidies; Public choice; Regulation; Rent seeking; Revenue seeking; Shadow pricing; Smuggling; Tariff seeking; Tariffs; Transfer problem; Voluntary export restrictions

JEL Classifications

F2

Directly unproductive profit-seeking (DUP) activities are defined (Bhagwati 1982a) as ways of making a profit (that is, income) by undertaking activities which are directly (that is, immediately, in their primary impact) unproductive, in the sense that they produce pecuniary returns but do not produce goods or services that enter a conventional utility function or inputs into such goods and services.

Typical examples of such DUP (pronounced appropriately as 'dupe') activities are (*i*) tariff-seeking lobbying which is aimed at earning pecuniary income by changing the tariff and therefore factor incomes; (ii) revenue-seeking lobbying which seeks to divert government revenues towards oneself as recipient; (iii) monopoly seeking lobbying whose objective is to create an artificial monopoly that generates rents; and (iv) tariff-evasion or smuggling which de facto reduces or eliminates the tariff (or quota) and generates returns by exploiting thereby the price differential between the tariff-inclusive legal and the tariff-free illegal imports.

While these are evidently profitable activities, their *output* is zero. Hence, they are wasteful in their primary impact, recalling Pareto's distinction between production and predation: they use real resources to produce profits but no output.

DUP activities of one kind or another have been analysed by several economic theorists, among them (i) the public-choice school's leading practitioners, their major work having been brought together in Buchanan et al. (1980), (ii) Lindbeck (1976) who has worked on 'endogeneous politicians', and (iii) the Chicago 'regulation' school, led by Stigler, Peltzman, Posner and also Becker (1983).

However, a central theoretical breakthrough has come from the work of trade theorists who have systematically incorporated the analysis of DUP activities in the main corpus of general equilibrium theory.

The early papers that defined this general-equilibrium-theoretic approach, and which were set in the context of the theory of trade and welfare, were: Bhagwati and Hansen (1973) which analysed the question of illegal trade (that is, tariff-evasion), Krueger (1974) which analysed the question of rent-seeking for rents associated with import quotas specifically and quotas more generally, and (iii) Bhagwati and Srinivasan (1980) who analysed the phenomenon of revenue-seeking, the 'price' counterpart of Krueger's rent-seeking, where a tariff resulted in revenues which were then sought by lobbies.

The synthesis and generalization of these and other apparently unrelated contributions, showing that they all related to diversion of resources to zero-output activities, was provided in Bhagwati (1982a) where they were called DUP activities. The following significant aspects of the theoretical analysis of DUP activities are noteworthy.

First, they are generally related to policy interventions (but they need not be: plunder, for

instance, pre-dates the organization of governments). In so far as policy interventions induce DUP activities, they are analytically divided into two appropriate categories (Bhagwati and Srinivasan 1982):

Category I: *Policy-triggered DUP activities.* One class consists of *lobbying* activities. Examples include: rent-seeking analysis of the cost of protection *via* import licences (Krueger 1974); revenue-seeking analysis of the cost of tariffs (Bhagwati and Srinivasan 1980), of shadow prices in cost–benefit analysis (Foster 1981), of price versus quantity interventions (Bhagwati and Srinivasan 1982), of non-economic objectives (Anam 1982), of rank-ordering of alternative distorting policies such as tariffs, production and consumption taxes (Bhagwati et al. 1984), of the optimal tariff (Dinopoulos 1984), of the transfer problem (Bhagwati et al. 1985), and of voluntary export restrictions relative to import tariffs (Brecher and Bhagwati 1987).

Another class consists of *policy-evading* activities. Examples include: analysis of smuggling (Bhagwati and Hansen 1973), its implication for optimal tariffs (Johnson 1974; Bhagwati and Srinivasan 1973), and alternative modelling by Kemp (1976), Sheikh (1974), Pitt (1981) and Martin and Panagariya (1984).

Category II: *Policy-influencing DUP activities.* The other generic class of DUP activities is not triggered by policies in place but is rather aimed at influencing the formulation of the policy itself. The most prominent DUP-theoretic contributions in this area relate to the analysis of tariff-seeking. Although Brock and Magee (1978, 1980) pioneered here, the general equilibrium analyses of endogeneous tariffs began with Findlay and Wellisz (1982) and Feenstra and Bhagwati (1982), the two sets of authors modelling the government and the lobbying activities in contrasting ways. Notable among the later contributions are Mayer (1984), who extends the analysis formally to include factor income-distribution and therewith voting behaviour, and Wellisz and Wilson (1984). Magee (1984) has an excellent review of many of these contributions. The implication of endogenizing the tariff for conventional measurement of the cost of protection has been analysed in Bhagwati (1980) and Tullock (1981).

The *choice* between alternative policy instruments when modelling the response of lobbies and governments to import competition has also been extensively analysed. The issue was raised by Bhagwati (1982b) and analysed further by Dinopoulos (1983) and Sapir (1983) in terms of how different agents (for example, 'capitalists' and 'labour') would profit from different policy responses such as increased immigration of cheap labour and tariffs when import competition intensified. It has subsequently been explored more fully by Rodrik (1986), who compares tariffs with production subsidies.

Second, Bhagwati (1982a) has noted, generalizing a result in Bhagwati and Srinivasan (1980), that DUP activities, while defined to be those that waste resources in their direct impact, cannot be taken as *ultimately* wasteful, that is, immiserizing, since they may be triggered by a suboptimal policy intervention. For, in that event, throwing away or wasting resources may be beneficial. The shadow price of a productive factor in such 'highly distorted' economies may be negative. This is the obverse of the possibility of immiserizing growth (Bhagwati 1980). Thus, Buchanan (1980), who has addressed the issue of DUP activities and *defined* them as activities that (ultimately) cause waste, has been corrected in Bhagwati (1983): the definition of DUP activities cannot properly exclude the possibility that DUP activities are ultimately beneficial rather than wasteful. This central distinction between the direct and the ultimate welfare impacts of DUP activities is now universally accepted. DUP activities are therefore defined now, as in Bhagwati (1982b) and subsequent contributions, as wasteful only in the direct sense.

Third, Bhagwati et al. (1984) have raised yet another fundamental issue concerning DUP activities. Thus, where DUP activities belong to Category II distinguished above, full endogeneity of policy can follow. If so, the conventional rank-ordering of policies is no longer possible. We have the *determinacy paradox*: policy is chosen in the solution to the full 'political-economy', DUP-theoretic solution and cannot be

varied at will. These authors have therefore suggested that, where full endogeneity obtains, the appropriate way to theorize about policy is to take variations around the observed DUP-theoretic equilibrium. Thus, traditional economic parameters such as factor supply could be varied; similarly now the DUP-activity parameters such as, say, the cost of lobbying could be varied. The impact on actual welfare resulting from such variations can then be a proper focus of analysis, implying a wholly different way of looking at policy questions from that which economists have employed to date.

Finally, DUP activities are related to Krueger's (1974) important category of rent-seeking activities. The latter are a subset of the former, in so far as they relate to lobbying for quota-determined scarcity rents and are therefore part of DUP activities of Category II distinguished above (Bhagwati 1983).

See Also

- ▶ Bribery
- ▶ Rent Seeking

Bibliography

Anam, M. 1982. Distortion-triggered lobbying and welfare: A contribution to the theory of directly-unproductive profit-seeking activities. *Journal of International Economics* 13 (August): 15–32.

Becker, G.S. 1983. A theory of competition among pressure groups for political influence. *Quarterly Journal of Economics* 93: 371–400.

Bhagwati, J. 1980. Lobbying and welfare. *Journal of Public Economics* 14: 355–363.

Bhagwati, J. 1982a. Directly-unproductive profit-seeking (DUP) activities. *Journal of Political Economy* 90: 988–1002.

Bhagwati, J. 1982b. Shifting comparative advantage, protectionist demands, and policy response. In *Import competition and response*, ed. J. Bhagwati. Chicago: Chicago University Press.

Bhagwati, J. 1983. DUP activities and rent seeking. *Kyklos* 36: 634–637.

Bhagwati, J., and B. Hansen. 1973. Theoretical analysis of smuggling. *Quarterly Journal of Economics* 87: 172–187.

Bhagwati, J., and T.N. Srinivasan. 1973. Smuggling and trade policy. *Journal of Public Economics* 2: 377–389.

Bhagwati, J., and T.N. Srinivasan. 1980. Revenue-seeking: A generalization of the theory of tariffs. *Journal of Political Economy* 88: 1069–1087.

Bhagwati, J., and T.N. Srinivasan. 1982. The welfare consequences of directly-unproductive profit-seeking (DUP) lobbying activities: Price *versus* quantity distortions. *Journal of International Economics* 13: 33–44.

Bhagwati, J., R. Brecher, and T.N. Srinivasan. 1984. DUP activities and economic theory. In *Neoclassical political economy: The analysis of rent-seeking and DUP activities*, ed. D. Colander. Cambridge, MA: Ballinger & Co.

Bhagwati, J., R. Brecher, and T. Hatta. 1985. The generalized theory of transfers and welfare: Exogenous (policy-imposed) and endogenous (transfer-induced) distortions. *Quarterly Journal of Economics* 100: 697–714.

Brecher, R. and Bhagwati, J. 1987. Voluntary export restrictions and import restrictions: A welfare-theoretic comparison. Essays in honour of W.M. Corden, H. Kierzkowski. Oxford: Basil Blackwell.

Brock, W., and S. Magee. 1978. The economics of special interest politics: The case of the tariff. *American Economic Review* 68: 246–250.

Brock, W., and S. Magee. 1980. Tariff formation in a democracy. In *Current issues in International Commercial Policy and Diplomacy*, ed. J. Black and B. Hindley. New York: Macmillan.

Buchanan, J. 1980. Rent seeking and profit seeking. In *Towards a general theory of the rent-seeking society*, ed. J. Buchanan, G. Tullock, and R. Tollison. College Station: Texas A&M University Press.

Buchanan, J., G. Tullock, and R. Tollison, eds. 1980. *Towards a general theory of the rent-seeking society*. College Station: Texas A&M University Press.

Dinopoulos, E. 1983. Import competition, international factor mobility and lobbying responses: The Schumpeterian industry cases. *Journal of International Economics* 14: 395–410.

Dinopoulos, E. 1984. The optimal tariff with revenue-seeking: A contribution to the theory of DUP activities. In *The neoclassical political economy: The analysis of rent-seeking and DUP activities*, ed. D. Colander. Cambridge, MA: Ballinger & Co.

Feenstra, R., and J. Bhagwati. 1982. Tariff seeking and the efficient tariff. In *Import competition and response*, ed. J. Bhagwati. Chicago: Chicago University Press.

Findlay, R., and S. Wellisz. 1982. Endogenous tariffs, the political economy of trade restrictions, and welfare. In *Import competition and response*, ed. J. Bhagwati. Chicago: Chicago University Press.

Foster, E. 1981. The treatment of rents in cost-benefit analysis. *American Economic Review* 71: 171–178.

Johnson, H.G. 1974. Notes on the economic theory of smuggling. In *Illegal transactions in International Trade*, Series in International Economics, ed. J. Bhagwati. Amsterdam: North-Holland.

Kemp, M. 1976. Smuggling and optimal commercial policy. *Journal of Public Economics* 5: 381–384.

Krueger, A. 1974. The political economy of the rent-seeking society. *American Economic Review* 64: 291–303.
Lindbeck, A. 1976. Stabilization policies in open economies with endogenous politicians. Richard Ely Lecture. *American Economic Review* 66: 1–19.
Magee, S. 1984. Endogenous tariff theory: A survey. In *Neoclassical political economy: The analysis of rent-seeking and DUP activities*, ed. D. Colander. Cambridge, MA: Ballinger & Co.
Martin, L., and A. Panagariya. 1984. Smuggling, trade, and price disparity: A crime-theoretic approach. *Journal of International Economics* 17: 201–218.
Mayer, W. 1984. Endogenous tariff formation. *American Economic Review* 74: 970–985.
Pitt, M. 1981. Smuggling and price disparity. *Journal of International Economics* 11: 447–458.
Rodrik, D. 1986. Tariffs, subsidies and welfare with endogenous policy. *Journal of International Economics* 21: 285–299.
Sapir, A. 1983. Foreign competition, immigration and structural adjustment. *Journal of International Economics* 14: 381–394.
Sheikh, M. 1974. Smuggling, production and welfare. *Journal of International Economics* 4: 355–364.
Tullock, G. 1981. Lobbying and welfare: A comment. *Journal of Public Economics* 16: 391–394.
Wellisz, S. and Wilson, J.D. 1984. *Public sector inefficiency, a general equilibrium analysis*. Discussion Paper No. 254, International Economics Research Center, Columbia University.

Director, Aaron (1901–2004)

Steven G. Medema

Keywords

Antitrust; Cartels; Director, A.; Law and economics; Oligopoly; Posner, R.; Predatory pricing; Resale price maintenance; Tie-in sales

JEL Classifications

B31

Aaron Director's enduring contribution to economics came via his role in the development of the Chicago law and economics tradition. Director was born in Charterisk (in present-day Ukraine) in 1901 and emigrated to the United States with his family in 1913. He received his undergraduate degree from Yale University and his graduate training at the University of Chicago. Although he came to Chicago in 1927 to work with Paul Douglas on labour economics, it was Frank Knight and Jacob Viner who, via their price theory courses, had the greatest influence on him. Director remained at Chicago as a graduate student and part-time instructor until 1934. The 1930s were a heady period at Chicago, where the student body included George Stigler, Paul Samuelson (who credits Director's teaching with stimulating his interest in economics), and Milton Friedman – each of whom helped to reshape economic thinking in the middle third of the 20th century – as well as Rose Director (Aaron's sister and, eventually, Rose Friedman). Aaron Director was very much part of this milieu. He left the University of Chicago for the US Treasury Department in 1934 and, save for an aborted attempt to complete a dissertation on the history of the Bank of England, remained in Washington, DC, until 1946, when he returned to the University of Chicago to take up a position in the Law School, where he remained until his retirement in 1966.

Director's appointment in the Law School was a result of the efforts of Henry Simons, the first economist on the law faculty at Chicago, and Friedrich Hayek, whose *Road to Serfdom* was published in the United States largely because of Director's intervention with the University of Chicago Press. The plan, as laid out by Simons, was for Director to head up the 'Free Market Study', a Volker Fundfinanced project, housed in the Law School and dedicated to undertaking 'a study of a suitable legal and institutional framework of an effective competitive system' (Coase 1998, p. 246). However, Simons committed suicide in the summer of 1946, and Director was asked to take on Simons's basic Law School price theory course, 'Economic Analysis of Public Policy'. This provided Director with an initial forum for bringing the perspective he had learned from Knight and Viner into the Law School's teaching programme.

The transition from having an economist on the Law School faculty to the establishment of a law and economics tradition at Chicago began not

long after this, when Edward Levi invited Director to collaborate in the teaching of the antitrust course. Levi would teach a traditional antitrust course for four days each week; Director would then come in on the fifth day and, using the tools of price theory, show that the traditional legal approach could not stand up to the rigours of economic analysis. The basic pattern was very simple: Director would ask whether the practice in question was, in general, consistent with monopolistic profit maximization. The answer was often negative, which meant that there had to be some sort of legitimate rationale for the supposedly anti-competitive practice in question. What Director's price theory showed was that the 'simple and obvious' answers were often wrong-headedly simplistic. This process had a profound impact on students and colleagues alike. Director's antitrust students – a group that included Robert H. Bork, Ward Bowman, Kenneth Dam, Edmund Kitch, Wesley J. Liebeler, John S. McGee, Henry Manne, and Bernard H. Siegan – have often spoken of the 'conversion' they experienced in this class, and even Levi himself became a partial convert (see Kitch 1983; Director and Levi 1951). What was perhaps Director's most significant contribution on the missionary front came after his retirement, when he and Richard Posner spent time together at Stanford in 1968 – Posner's first year on the Stanford Law School faculty. It was Director who taught Posner to think like a Chicago economist, introduced him to Stigler and Ronald Coase, and in this and other ways was instrumental in Posner's move to the Chicago Law School after only one year on the Stanford faculty. The rest, as they say, is history.

Although Director's published output was slight, his influence extended well beyond the classroom. His insights made their way into the antitrust literature – and, eventually, antitrust policy – through the writings of students and colleagues, as Sam Peltzman (2005) has detailed. Director's primary legacies are in the analysis of predatory pricing (via McGee 1958), resale price maintenance (via Telser 1960), and tie-in sales (see Director and Levi 1951; Bowman 1957; Burnstein 1960), but his influence was also prominent in Stigler's view of oligopoly and antitrust policy, Posner's (1969) perspective on oligopoly and cartels, and Robert Bork's influential articles on antitrust (for example, Bork and Bowman 1965; Bork 1967). These contributions coalesced in a distinctive Chicago approach to antitrust analysis, an approach that Herbert Hovenkamp (1986, p. 1020) says 'has done more for antitrust policy than any other coherent economic theory since the New Deal', and whose influence is inescapable.

Director's impact at the Law School went far beyond antitrust: He was also the prime mover in the early professionalization of law and economics. Director formally established the nation's first law and economics programme, which maintained visiting fellowships for law and economics scholars, and, in 1958, founded the *Journal of Law and Economics*. Within a few decades, Director's efforts at Chicago had been replicated in a set of thriving and well-funded law and economics programmes at major law schools around the country. One would be hard pressed to name an individual in our discipline who has had as much influence as Director without a much more extensive bibliography.

See Also

- ▶ Anti-trust Enforcement
- ▶ Chicago School
- ▶ Law, Economic Analysis of

Selected Works

1931. (With P. Douglas.) *The problem of unemployment*. New York: Macmillan.
1933. The economics of technocracy. Public Policy Pamphlet No. 2. Chicago: University of Chicago Press.
1940. Does inflation change the economic effects of war? *American Economic Review* 35, 351–361.
1951. (With E. H. Levi.) Law and the future: Trade regulation. *Northwestern Law Review* 51, 281–296
1964. Parity and the economic marketplace. *Journal of Law and Economics* 7, 1–10.

Bibliography

Bork, R.H. 1967. Antitrust and monopoly: The goals of antitrust policy. *American Economic Review* 57: 242–253.

Bork, R.H., and W.S. Bowman Jr. 1965. The crisis in antitrust. *Columbia Law Review* 65: 363–376.

Bowman, W.S. Jr. 1957. Tying arrangements and the leverage problem. *Yale Law Journal* 67: 19–36.

Burnstein, M.L. 1960. The economics of tie-in sales. *Review of Economics and Statistics* 42: 68–73.

Coase, R.H. 1998. Director, Aaron. In *The new palgrave dictionary of economics and the law.* London: Macmillan.

Hovenkamp, H. 1986. Chicago and its alternatives. *Duke Law Journal* 1986: 1014–1029.

Kitch, E.W. 1983. The fire of truth: a remembrance of Law and Economics at Chicago, 1932–1970. *Journal of Law and Economics* 26: 163–233.

McGee, J.S. 1958. Predatory price cutting: The standard oil (NJ) case. *Journal of Law and Economics* 1: 137s–1369.

Peltzman, S. 2005. Aaron Director's influence on antitrust policy. *Journal of Law and Economics* 48: 313–330.

Posner, R.A. 1969. Oligopoly and the antitrust laws: A suggested approach. *Stanford Law Review* 21: 1562–1606.

Stigler, S.M. 2005. Aaron Director remembered. *Journal of Law and Economics* 48: 307–311.

Telser, L.G. 1960. Why should manufacturers want fair trade? *Journal of Law and Economics* 3: 86–105.

Discrete Choice Models

Takeshi Amemiya

These are those statistical models which specify the probability distribution of discrete dependent variables as a function of independent variables and unknown parameters. They are sometimes called *qualitative response models,* and are relevant in economics because the decision of an economic unit frequently involves discrete choice: for example, the decision regarding whether a person joins the labour force or not, the decision as to the number of cars to own, the choice of occupation, the choice of the mode of transportation, etc.

Despite their relevance, however, it is only recently (approximately in the last twenty years) that economists have started using them extensively. There seem to be three reasons for a recent surge of interest in such models: (1) Economists have realized that econometric models using only aggregate data cannot accurately explain economic phenomena nor predict the future values of economic variables well. (2) Large scale disaggregated data on consumers and producers have become available. (3) The rapid development of computer technology has made possible estimation of realistic models of this kind.

Note that when aggregated over many individuals, discrete variables behave almost like continuous variables and therefore can be subjected to standard regression analysis. A discrete choice model becomes necessary when we want to model the behaviour of an individual economic unit.

As econometric applications of these models have increased, we have also seen an increase of theoretical papers which address the problem of their specification and estimation. Biometricians have in fact used such models longer than have econometricians, using them, for example, to analyse the effect of an insecticide or the effect of a medical treatment. However, since the versions that econometricians use are generally more complex than those used by biometricians, it has been necessary for the former to develop new models and new methods of statistical inference.

There are cases where a discrete decision of an economic unit is closely interrelated with the determination of the value of a continuous variable. For example, a decision to join the labour force necessitates the decision of how many hours to work and at what wage rate. A decision to buy a car cannot be separated from the decision of how much to spend on a car. The joint determination of the values of discrete variables and continuous variables belongs to the topic of *limited dependent* variables.

Other closely related topics are *Markov chain models* and *duration* (or survival) *models*. These models introduce the time domain into discrete choice models thereby making the models dynamic. In Markov chain models time changes discretely, whereas in duration models time moves continuously.

Those who wish to study the subject in more detail than the present entry are referred to Amemiya (1981, 1985), Maddala (1983), and McFadden (1984).

Univariate Binary Models

Model Specification

The simplest type of a discrete choice model is a univariate binary model which specifies the binary (1 or 0) outcome of a single dependent variable. Let y_i be the ith observation on the binary dependent variable and x_i the ith observation on the vector of independent variables. Then a general univariate binary model is defined by

$$P(y_i = 1) = F(x_i' \beta), \quad i = 1, 2, \ldots, n, \qquad \text{(I)}$$

where P stands for probability, F is a particular distribution function, and β is a vector of unknown parameters. For example, the event $y_i = 1$ may signify that the ith individual buys a car and the elements of the vector x_i may include the income of the ith individual and the price of the car the individual must pay if he decides to buy a car.

Note that we have assumed the argument of F in (I) to be a linear function of the independent variables. As in the linear regression model, this linearity assumption is more general than appears at first, because x_i need not be the original economic variables like income and price, but instead could contain various transformations of the original variables. However, the model in which the function F depends on a nonlinear function of the independent variables and unknown parameters can be handled with only a slight modification of the subsequent analysis.

A variety of models arises as we choose different distribution functions for F. The most commonly used functions are the standard normal distribution function Φ and the logistic distribution function Λ. These functions are defined by

$$\Phi(x) \int_{-\infty}^{x} (2\pi)^{-1/2} \exp(-2^{-1} t^2) \mathrm{d}t$$

and

$$\Lambda(x) = (1 + \mathrm{e}^{-x})^{-1}$$

When $F = \Phi$, the model is called the *probit* model, and when $F = \Lambda$, it is called the *logit* model.

The decision regarding which function to use should be based both on theoretical considerations and on how well a model fits the data. However, as long as a researcher experiments with various independent variables and with various ways in which the independent variables appear in the argument of F, the particular choice of F is not crucial.

Let us consider by way of an example how this model arises as the result of an individual maximizing a utility function. Consider the decision of a person regarding whether he drives a car to work or travels by public transport. We suppose that a level of utility is associated with each alternative and the person is to choose the alternative for which the utility is greater. Let U_{i1} and U_{i0} be the ith person's utilities associated with driving a car and travelling by public transport respectively. We assume that they are linear functions of independent variables with additive error terms as follows:

$$U_{i1} = x_{i1}' \beta_1 + \varepsilon_{i1},$$

and

$$U_{i0} = x_{i0}' \beta_0 + \varepsilon_{i0}.$$

Here, the vector x_{i1} may be thought of as consisting of the time and the cost which would be incurred if the ith person were to drive a car, plus his socio-economic characteristics. The error term may be regarded as the sum of all the unobserved independent variables. Defining $y_i = 1$ if the ith person travels by car and $y_i = 0$ if he travels otherwise, we have

$$\begin{aligned} P(y_i = 1) &= P(U_{i1} > U_{i0}) \\ &= F(x_{i1}' \beta_1 - x_{i0}' \beta_0), \end{aligned}$$

where F is the distribution function of $\varepsilon_{i0} - \varepsilon_{i1}$. Thus, a probit model will result from the normality of $\varepsilon_{i0} - \varepsilon_{i1}$. The normality may be justified on the ground of a central limit theorem.

If a probit model fits the data well, so will a logit model because the logistic distribution function is similar to the standard normal distribution function.

Estimation

Let us consider the estimation of the parameter vector β in the model (I). We shall first discuss the maximum likelihood (ML) estimator and second, the minimum chi-square (MIN χ^2) estimator.

The likelihood function based on n independent binary observations $y_1, y_2, \ldots, y_n$ is given by

$$L = \prod_{i=1}^{n} F\left(x_i'\beta\right)^{yi}\left[1 - F\left(x_i'\beta\right)\right]^{1-yi}.$$

The ML estimator $\widehat{\beta}$ is obtained by maximizing ln L. Under general conditions $\widehat{\beta}$ is consistent and asymptotically normal with the asymptotic variance–covariance matrix given by

$$V\widehat{\beta} = \left\{\sum_{i=1}^{n}\left[\frac{f^2\left(x_i'\beta\right)}{F\left(x_i'\beta\right)\left[1 - F\left(x_i'\beta\right)\right]}x_i x_i'\right]\right\}^{-1},$$

where f is the derivative of F.

Since an explicit formula for the ML estimator cannot be obtained for this model, the calculation of the estimator must be done by an iterative method. The log likelihood function can be shown to be globally concave in the probit and logit models. In these models, therefore, a standard iterative algorithm such as the Newton–Raphson method will generally converge to the global maximum.

The MIN χ^2 estimator, first proposed by Berkson (1944) for the logit model, works only if there are many observations on y for each of the values taken by the vector x. Let us suppose that x_i takes T vector values $x_1, x_2, \ldots, x_T$ and classify integers 1, 2,…, n into T disjoint sets $I_1, I_2, \ldots, I_T$ by the rule: $i \in I_t$ if $x_i = x_t$. Define n_t = number of integers contained in I_t and $\widehat{P}_t = n_t^{-1}\sum_{i \in I_t} y_i$. Then, by a Taylor expansion, we have approximately

$$F^{-1}\left(\widehat{P}_t\right) \square x_t'\beta + \left\{f\left[F^{-1}(P_t)\right]\right\}^{-1}\left(\widehat{P}_t - P_t\right),$$

where F^{-1} denotes the inverse function of F. The MIN χ^2 estimator $\tilde{\beta}$ is the weighted least squares estimator applied to this last heteroscedastic regression equation; that is,

$$\tilde{\beta} = \left[\sum_{t=1}^{T} w_t x_t x_t'\right]^{-1}\sum_{t=1}^{T} w_t x_t F^{-1}\left(\widehat{P}_t\right),$$

where

$$w_t = n_t f_t^2\left[F^{-1}\left(\widehat{P}_t\right)\right] / \left[\widehat{P}_t\left(1 - \widehat{P}_t\right)\right].$$

The MIN χ^2 estimator has the same asymptotic distribution as the ML estimator. Its advantage over the latter is computational simplicity, while its weakness is that it requires many observations for each value of the independent variables. The required number of observations increases with the number of the independent variables. If an independent variable takes many values it may be necessary to group the values into a small number of groups in order to define the MIN χ^2 estimator. But such a procedure will introduce a certain bias to the estimator.

Multinomial Models

A multinomial model is a statistical model for independent discrete variables, some of which take more than two values: Supposing that y_i takes $m_i + 1$ integer values 0, 1,…, m_i, the model is defined by specifying the $\sum_{i=1}^{n} m_i$ probabilities:

$$P(y_i = j) = F_{ij}(x, \beta), \quad i = 1, 2, \ldots, n \quad j = 1, 2, \ldots, m_i. \tag{II}$$

Note that $P\ (y_i = 0)$ need not be specified because the sum of the $m_i + 1$ probabilities is

one for each *i*. It is important to let *m* depend on *i* because the number of alternatives available to different individuals may differ.

Defining $\sum_{i=1}^{n}(m_i+1)$ binary variables

$$\begin{aligned} y_{ij} &= 1 \text{ if } \quad y_i = j \\ &= 0 \text{ if } y_i \neq j, \quad i = 1,2,\ldots,n \\ &\qquad\qquad\quad j = 0,1,\ldots,m_j, \end{aligned}$$

the likelihood function of the model can be written as

$$L = \prod_{i=1}^{n}\prod_{j=0}^{m_i} F_{ij}(x,\beta)^{yij}.$$

Note that this reduces to the *L* equation of Section "Univariate Binary Models" if $m_i = 1$ for all *i*.

The ML estimator of β is consistent and asymptotically normal with its asymptotic variance–covariance matrix given by

$$V\widehat{\beta} = -\left[E\frac{\partial^2 \log L}{\partial\beta\partial\beta'}\right]^{-1},$$

which will be equal to $V\widehat{\beta}$ equation in Section "Univariate Binary Models" in the binary case. The MIN χ^2 estimator can be also defined for the multinomial model, although the definition will not be given here.

Ordered Models

An ordered multinomial model arises when there is an unobserved continuous random variable y_t^* which determines the outcome of y_i by the rule

$$\begin{aligned} &y_i = j \text{ if and only if } \alpha_j < y_t^* < \alpha_{j+1}, \\ &j = 0,1,\ldots,m, \quad \alpha_0 = -\infty, \quad \alpha_{m+1} = \infty. \end{aligned}$$

Such a rule may be appropriate, for example, if $y_i = j$ signifies the event that the *i*th individual owns *j* cars and y_t^* refers to a measure of the intensity of the *i*th individual's desire to own cars. If the distribution function of by $y_t^* - x_t'\beta$ is *F*, the last equation leads to an ordered model defined by

$$P(y_i = j) = F(\alpha_{j+1} - x_t'\beta) - F(\alpha_j - x_t'\beta).$$

As in the binary case, the choice of Φ and Λ for *F* is most frequently used.

An ordered model is attractive because of its simplicity. However, in many economic applications it may be an oversimplification to assume that the outcome of a multinomial variable can be completely determined by the outcome of a simple continuous variable. For example, for owning cars it is probably more realistic to assume that the *i*th person owns *j* cars if $U_{ij} > U_{ik}$ for all $k \neq j$, where U_{ij} is the utility that accrues to the *i*th person if he owns *j* cars. In this case *m* continuous variables $U_{ij} - U_{i,j+1}$, $j = 0, 1, \ldots, m$, determine the outcome of the discrete variable.

A multinomial model which is not an ordered model is called an unordered model. The models discussed in the next parts of this section are all unordered.

Multinomial Logit Model

A multinomial logit model is described below by defining the probabilities of the *i*th individual who faces three alternatives $j = 0$, 1, and 2. A generalization to the case of more alternatives can be easily inferred. The three probabilities are given by

$$\begin{aligned} P(y_i = 2) &= D^{-1}\exp(x_{i2}'\beta) \\ P(y_i = 1) &= D^{-1}\exp(x_{i1}'\beta), \\ P(y_i = 0) &= D^{-1}, \end{aligned}$$

where $D = 1 + \exp(x_{i1}'\beta) + \exp(x_{i2}'\beta)$.

McFadden (1974) showed how a multinomial logit model can be derived from the maximization of stochastic utilities. Suppose that the *i*th individual's utility U_{ij} associated with the *j* th alternative is the sum of the nonstochastic part μ_{ij} and the stochastic part ε_{ij} and that the individual chooses the alternative for which the utility is a maximum. Suppose further that ε_{i0}, ε_{i1} and ε_{i2} are independent and identically distributed according to the distribution function exp[−exp(−ε)] – called the

type I extreme value distribution. Then we can show

$$P(y_i = 2) = P(U_{i2} > U_{i1},\ U_{i2} > U_{i0}) = \exp(\mu_{i2})\ /[\exp(\mu_{i0})\ + \exp(\mu_{i1}) + \exp(\mu_{i2})],$$

and similarly for $P\ (y_i = 1)$ and $P\ (y_i = 0)$. Thus, the model defined by three equations above follows from putting $\mu_{i2} - \mu_{i0} = x'_{i2}\beta$ and $\mu_{i1} - \mu_{i0} = x'_{i1}\beta$.

The multinomial logit model has been extensively used in economic applications, such as the choice of modes of transportation, the choice of occupations, and the choice of types of appliances. The likelihood function of the model can be shown to be globally concave; consequently, the ML estimator can be computed with relative ease.

A major limitation of the multinomial logit model lies in its independence assumption. Consider the choice of transportation modes and suppose first that the alternatives consist of car, bus, and train. Then the assumption of independent utilities may be reasonable. Next, to use McFadden's famous example, suppose instead that the choice is among a car, a red bus, and a blue bus. Then it is clearly unreasonable to assume that the utilities associated with the red bus and the blue bus are independent. In the next subsection we shall consider a multinomial model which corrects this deficiency.

Nested Logit Model

We continue the last example. Let $U_j = \mu_j + \varepsilon_j$, $j = 0$, 1 and 2, be the utilities associated with car, red bus, and blue blus, respectively. (The subscript i is suppressed to simplify notation.) Following McFadden (1977), suppose that ε_0 is distributed according to the type I extreme value distribution and independent of ε_1 and ε_2 and that the joint distribution of ε_1 and ε_2 is given by

$$F(\varepsilon_1,\ \varepsilon_2)\ = \exp - \left[\exp\left(-\rho^{-1}\varepsilon_1\right)\right. + \left.\left[\exp\left(-\rho^{-1}\varepsilon_2\right)\right]^{\rho}, 0 \le \rho \le 1.$$

This distribution is called *Gumbel's type B bivariate extreme value distribution.* The correlation coefficient is $1 - \rho^2$, and if $\rho = 1$ (the case of independence), $F(\varepsilon_1,\ \varepsilon_2)$ becomes the product of two type I extreme value distributions.

Under these assumptions it can be shown that

$$P(y = 0) = \exp(\mu_0)\ /\left\{\exp(\mu_0)\ + \exp(\rho^{-1}\mu_1) + \exp\left(\rho^{-1}\mu_2\right)\right]^{P}\}$$

and

$$P(y = 1|y \neq 0) = \exp(\rho^{-1}\mu_1)\ /[\exp(\rho^{-1}\mu_1)\ + \exp(\rho^{-1}\mu_2)].$$

The other probabilities can be deduced from the above. Note that the last equation shows that the choice between red bus and blue bus is made according to a binary logit model, while the previous equation shows that the choice between car and noncar is also like a logit model except that a certain weighted average of $\exp(\mu_1)$and $\exp(\mu_2)$ is involved.

Multinomial Probit Model

A multinomial probit model is derived from the assumption that the utilities U_{i0}, U_{i1},. . ., U_{imi} are multivariate normal for every i. Its advantage is that general assumptions about the correlations among the utilities are allowed. Its major disadvantage is that the calculation of the choice probability requires the evaluation of multiple integrals of joint normal densities, which is feasible only for a small number of alternatives.

Multivariate Models

A multivariate discrete choice model specifies the joint probability distribution of two or more discrete dependent variables. For example, the joint distribution of two binary variables y_1 and y_2 each of which takes values 1 or 0 is determined by the four probabilities $P_{jk} = P\ (y_1 = j,\ y_2 = k)$, j, $k = 0$, 1. (Of course, the sum of the probabilities must be equal to 1.)

A multivariate model is a special case of a multinomial model. For example, the model of two binary variables mentioned in the preceding paragraph may be regarded as a multinomial model for a single discrete variable which takes four values with probabilities P_{11}, P_{10}, P_{01}, and P_{00}. Therefore, all the results given in section "Multinomial Models" apply to multivariate models as well. In this section we shall discuss three types of models which specifically take into account the multivariate feature of the model.

Log-Linear Model

A log-linear model refers to a particular parameterization of a multivariate discrete choice model. In the previous bivariate binary model, the log-linear parameterization of the four probabilitites is given as follows:

$$\begin{aligned} P_{11} &= D^{-1}\exp(\alpha_1 + \alpha_2 + \alpha_{12}), \\ P_{10} &= D^{-1}\exp(\alpha_1), \\ P_{01} &= D^{-1}\exp(\alpha_2), \end{aligned}$$

and

$$P_{00} = D^{-1}, \tag{III}$$

where $D = 1 + \exp(\alpha_1) + \exp(\alpha_2) + \exp(\alpha_1 + \alpha_2 + \alpha_{12})$

There is a one-to-one correspondence between any three probabilities and the three α parameters of the log-linear model; thus, the two parameterizations are equivalent. An advantage of the log-linear parameterization lies in its feature that $\alpha_{12} = 0$ if and only if y_1 and y_2 are independent.

Equations (III) may be represented by the following single equation:

$$P(y_1, y_2) \propto \exp(\alpha_1 y_1 + \alpha_2 y_2 + \alpha_1 \alpha_2 y_1 y_2).$$

Each equation of (III) is obtained by inserting values 1 or 0 into y_1 and y_2 in this equation. This formulation can be generalized to a log-linear model of more than two binary variables. The case of three variables is given below:

$$\begin{aligned} P(y_1, y_2, y_3) \propto \exp(&\alpha_1 y_1 + \alpha_2 y_2 + \alpha_3 y_3 + \alpha_{12} y_1 y_2 \\ &+ \alpha_{13} y_1 y_3 + \alpha_{23} y_2 y_3 \\ &+ \alpha_{123} y_1 y_2 y_3). \end{aligned}$$

The first three terms in the exponential function are called the main effects. Terms involving the product of two variables are called second-order interaction terms, the product of three variables third-order interaction terms, and so on.

Note that the last equation has seven parameters, which can be put into one-to-one correspondence with the seven probabilities that completely determine the distribution of y_1, y_2, and y_3. Such a model, without any constraint among the parameters, is called a *saturated* model. Researchers often use a constrained log-linear model, called an *unsaturated* model, which is obtained by setting some of the higher-order interaction terms to zero; e.g. Goodman (1972). See also Nerlove and Press (1973) for an example of a log-linear model in which some of the α parameters are specified to be functions of independent variables and unknown parameters.

Multivariate Nested Logit Model

The multivariate nested logit model is a special case of the nested logit model discussed in section "Nested Logit Model", which is useful whenever a set of alternatives can be classified into classes each of which contains similar alternatives. It is useful in a multivariate situation because the alternatives can be naturally classified according to the outcome of one or more of the variables.

For example, in the bivariate binary case, the four alternatives can be classified according to whether $y_1 = 1$ or 0. Let U_{jk} be the utility associated with the choice $y_1 = j$ and $y_2 = k$, $j, k = 0, 1$ and assume as before that $U_{jk} = \mu_{jk} + \varepsilon_{jk}$, where μ's are nonstochastic and ε's are random. As a slight generalization of the Gumbel distribution in section "Nested Logit Model" assume that

$$F(\varepsilon_{j1}, \varepsilon_{j0}) = a_j \exp\Big\{-\Big[\exp\Big(-\rho_j^{-1}\varepsilon_{j1}\Big) + \exp\Big(-\rho_j^{-1}\varepsilon_{j2}\Big)\Big]^{\rho j}\Big\}, \quad j = 1, 0$$

and that $(\varepsilon_{11}, \varepsilon_{10})$ are independent of $(\varepsilon_{01}, \varepsilon_{00})$. Then the resulting multivariate nested logit model is characterized by the following probabilities:

$$P(y_1 = 1) = a_1\left[\exp\left(\rho_1^{-1}\mu_{11}\right) + \exp\left(\rho_1^{-1}\mu_{10}\right)\right]^{\rho 1} \div \Big\{a_1\left[\exp\left(\rho_1^{-1}\mu_{11}\right) + \exp\left(\rho_1^{-1}\mu_{10}\right)\right]^{\rho 1} + a_0\left[\exp\left(\rho_0^{-1}\mu_{01}\right) + \exp\left(\rho_0^{-1}\mu_{00}\right)\right]^{\rho 0}\Big\},$$

$$P(y_2 = 1|y_1 = 1) = \exp\left(\rho_1^{-1}\mu_{11}\right)/\left[\exp\left(\rho_1^{-1}\mu_{11}\right) + \exp\left(\rho_1^{-1}\mu_{10}\right)\right],$$

$$P(y_2 = 1|y_1 = 0) = \exp\left(\rho_0^{-1}\mu_{01}\right)/\left[\exp\left(\rho_0^{-1}\mu_{01}\right) + \exp\left(\rho_0^{-1}\mu_{00}\right)\right].$$

We may further specify $\mu_{jk} = x'_{jk}\beta$.

Multivariate Probit Model

This model is conceptually different from the models of the preceding two sections in that here the marginal probabilities are specified first and the joint probabilities are then defined in a certain natural way.

As an example of a bivariate binary probit model, let us suppose $y_t^* \tilde{} N\left(\mu_j, 1\right)$, $j = 1$, and 2, and y_t^* is unobservable and its value determines the value of the observable binary variable y_j bythe rule

$$y_j = 1 \quad \text{if } y_j^* > 0$$
$$= 0 \,\text{otherwise.}$$

This rule determines the marginal probabilities

$$P(y_i = 1) = \Phi\left(\mu_j\right), j = 1 \text{ and } 2.$$

Thus, the model will be complete when we specify the joint probability $P\ (y_1 = 1,\ y_2 = 1)$. A natural way to specify it would be to assume that y_1^* and y_2^* are jointly normal with a correlation coefficient ρ and define

$$P(y = 1, y_2 = 1) = P\left(y_1^* > \mu_1, y_2^* > \mu_2\right).$$

Usually, a researcher will further specify $\mu_1 = x'_1\beta$ and $\mu_2 = x'_2\beta$ and estimate the unknown parameters β and ρ; see Morimune (1979) for an econometric example of this model.

A bivariate logit model may be defined similarly. But, unlike the probit case, there is no natural choice among many bivariate logistic distributions with the same marginal univariate logistic distributions.

Choice-Based Sampling

In models (I) or (II), the independent variables x_i were treated as known constants. This is equivalent to considering the conditional distribution of y_i given x_i. This practice was valid because it was implicitly assumed that y_i and x_i were generated according to either *random sampling* or *exogenous sampling.*

Under random sampling, y and x are sampled according to their true joint distribution $P\ (y|x)$ $f\ (x)$. Thus the likelihood function denoted L_R, is given by

$$L_R = \prod_{i=1}^{n} P(y_i|x_i)\, f(x_i).$$

Under exogenous sampling, a researcher samples x according to a certain distribution $g\ (x)$, which may not be equal to the true distribution $f\ (x)$ of x in the total population, and then samples y according to its true conditional probability $P\ (y|x)$. Thus the likelihood function, denoted L_E, is given by

$$L_E = \prod_{i=1}^{n} P(y_i|x_i)\ g(x_i).$$

In either case, as long as the parameters that characterize $P\ (y|x)$ are not related to the

parameters that characterize $f(x)$ or $g(x)$, the maximization of L_R or L_E is equivalent to the maximization of

$$L = \prod_{i=1}^{n} P(y_i|x_i) ,$$

which is equivalent to the L of Section "Multinomial Models".

Under choice-based sampling, a researcher samples y according to fixed proportions $H(y)$, and then, given y, samples x according to the conditional density $f(x|y)$. By the formula of conditional density,

$$f(x|y) = P(y|x)f(x)/Q(y),$$

where $Q(y) = E_x P(y|x)$, and E_x denotes the expectation taken with respect to the random vector x. Thus, the likelihood function under choice-based sampling, denoted L_c, is

$$L_c = \prod_{i=1}^{n} Q(y_i)^{-1} P(y_i \mid x_i) f(x_i) H(y_i).$$

Unlike random sampling or exogenous sampling, choice-based sampling requires new analysis because the maximization of L_c is not equivalent to the maximization of L on account of the fact that $Q(y)$ depends on the same parameters that characterize $P(y|x)$.

In particular, it means that the standard ML estimator which maximizes L is not even consistent under choice-based sampling. The reader should consult Amemiya (1985) or Manski and McFadden (1981) for the properties of the choice-based sampling ML estimator which maximizes L_c in various situations.

Choice-based sampling is useful when only a small number of people sampled according to random sampling are likely to choose a particular alternative. For example, in a transportation study, random sampling of individual households in a community with a small proportion of bus riders may produce an extremely small number of bus riders. In such a case a researcher may be able to attain a higher efficiency of estimation by sampling bus riders at a bus depot to augment the data gathered by random sampling.

An interesting problem in choice-based sampling is how to determine $H(y)$ to maximize the efficiency of estimation. Although there is no clear-cut solution to this problem in general, it is expected that if $Q(j)$ is small for some j then the value of $H(j)$ which is larger than $Q(j)$ will yield a more efficient estimator than the value of $H(j)$ which is equal to $Q(j)$. Note that if in the formula for L_c, $H(j) = Q(j)$ for every j, then L_c is reduced to L_R.

Distribution-Free Methods

Consider the univariate binary model (I). There, we assumed that the function $F(\cdot)$ is completely specified and known. Recently, Manski (1975) and Cosslett (1983) have shown how to estimate β consistently (subject to a certain normalization) without specifying $F(\cdot)$.

Manski's estimator is based on the idea that as long as F satisfies the condition $F(0) = 0.5$, one can predict y_i to be 1 or 0 depending on whether $x_i'\beta$ is positive or negative. His estimator of β is chosen so as to maximize the number of correct predictions. If we define the characteristic function χ of the event E by

$$\begin{aligned} \chi(E) &= 1 \text{ if } E \text{ occurs} \\ &= 0 \text{ otherwise,} \end{aligned}$$

the number of correct predictions can be mathematically expressed as

$$S(\beta) = \sum_{i=1}^{n} \left[y_i \chi\left(x_i'\beta \geqq 0\right) + (1 - y_i)\chi\left(x_i'\beta < 0\right) \right].$$

Manski calls this the score function – and hence his estimator the maximum score estimator. The estimator has been shown to be consistent, but its asymptotic distribution is unknown.

Cosslett proposed maximizing the likelihood function L in Section "Estimation" with respect to

both β and F, and called his estimator the generalized ML estimator. For a given value of β, the value of F which maximizes that L is a step function, and Cosslett showed a simple method of determining it. Finding the optimal value of β, however, is the computationally difficult part. Like the maximum score estimator, the generalized ML estimator of β is consistent but its asymptotic distribution is unknown.

See Also

- ▶ Censored Data Models
- ▶ Labour Supply of Women
- ▶ Limited Dependent Variables
- ▶ Logit, Probit and Tobit
- ▶ Selection Bias and Self-Selection

Bibliography

Amemiya, T. 1981. Qualitative response models: A survey. *Journal of Economic Literature* 19: 1483–1536.

Amemiya, T. 1985. *Advanced econometrics*. Cambridge, MA: Harvard University Press.

Berkson, J. 1944. Application of the logistic function to bioassay. *Journal of the American Statistical Association* 39: 357–365.

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

Goodman, L.A. 1972. A modified multiple regression approach to the analysis of dichotomous variables. *American Sociological Review* 37: 28–46.

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

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

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

McFadden, D. 1974. Conditional logit analysis of qualitative choice behavior. In *Frontiers in econometrics*, ed. P. Zarembka, 105–142. New York: Academic Press.

McFadden, D. 1977. Qualitative methods for analyzing travel behavior of individuals: Some recent developments. Cowles Foundation Discussion Paper No. 474.

McFadden, D. 1984. Econometric analysis of qualitative response models. In *Handbook of econometrics*, vol. 2, ed. Z. Griliches and M.D. Intriligator, 1385–1457. Amsterdam: North-Holland.

Morimune, K. 1979. Comparisons of normal and logistic models in the bivariate dichotomous analysis. *Econometrica* 47: 957–976.

Nerlove, M., and S.J. Press. 1973. *Univariate and multivariate log-linear and logistic models*, R-1306-EDA/NIH. Santa Monica: Rand Corporation.

Discriminating Monopoly

John M. Hartwick

In *The Wealth of Nations,* Adam Smith refers to two instances of price discrimination. In Book V, Chapter I, Part III, he ruminates on the problem of finding the best set of levies for toll roads and commends the practice of charging for luxurious carriages more than for working men's wagons even though the vehicles are of the same weight. He suggests that the rich can subsidize the poor by this tariff scheme. In Book IV, Chapter V, he notes that some groups of producers have sold their produce abroad at lower prices than at home. He views this as cross-subsidization and deplores the high prices which he sees as resulting in the domestic market. Smith's first problem, how to set tolls, has occupied economists to this day, although the solution was laid out in principle by Dupuit (1844) and with considerable precision by Edgeworth (1910): let each user's levy in excess of his or her marginal cost of usage (which may be zero on a toll bridge) be proportional to his or her intensity of preference as expressed by his or her elasticity of demand. Edgeworth in fact worked out details of two sorts of price discrimination – that practised by a private profit-maximizing monopolist and that practised by a 'state monopoly' interested in raising Z dollars of profit from the users of the monopoly while at the same time reducing welfare as little as possible. The solution to this state monopoly or public utility pricing problem we refer to today as Ramsey pricing (Ramsey 1927, who attributes the idea for his paper to Pigou).

Smith's second instance of price discrimination can most usefully be viewed as the case of a monopolist selling at distinct prices in two

separate markets, domestic and foreign, with distinct demand curves. Barone (1921, pp. 291–2) analysed it from this perspective diagrammatically and Yntema (1928) filled in the algebraic details.

Pigou (1920) presented his synthesis of results and introduced the terms first, second and third degree price discrimination – degrees referring to the fineness with which separate prices can be assigned to separate units demanded of the monopolist. He graphically worked out the two market case with linear demands in his Appendix III, pointing out that given two markets, only one might be served under uniform pricing whereas both might be served under price discimination. In 1904 he had in fact independently of Dupuit analysed what we now call perfect price discrimination, or the situation in which each unit produced by a monopolist fetches a different price, each of which being bounded above by the buyer's willingness to pay. In her synthesis, J. Robinson (1933, p. 205) asked whether the monopolist would produce more under third degree price discrimination relative to his output under a uniform price.

The toll-setting problem turns on the fact that if users were charged commensurate with the wear and tear they cause (marginal cost pricing), insufficient revenue would generally be raised to cover the cost of building on obviously desirable road, bridge, railroad, telephone network etc. For example, the wear and tear caused by the marginal bridge user is approximately zero and thus no revenue would be raised by charging according to costs of usage. Dupuit realized that an individual's willingness to pay for a trip could far exceed his or her incremental cost of usage and suggested collecting revenue on the basis of each individual's maximum willingness to pay. The revenue so collected 'would not have the slightest relation to the costs of production' (p. 271) but would reflect the total utility in dollars per day's use of the project. This is price discrimination: each user pays generally a different price for the same service.

To sharpen his exposition Dupuit turned to shipping tons of ore across a bridge. At high prices obviously fewer tons will be shipped, since buyers of the ore will be obliged to absorb the charges and will demand less at high prices inclusive of delivery charges. Dupuit discussed the hypothetical case of each additional ton crossing the bridge 'paying' a slightly lower toll evaluated at the maximum willingness to pay for the ton in question. This is the case of perfect price discrimination and a variant is practised in the form of *block pricing*. Firms occasionally sell the first say 1000 bricks at \$3 each, the second thousand at \$2.50 each, the third thousand at \$2.25 and so on. Robinson reflected on the issue of the monopolist brick seller selecting the break points (1000 bricks, 2000 bricks etc.) simultaneously with price per brick in order to maximize profits. One can see that perfect price discrimination is a procedure for transferring consumer surplus (the area under an individual's demand curve up to the quantity consumed less the amount actually paid) to the seller of the product.

Price discrimination is practised by a monopolist because it permits profits to rise above what they would be if a single or uniform price were charged. To see this suppose that in two separate markets the monopolist were practising price discrimination and maximizing aggregate profit. If he were now obliged to sell in both markets at a single uniform price his optimand can never rise since the single price represents a new constraint on his pursuit of maximum profit. Pursuing this case in more detail, let $Q_1(p)$ be the demand curve for gadgets by citizens abroad (or for return rail car trips for wheat shippers on a line) and $Q_2(p)$ the demand curve for gadgets from local people (or for return rail car trips for potash producers located in the wheat farming area). Then the monopolist's profit under price discrimination is $\pi = p_1Q_1(p_1) + p_1Q_2(p_2) - C(Q_1 + Q_2)$ where $C(\cdot)$ is total cost, increasing and convex in $Q = Q_1 + Q_2$, and p_i is the price in market i with quantity sold $Q_i(p_i)$. Profits attain a maximum when $C_Q = p_i[1 + (1/\in_i)]$, where

$$C_Q \equiv dC/dQ \text{ and } \in_i = \frac{\mathrm{d}Q_i}{\mathrm{d}p_i} \cdot \frac{p_i}{Q_i} < -1$$

is the elasticity of demand in market *i*. This profit-maximizing condition is referred to as the

Robinson–Yntema condition and was first set out by Edgeworth (1910). The left-hand side is marginal cost and the right-hand side is the marginal revenue in market *i*. For $|\in_1| \gtreqless |\in_2|$, $p_1 \lesseqgtr p_2$. For $n > 2$, the analysis is the same. (Edgeworth made C_Q a constant at *c,* defined his elasticity as

$$\frac{\mathrm{d}Q}{\mathrm{d}(p_i - c)} \cdot \frac{(p_i - c)}{Q}$$

and arrived at his 'equal elasticity condition' for profit-maximizing monopoly price discrimination.) A monopolist forced to sell at a single price (presumably because the product can be readily resold) will maximize profit when $C_Q = p\{1 + [1/(\omega_1 \in_1 + \omega_2 \in_2)]\}$, where $\omega_i = Q_i/Q$.

Edgeworth investigated when deviations in p_1 and p_2 from a uniform p would increase welfare (consumer surplus), while Robinson argued that price discrimination would raise Q from the level corresponding to a uniform profit-maximizing price if the more elastic demand curve is concave and the less elastic demand curve is convex.

The basic first order condition for monopoly price discrimination can be written as

$$\frac{\Delta Q_1}{Q_1} = \frac{\Delta Q_2}{Q_2} = 1,$$

where

$$\Delta Q_i = \left[p_i - C_Q\right] \frac{\mathrm{d}Q_i}{\mathrm{d}p_i}.$$

This illuminating formula indicates that each output would to a first approximation rise proportionately if there were no monopoly and no price discrimination, and it can orient one's intuition in viewing Robinson's result on concavity and convexity of demand schedules. Schmalensee (1981) and Varian (1985) have pointed out that a necessary condition for total net consumer surplus to rise as the monopolist switches from a uniform price to profit-maximizing price discrimination is that there be a rise in total output delivered.

The public utility or state monopoly problem in price discrimination is to raise Z dollars of profit by charging diverse prices to distinct customers while reducing welfare least. These profits might be assigned to cover the fixed costs of a public facility. Let $B_1(p)$ and $B_2(p)$ be the areas under each demand curve for price p. The state monopoly pricing problem is to maximize $W = B_1(p_1) + B_2(p_2) - C(Q_1 + Q_2)$ subject to $Z = p_1Q_1 + p_2Q_2 - C(Q_1 + Q_2)$. The first order condition can be expressed as

$$\frac{\Delta Q_1}{Q_1} = \frac{\Delta Q_2}{Q_2} = \lambda,$$

where λ is a function of the level Z of profit sought and the ΔQ_i's were defined above. This problem is an instance of Ramsey (1927) optimal excise tax analysis and has been put into a general equilibrium context in Hartwick (1978). Contemporary price discrimination schemes (e.g. Oi 1971) have incorporated income elasticities of demand as well as price elasticities and arrived at two-part tariffs involving a 'membership' fee and a user's fee for service.

In closing, we note that high prices in peak times and low in off-peak times are not forms of monopoly price discrimination, since such peak-load prices vary with changes in the marginal costs of production. Under price discrimination, prices deviate from marginal costs of production only in accord with variations across buyers in their 'intensities' of demand. Marginal cost remains the same for each buyer.

See Also

- Basing Point System
- Consumer Surplus
- Monopoly
- Price Discrimination
- Public Utility Pricing

References

Barone, E. 1921. Les syndicats (cartels et trusts). *Revue de Métaphysique et de Morale* 28(2): 279–309.

Dupuit, J. 1844. On the measurement of the utility of public works. Reproduced in *Readings in welfare*

economics, ed. K.J. Arrow, T. Scitovsky, and American Economic Association. Homewood: Irwin, 1969.
Edgeworth, F.Y. 1910. Applications of probabilities to economics. *Economic Journal* 20: 284–304, 441–465. Reprinted in F.Y. Edgeworth, *Papers relating to political economy*, vol. II. New York: Burt Franklin, 1925.
Hartwick, J.M. 1978. Optimal price discrimination. *Journal of Public Economics* 9(1): 83–89.
Oi, W.Y. 1971. A Disneyland dilemma: Two-part tariffs for a Mickey Mouse monopoly. *Quarterly Journal of Economics* 85(1): 77–96.
Pigou, A.C. 1904. Monopoly and consumers' surplus. *Economic Journal* 14: 388–394.
Pigou, A.C. 1920. *The economics of welfare*. London: Macmillan.
Ramsey, F. 1927. A contribution to the theory of taxation. *Economic Journal* 37: 47–61.
Robinson, J. 1933. *The economics of imperfect competition*. London: Macmillan.
Schmalensee, R. 1981. Output and welfare implications of monopolistic third-degree price discrimination. *American Economic Review* 71(1): 242–247.
Smith, A. 1776. *The wealth of nations*. Ed. E. Cannan, reprinted, New York: Modern Library, 1937.
Varian, H. 1985. Price discrimination and social welfare. *American Economic Review* 75(4): 870–875.
Yntema, T. 1928. The influence of dumping on monopoly price. *Journal of Political Economy* 36(6): 686–698.

Discrimination

Peter Mueser

Discrimination may be said to occur in a market where individuals face terms of trade that are determined by personal characteristics which do not appear directly relevant to the transaction. Most concern has centred on differential treatment by race or ethnic group, and by sex. The primary focus has been on the labour market and housing market, with research motivated, in large part, by controversy over the role of government in maintaining or eliminating observed differentials.

The first extensive literature on the economics of discrimination dates to the equal pay controversy in Britain beginning before the turn of the century, focusing on the lower wages of women. Although interest in the economics of pay differentials by sex abated in the two decades following World War II, many aspects of the more recent theory appeared in this literature. The modern development of systematic models of economic discrimination began with the publication of Gary Becker's *The Economics of Discrimination* (1957). With the passage of laws prohibiting discrimination in the US, Britain and other countries in the 1960s and 1970s, research in the area has again grown.

Market Discrimination and Personal Preferences

Becker's (1957) treatment took market discrimination to be the result of personal tastes of participants, providing a simple, closed model with a variety of testable implications. Earnings differentials, and discrimination in housing and other markets stem from the attempts of owners, workers and customers to avoid contact or interaction with certain groups.

Consider first the influence of employer preferences. Rather than maximizing profits, employers maximize a utility that incorporates the personal characteristics of employees. If employers prefer to hire workers from group A rather than group B and are willing to sacrifice profits to do so, they may be said to have discriminatory tastes. Where such employers dominate the market, the relative wages of group B workers must adjust downward if any are to be hired, and the resulting difference equals the pecuniary value of the employed preference. Where there are variations in taste among employers, relative wages are determined by the shape of the taste distribution, and the proportion of A and B workers to be hired.

Employees may also be taken to have discriminatory preferences over the group membership of their co-workers. If discriminating workers and members of the group they shun are perfect substitutes in production, employers have an incentive to provide separate facilities for groups, but no wage differential between groups will occur. In order for such preferences to cause wage differentials, the technology of production must preclude complete separation. For example, if it is

necessary for supervisors to interact with assembly line workers, and supervisors prefer one group of workers to another, an employer who has no taste for discrimination and faces a competitive labour market will hire members of both groups only if their wages differ correspondingly.

Customers' tastes may also influence market wages in the absence of employer preferences. The extension of this approach to a variety of markets is clear. Housing discrimination would occur if owners required a premium in order to sell or rent to individuals in certain groups. However, in markets for goods, in contrast to services, no appreciable price differential could survive unless restrictions on resale were binding.

In any of these cases, some market participants may have a taste for discrimination without market differentials occurring if there are a sufficient number of non-discriminating participants to interact with the disliked group. One obvious source of non-discriminating participants is fellow group members. If groups are equally represented among employers, various kinds of workers, and customers, complete segregation can occur without economic loss to any group. Preferences against trading with those outside one's own group can only affect terms of trade where groups have different resources or skills. The formal model is, however, silent on the source of such differences.

Competition and Discriminatory Preferences

It is widely argued that, in the long run, competition in markets for output and capital will drive out discriminating employers. Since discrimination by race and sex has long existed, this result has frequently been taken as grounds for rejecting the model (Arrow 1972). In fact, this conclusion follows only from a particular version of the model, in which the taste for discrimination imposes a direct utility loss on the employer for each employee hired from the disliked group. Since non-discriminating employers suffer no such utility loss, under free competition, where they can expand production or buy out discriminators, they will take over the market. This need not be the case. A polar example has been termed nepotism, in which preferences for hiring one group act much like a net subsidy for the employer. In this case, those with the strongest discriminatory preferences ultimately dominate the market.

In general, discriminating employers earn lower money profits than those who do not discriminate, but this does not imply that under competition they will be driven from the market. Foregone profits must be recognized as consumption expenditures, and, so long as employer resources are sufficient to permit any consumption, there is no inconsistency between perfect competition and the existence of stable, long-run wage differentials stemming from employer preferences.

While market discrimination may survive competition, restrictions on competition will often result in more severe discrimination. In a competitive market, any market differential translates into a pecuniary cost that the discriminating employer must pay. But where prices do not equalize supply and demand, the employer's cost of discrimination may decline.

An effective minimum wage or, during times of economic decline, a wage that is downwardly rigid, allows the employer to hire both more productive workers and those most preferred without paying a higher wage. In contrast, if wages do not adjust immediately in an economic upturn, the cost of discrimination will increase.

Where a union successfully bargains for wages above the competitive level, discriminatory hiring, on the basis of either employers' or union members' preferences may take place at lower cost. In fact, discrimination against blacks by unions in the US was explicit and widespread until the 1930s, but by the 1960s union representation for blacks and whites was nearly identical. Black representation has been greatest in industrial unions, where unionization often hinges on the ability to organize both black and white workers, and proportionally lowest in craft unions, where unions frequently exercise power by limiting membership. Despite the historically high level of union discrimination, it appears to

have contributed relatively little to observed black–white wage differentials in the US.

The cost of discrimination to firms may decline where there are restrictions on profit maximization. Those who manage non-profit organizations, regulated monopolies or government bureaucracies will devote more resources to improving their own working conditions; unless faced with direct constraints, they will be more likely to exercise personal preferences in the kinds of workers they hire (Alchian and Kessel 1962).

Finally, in markets with search costs, discrimination may occur even if there are sufficient non-discriminating participants to trade with members of the disliked group, since the appropriate matching cannot occur.

Discrimination as Exploitation

It is frequently asserted that discrimination is engaged in because it is profitable. In general, there is some level of discrimination by members of any group that will improve the terms of trade so as to increase their money incomes. Discrimination by white employers, under some conditions, may increase the incomes of whites by increasing both employer profits and the earnings of white workers. Similarly, tastes that restrict blacks and women to certain kinds of jobs may increase money income both for employers and white male employees.

If discriminators' preferences are taken seriously, however, the impact of discriminatory preferences on money income is irrelevant. Although Becker's original treatment calculated such welfare effects, like any such comparison it required an arbitrary normalization to compare individuals with differing preferences. Changes in money income due to discrimination can be taken to represent group welfare if discriminatory behaviour does not reflect actual personal preference. However, individual incentives in a competitive market can no longer explain discrimination, since those who discriminate least receive the greatest gains. It is necessary that some process exist by which the group enforces its will on individual choices. Exploitation must have its roots in a social or political process.

Historically, there is no question that the enactment of discriminatory laws and provision of unequal public services has often represented the exploitation of groups with little political power. J.S. Mill (1869) argued that limitations on women's legal rights and the restrictions they faced in entering certain occupations, were part of a policy to provide men both with higher earnings in the labour market and greater authority over their wives at home. The history of governmental action regarding blacks in America since the Civil War is replete with examples of policies designed to benefit whites with political power at the expense of blacks.

Despite the government's often central role in furthering dominant group interests, there are clearly other channels by which groups exercise influence. An ethnic group is generally bound together by an ideology that dictates members' actions in a wide variety of contexts. Although some members may internalize such ideology, compliance is enforced by systems of social norms and sanctions within the group. The process by which such systems develop is not well understood, although it has been shown that discriminatory norms may be self-enforcing once established (Akerlof 1976). Nonetheless, it is clear empirically that economic relations among groups, and their relations within the power structure, are critical in determining group ideology and, in turn, individual actions.

For example, it is a recurrent observation that severe ethnic or racial antagonism often can be traced to the point at which groups first find themselves competing in the labour market. Some writers have argued that all discrimination by race or ethnic group can be traced to such a dynamic, in which groups mobilize political and economic resources to further their material interests. The goal of such action is seen to be the exclusion of the competing group from the labour market or, failing this, the creation of a caste system providing the dominant group with preferential treatment (Bonacich 1972).

As a rule, it is among lower income groups that racism appears most virulent and associated

violence most common. In part, this reflects the fact that racism is a source of power to those groups whose alternatives are limited. In some measure, social norms, personal animosity and collective violence substitute for political power and state action.

In contrast to the assumption of the preference-based model, the treatment here implies that discriminatory preferences cannot be taken as exogenous. Tastes, or apparent tastes, may develop to further group interests. This is not to say that individual actions are ever completely determined by group interests, even where these are unambiguous. Within the most tightly structured groups, for example where ethnic identity is strong, discriminatory collusion against outsiders relies heavily on the availability of explicit policing mechanisms. Where individual behaviours are difficult to observe, and the benefits of violating collusive rules are great, discrimination will be less successful.

It must be stressed that many of the conclusions of the taste-based model may apply even where groups' interests play a critical role in shaping individuals' actions. For example, the model tells us that if white workers who compete with black workers merely refuse to work with them, white workers obtain no net gain in income. It is only through the adoption of discriminatory practices by employers that white workers realize gains.

Statistical Discrimination

Participants in a market have an incentive to consider personal characteristics if these provide information that is relevant to the exchange but costly to obtain by other means. Statistical discrimination occurs where an ascribed characteristic serves this function. The widely accepted use of sex in markets for various kinds of insurance is an obvious example. Markets for credit and rental housing have similar structures, as does the market for labour. Initial screening is particularly critical in hiring for entry level positions in firms with internal labour markets, where a firm often undertakes extensive worker training, and implicit contracts limit the employer's ability to adjust wages in accord with realized productivity.

Some labour market analysts have attempted to limit the term statistical discrimination to contexts in which an employer distinguishes groups that do not differ in average productivity (Aigner and Cain 1977). For example, where an employer favoured men over women because women were more likely to quit after receiving firm-specific training, this would not be labelled discrimination. In contrast, statistical discrimination would be said to occur if an employer screened by race in jobs where expected ability was critical because he was unable to judge the abilities of blacks. Such a distinction becomes muddied when it is recognized that matching the worker to the job is part of the productive process.

In certain respects, observed patterns of employment for women and blacks are consistent with statistical discrimination. Both are seriously under-represented in jobs offering extended promotion ladders, and, historically, firms often explicitly reserved for white males the training that prepared an employee for promotion.

Since such persistent statistical discrimination results from the efficient use of information, the basis of wage differentials would seem to rest on pre-market influences, not market dynamics. However, it is possible that group differences are themselves the result of employer expectations. Assume employers believe that members of a particular group have lower levels of the skills necessary for success in screened jobs. In so far as performance is ultimately rewarded for those placed in screened positions, members of this group, because they are less likely to be hired into such positions, will have reduced incentives to invest in relevant skills. Any one employer who hired members of that group into these positions would find workers to be less productive, so beliefs would be confirmed.

In addition to a number of technical conditions (Arrow 1972), in order for such a 'self-fulfilling prophecy' to be stable, the actions of a single firm must not alter individual incentives. If it were possible for a firm to contract with individuals to fill positions prior to the point when they acquire such skills, individuals who entered contracts,

whatever their group membership, would face the same incentives to obtain skills, and the vicious circle would be broken. The acquisition of such skills must therefore occur well in advance of the point that individual workers and firms can easily enter into agreements. Differences in socialization by sex, or cultural differences by race or ethnic group, if in response to disparate treatment in the labour market, could reflect this kind of vicious circle.

Explaining Market Differentials

Earnings for women have been appreciably below those of men in almost all societies, past and present. Differences in levels of market participation and other observable personal attributes explain only a portion of the differential. Historically, some of the difference may be identified with governmental or institutional discrimination. Nevertheless, the enactment of laws in many countries prohibiting discrimination in the 1960s and 1970s have had little effect on the overall distribution of wages by sex. There is no obvious way to identify the impact of discrimination in explaining observed differences. Any unobserved direct market discrimination may induce differences in labour market participation and measurable pre-market factors, yet any unmeasured differences between men and women that are not due to discriminatory treatment, may also contribute to the wage differential.

Women have historically performed the bulk of household work and child care, participating in the labour market less continuously and less intensively than have men. Given this division of labour within the family, women who expect to marry have less incentive to develop skills requiring continuous labour market participation. Non-discriminating employers may simply pay women less because they have not invested in those skills that are most valuable in the market.

How the earnings gap is viewed must depend partly on the source of the family's division of labour. If it results from market discrimination, or social norms constructed to benefit dominant males, it may be analysed in terms of the models of discrimination. However, such a division is also consistent with joint optimization by husband and wife: if the bearing and rearing of children are even weakly complementary, it is efficient for the family to have the women specialize in both these non-market tasks. Sex-typed socialization would then merely reflect preparation for anticipated roles.

That perfect equality would occur in the absence of all labour market discrimination seems unlikely. Nonetheless, unless there are strong sanctions, employers have an incentive to practise statistical discrimination, magnifying whatever sex differences would occur in its absence.

The labour market disadvantages suffered by many ethnic and racial groups is similarly open to interpretation. Thomas Sowell (1981) has argued that cultural differences between arriving immigrant groups and blacks in the US are more important in explaining their economic progress than the levels of discrimination they faced. While it is clear that cultural factors are critically important, the degree to which these or other pre-market differences explain observed earnings differentials is unclear. The theory implies that discrimination will be most common and most damaging against groups with low levels of resources, those who would be disadvantaged in its absence.

For blacks in the US, slavery and subsequent governmental discrimination induced shortfalls in human resources that would have limited black achievement under the best of conditions. Nonetheless, up through the 1960s, measured pre-market differences explained only a modest portion of observed earnings differentials. Although unobserved premarket differences may have played a role, given the pervasiveness of explicit market discrimination, it seems likely that discrimination further depressed the black position. To what degree labour market differences that persist despite the prohibition of discrimination since 1965 – most notably in rates of unemployment – are due to unmeasured pre-market differences, possibly associated with statistical discrimination, or to other market discrimination, is an open question.

See Also

- ▶ Equality
- ▶ Gender
- ▶ Human Capital
- ▶ Inequality Between Persons
- ▶ Inequality Between the Sexes
- ▶ Labour Supply of Women
- ▶ Signalling
- ▶ Women's Wages

Bibliography

Akerlof, G. 1976. The economics of caste and of the rat race and other woeful tales. *Quarterly Journal of Economics* 90(4): 599–617.

Aigner, D., and G. Cain. 1977. Statistical theoies of discrimination in labor markets. *Industrial and Labor Relations Review* 30(2): 175–187.

Alchian, A., and R. Kessel. 1962. Competition, monopoly, and the pursuit of pecuniary gain. In *Aspects of labor economics*, NBER Special Conference Series. Princeton: Princeton University Press.

Arrow, K. 1972. Models of job discrimination. In *Racial discrimination in economic life*, ed. A. Pascal. Lexington: D.C. Heath.

Becker, G. 1957. *The economics of discrimination*, 2nd ed. Chicago: University of Chicago Press, 1971.

Bonacich, E. 1972. A theory of ethnic antagonism: The split labor market. *American Sociological Review* 37: 547–559.

Mill, J.S. 1869. *The subjection of women*. New York: Stokes, 1911.

Sowell, T. 1981. *Ethnic America: A history*. New York: Basic Books.

Disequilibrium Analysis

Jean-Pascal Benassy

A convenient way to define 'disequilibrium' is of course as the contrary of 'equilibrium'. Unfortunately this leaves us with no unique definition as the word equilibrium itself has been used in the economic literature with at least two principal meanings. The first one refers to market equilibrium, i.e. the equality of supply and demand on markets. This is the meaning we shall retain in this entry, and therefore the disequilibrium analysis we shall be concerned with here is the study of nonclearing markets, also called non-Walrasian analysis by reference to the most elaborate model of market clearing, the Walrasian model.

The second meaning of equilibrium is somewhat more general. A typical definition is given by Machlup (1958) as '... a constellation of selected interrelated variables, so adjusted to one another that no inherent tendency to change prevails in the model which they constitute'. Dealing with disequilibrium in this second meaning would be a quite formidable (and actually extremely imprecise) task, which is why we want to limit ourselves in this entry to disequilibrium analysis in the first sense.

We should note that the entry RATIONED EQUILIBRIA presents concepts of equilibria in the second, but not in the first sense of the word, i.e. more specifically equilibria without market clearing or non-Walrasian equilibria.

The Essence of the Theory

Disequilibrium analysis is best appraised by reference to the standard equilibrium market clearing paradigm, corresponding to the notions of Marshallian or Walrasian equilibrium. There all private agents receive a price signal and assume that they will be able to exchange whatever they want at that price. They express demands and supplies, sometimes called 'notional', which are functions of this price signal. An equilibrium price system is a set of prices for which demand and supply match on all markets. Transactions are equal to the demands and supplies at the equilibrium price system.

Two characteristics deserve to be stressed: all private agents receive price signals and make rational quantity decisions with respect to them. But no agent makes any use of the quantity signals sent to the market. Also no agent actually sets prices, the determination of which is left to the 'invisible hand' or to the implicit Walrasian auctioneer. This logical hole of the theory was pointed out by Arrow (1959) when he noted

there was '. . . a logical gap in the usual formulations of the theory of the perfectly competitive economy, namely, that there is no place for a rational decision with respect to prices as there is with respect to quantities', and more specifically 'each individual participant in the economy is supposed to take prices as given and determine his choices as to purchases and sales accordingly; there is no one left over whose job is to make a decision on price'.

Disequilibrium analysis takes this strong logical objection quite seriously, and its purpose is to build a consistent theory of the functioning of decentralized economies when market clearing is not axiomatically assumed. The consequences of abandoning the market clearing assumption are actually quite far-reaching: (i) The transactions cannot be all equal to demands and supplies expressed on markets. Rationing will be experienced and quantity signals will be formed in addition to price signals (ii) Demand and supply theory must be substantially modified to take into account these quantity signals. One thus obtains a theory of effective demand, as opposed to notional demand which only takes price signals into account (iii) Price theory must also be amended in a way that integrates the possibility of non-clearing markets, the presence of quantity signals, and makes agents themselves responsible for price making, (iv) Finally expectations, which in market clearing models are concerned with price signals only, must now include quantity signals expectations as well.

History

Though roots may be found earlier, an uncontestable grandfather of disequilbrium analysis in the sense we use here is of course Keynes (1936). He rightfully perceived that one of his main contributions in the *General Theory* was the introduction of quantity adjustments, and more specifically income adjustments, in the economic process, whereas the then dominant 'classical' economists focused on price adjustments only. As Keynes (1937) wrote 'As I have said above, the initial novelty lies in my maintaining that it is not the rate of interest, but the level of incomes which ensures equality between savings and investment'.

Unfortunately for many decades things did not go much further: macroeconomists added the level of income in their equations, thereby allowing for unemployment. But concentration on the 'equilibrium' of the goods and money markets, exemplified by the dominant IS–LM model, obscured the 'disequilibrium' nature of the model. As for microeconomics, it was basically unaffected by the Keynesian revolution, and correlatively a growing gap developed between microeconomics and macroeconomics.

A few isolated contributions in the post-war period made some steps toward modern disequilibrium theories. Samuelson (1947), Tobin and Houthakker (1950) studied the theory of demand under conditions of rationing. Bent Hansen (1951) introduced the ideas of active demand, close in spirit to that of effective demand, and of quasi-equilibrium where persistent disequilibrium created steady inflation. Patinkin (1956, ch. 13) considered the situation where the firms might not be able to sell all their 'notional' output. Hahn and Negishi (1962) studied non-tâtonnement processes where trade could take place before a general equilibrium price system was reached. Hicks (1965) discussed the 'fixprice' method as opposed to the flexprice method.

A main impetus came from the stimulating works of Clower (1965) and Leijonhufvud (1968). Both were concerned with the microeconomic foundations of Keynesian theory. Clower showed that the Keynesian consumption function made no sense unless reinterpreted as the response of a rational consumer to a disequilibrium on the labour markets. He introduced the 'dual-decision' hypothesis, a precursor of modern effective demand theory, showing how the consumption function could have two different functional forms, depending on whether the consumer was rationed on the labour market or not. Leijonhufvud (1968) insisted on the importance of short-run quantity adjustments to explain the establishment of an equilibrium with involuntary unemployment.

These contributions were followed by the macroeconomic model of Barro and Grossman (1971,

1976), integrating the 'Clower' consumption function and the 'Patinkin' employment function in the first 'disequilibrium' macroeconomic model.

Then the main development was that of microeconomic concepts of non-Walrasian equilibrium proposed notably by Benassy (1975, 1976, 1977, 1982), Drèze (1975) and Younès (1975). These, which generalize the notion of Walrasian general equilibrium to non-market clearing situations, gave solid microeconomic foundations to the field. The main concepts are reviewed in the entry RATIONED EQUILIBRIA.

From then on, the field has developed quite rapidly, notably in the direction of macroeconomic applications and econometrics, as we shall outline below. We shall now review quickly the main elements of disequilibrium analysis. Longer developments can be found notably in Benassy (1982).

Non-Clearing Markets and Quantity Signals

A most important element of the theory is obviously to show how transactions can occur in a market in disequilibrium, and how quantity signals are generated in the decentralized trading process. To make things clear and intuitive, we shall start with the simple case of two agents, one demander and one supplier in a market which does not necessarily clear. They meet and express, respectively, an effective demand $\tilde{d}$ and supply $\tilde{s}$ (note that we do not use notional demands and supplies which are fully irrelevant in this context). We shall now indicate how transactions and quantity signals are formed in this example. Transactions will be denoted d^* and s^* respectively and they must of course satisfy $d^* = s^*$.

The first principle we shall use is that of voluntary exchange, i.e. that no agent can be forced to trade more than he wants on a market. This condition is quite natural and actually verified on most markets, except maybe for some labour markets which are regulated by more complex contractual arrangements. It is written in this example:

$$d^* \leq \tilde{d} \quad s^* \leq \tilde{s}$$

which implies that:

$$d^* = s^* \leq \min\ (\tilde{d}, \tilde{s})$$

Actually in this simple example, there is not reason why these two agents would exchange less than the minimum of demand and supply, as they would be both frustrated in their desires of exchange. This simple 'efficiency' assumption leads us to take the transaction as:

$$d^* = s^* = \min\ (\tilde{d}, \tilde{s})$$

the well-known 'rule of the minimum'.

Now at the same time as transactions take place, quantity signals are set across the market: faced with the supply $\tilde{s}$, and under voluntary exchange, the demander knows that he will not be able to purchase more than $\tilde{s}$.Symmetrically the supplier knows that he cannot sell more than $\tilde{d}$. Each agents thus receives from the other a quantity signal, respectively denoted as $\overline{d}$ and $\overline{s}$, which tells him the maximum quantity he can respectively buy and sell. In this example:

$$\overline{d} = \tilde{s} \quad \overline{s} = \tilde{d}$$

and the transactions can thus be expressed as:

$$d^* = \min(\tilde{d}, \overline{d})$$
$$s^* = \min(\tilde{s}, \overline{s})$$

Let us move now to the general case (which is explored more formally in the entry on RATIONED EQUILIBRIA). Agents, indexed by *i,* exchange goods indexed by *h.* On each market *h* a rationing scheme transforms inconsistent demands and supplies, denoted $\tilde{d}_{ih}$ and $\tilde{s}_{ih}$, into consistent transactions, denotedand d^*_{ih} and s^*_{ih}, which balance identically (i.e. total purchases always equal total sales). At the same time, and continuing to assume voluntary exchange, each agent receives a quantity signal, respectively $\overline{d}_{ih}$ or $\overline{s}_{ih}$ for demanders and suppliers, which tells him

the maximum quantity he can buy or sell, and the rationing scheme is equivalently written:

$$d_{ih}^{*} = \min\left(\tilde{d}_{ih}, \overline{d}_{ih}\right)$$
$$s_{ih}^{*} = \min\left(\tilde{s}_{ih}, \overline{s}_{ih}\right)$$

where $\overline{d}_{ih}$ and $\overline{s}_{ih}$ are functions of the demands and supplies of the other agents on the market. These quantity signals may result from the signals sent to each other by agents in decentralized pairwise meetings (as in the above two agents example) or result from a more centralized process (as in a uniform rationing scheme).

We thus see that on a market we may have unrationed demanders or suppliers, or rationed ones. The rationing scheme is called efficient if there are not both rationed demanders and rationed suppliers in the same market. An efficient rationing scheme implies the well-known 'rule of the minimum', according to which aggregate transactions equal the minimum of supply and demand. Such an assumption, which was very natural for our example with two agents, may not be valid if one considers a macroeconomic market, as not all demanders and suppliers meet pairwise. In particular it is well known that the property of market efficiency may be lost in the process of aggregating submarkets, whereas voluntary exchange remains. Note, however, that the concepts that follow do not require that property of market efficiency.

Now it is clear that the quantity signals received by the agents hould have an effect on demand, supply and price formation. This is what we shall explore now.

Effective Demand and Supply

Demands and supplies are signals that agents send to the 'market' (i.e. to the other agents) in order to obtain the best transactions according to their criterion. The traditional 'notional' or Walrasian demands and supplies are constructed under the assumption (which is actually verified ex-post in a Walrasian equilibrium) that each agent can buy and sell as much as he wants on each market. There is thus an equality between the signal the agent sends to the market (demand or supply) and the transaction he will obtained from it.

In disequilibrium analysis there is of course a difference between the signals sent (effective demands and supplies) and their consequences (the transactions actually realized). Effective demands and supplies expressed by an agent in the various markets are the signals which maximize his expected utility of the resulting transactions, knowing that these transactions are related to the demands and supplies by equalities of the type seen above, i.e.:

$$d_{ih}^{*} = \min\left(\tilde{d}_{ih}, \overline{d}_{ih}\right)$$
$$s_{ih}^{*} = \min\left(\tilde{s}_{ih}, \overline{s}_{ih}\right)$$

The results of such expected utility maximization programmes may be quite complex, depending for example on whether quantity constraints are expected deterministically or stochastically, or whether agents act or not as price markers, as we shall see in the next section. In the case of deterministic constraints, there exists a simple and workable definition of effective demand, which generalizes Clower's original 'dual decision' method: effective demand (or supply) on one particular market is the trade which maximizes the agent's criterion subject to the constraints encountered or expected on the other markets. This definition thus naturally integrates the well-known 'spillover effects', which show how disequilibrium in one market affects demands and supplies in the other markets.

We shall immediately give an illustrative example of this definition, due to Patinkin (1956) and Barro and Grossman (1971), that of the employment function of the firm. Consider a firm with a production function $y = F(l)$ exhibiting diminishing returns, and faced with a price p on the output market and a wage w on the labour market. The traditional 'notional' labour demand results from maximization of profit $py - wl$ subject to the production constraint $y = F(l)$, which yields immediately the usual Walrasian labour demand $F'^{-1}(w/p)$. Assume now that the firm faces a constraint $\overline{y}$ on its sales

of output (i.e. a total demand $\overline{y}$). According to the above definition the effective demand for labour $\tilde{l}$ is the solution in l of the following programme:

$$\begin{aligned} \text{Maximize} \quad & py - wl \quad \text{s.t.} \\ & y = F(l) \\ & y \leq \overline{y} \end{aligned}$$

the solution of which is:

$$\tilde{l} = \min\left\{F'^{-1}(w/p), \quad F^{-1}(\overline{y})\right\}$$

We see that the effective demand for labour may have two forms: the Walrasian demand just seen above if the sales constraint is not binding, or, if this constraint is binding, a more 'Keynesian' form equal to the quantity of labour just necessary to produce the output demand. We see immediately on this example that effective demand may have various functional forms, which intuitively explains why disequilibrium models often have multiple regimes (see for example the three goods–three regimes model in the entry FIX-PRICE MODELS.

In the case of stochastic demand, the programme yielding the effective demand for labour becomes evidently more complex. One obtains some results quite reminiscent of the inventories literature as developed for example by Arrow et al. (1958) or Bellman (1957). See Benassy (1982) for the link between these two lines of work.

Price Making

We shall now address the problem of price making by decentralized agents, and we shall see that there too quantity signals play a prominent role. It is actually quite intuitive that quantity signals must be a fundamental part of the competitive process in a truly decentralized economy. Indeed, it is the inability to sell as much as they want that leads suppliers to propose, or to accept from other agents, a lower price, and conversely it is the inability to buy as much as they want that leads demanders to propose, or accept, a higher price. Various modes of price making integrating these aspects can be envisioned. We shall deal here with a particular organization of the pricing process where agents on one side of the market (usually the suppliers) quote prices and agents on the other side act as price takers. Other modes of pricing (bargaining, contracting) are currently studied, but have not yet been integrated in this line of work. As we shall see this model of price making is quite reminiscent of the imperfect competition line: Chamberlain (1933), Robinson (1933), Triffin (1940), Bushaw and Clower (1957), Arrow (1959), Negishi (1961).

Consider thus, to fix ideas, the case where sellers set the prices (things would be quite symmetrical if demanders were setting the prices), and in order to have only one price per market, let us characterize a market by the nature of the good sold and its seller (we thus consider two goods sold by different sellers as different goods, a fairly usual assumption in microeconomic theory since these goods differ at least by location, quality, etc. . .). On each 'market' so defined we thus have one seller, the price maker, facing several buyers. As we saw above, for a given price this seller faces a quantity constraint s, actually equal to the demand of the other agents on that market. But the price level is now a decision variable for the seller, and this quantity constraint (the others' total demand) can be modified by changing the price: for example in general the seller who wants to sell more knows that, others things being equal, he should lower the price. The relation between the maximum quantity he expects to sell and the price set by the price maker is called the expected demand curve. If demand is forecasted deterministically, this expected demand curve will be denoted as:

$$\overline{S}(p, \theta)$$

where θ is a vector of parameters depending on the exact functional form of that curve (for example elasticity and a position factor for isoelastic curves). If demand is forecast stochastically, the expected demand curve will have the form of a probability distribution on $\overline{s}$ (i.e. total demand) conditional on the price.

For a given expected demand curve, the price maker chooses the price which will maximize profits, given the relation between price and

maximum sales. For example, continuing to consider a firm with production function $F(l)$, the programme yielding the optimum price is the following in the case of a deterministic expected demand curve:

$$\begin{aligned} &\text{Maximize} \quad py - wl \quad \text{s.t.} \\ &y = F(l) \\ &y \leq \overline{S}(p, \theta) \\ &l \leq \overline{l} \end{aligned}$$

where $\overline{l}$ is the constraint the firm possibly faces on the labour market, where it is a 'wage taker'. Note that, according to our definition above, the effective demand for labour of this same firm would be given by the above programme, from which the last constraint would be deleted.

Both the price and quantity decisions of price makers depend on the parameters θ. Of course it would require quite heroic assumptions on the computational ability and information available to price setters to assume that they know the 'true' demand function (i.e. the 'true' functional form with the 'true' parameters). But the theory developed here gives a natural way of learning about the demand curve. Indeed, each realization p, s in a period is a point on the 'true' demand curve in that period (Bushaw and Clower 1957). Using the sequence of these observations, plus any extra information available (including for example the price of its competitors), the price maker can use statistical techniques to yield an estimation of the demand curve. Whether this learning would lead to the 'true' demand curve is still an unresolved problem.

Expectations

Of course, the modifications we outlined concerning the signal structure affect not only the current period, but the future periods as well, and as compared to traditional 'competitive' analysis, disequilibrium analysis introduces expected quantity signals in addition to expected price signals. Such an introduction allows for example to rationalize the traditional Keynesian accelerator (Grossman 1972). The introduction of such quantity expectations into the microeconomic setting was made in Benassy (1975, 1977b, 1982). Macroeconomic applications of the corresponding concepts can be found in Hildenbrand and Hildenbrand (1978), Muellbauer and Portes (1978), Benassy (1982, 1986), Neary and Stiglitz (1983).

Scope and Uses of Disequilibrium Analysis

We have briefly outlined the basic elements or building blocks of disequilibrium analysis. We saw that it generalizes the traditional theories of demand, supply and price formation to cases where, in the absence of an auctioneer, markets do not automatically clear. This theory is thus a quite general one, and the scope of its applications very broad. Up to now there have been in the literature three particularly active areas of development: (1) The construction of various concepts of equilibria with rationing, or non-Walrasian equilibria. These concepts, which generalize the traditional notion of Walrasian equilibrium to the cases where not all markets clear, show how mixed price-quantity adjustments can bring about a new type of equilbrium in the short run. (2) The development of numerous macroeconomic applications, which basically use the above concepts in the framework of aggregated macromodels, and derive policy implications, for example to fight involuntary unemployment. (3) Finally new econometric methods have been developed to deal with such models, as traditional methods were more suited to the study of equilibrium markets.

Microeconomic concepts of non-Walrasian equilibria are reviewed in the entry RATIONED EQUILIBRIA. We shall now very briefly outline the macroeconomic and econometric developments.

Macroeconomic Applications

Many contributions in the field started from a reconsideration of Keynesian models, and it is therefore no surprise that many macroeconomic

applications have been made. The early model of Barro and Grossman (1971) has been followed by a huge macroeconomic literature, notably aimed at policy analysis and the study of involuntary unemployment. A very valuable feature of disequilibrium macromodels is that, like the microeconomic models, they endogenously generate multiple regimes in which various policy tools may have quite different impacts. These models are thus a particularly useful tool for synthesizing hitherto disjoint macroeconomic theories. One finds a number of macroeconomic applications in books by Barro and Grossman (1976), Benassy (1982, 1986), Cuddington et al. (1984), Malinvaud (1977), Negishi (1979). We may note that the same methods can also be used to study the problems of centrally planned economies (Portes 1981).

A few lessons can be drawn from these macrodisequilibrium models. The first is that, even though these models were at the very beginning aimed at bridging the gap with Keynesian analysis, they proved to be of more general relevance, and were able to generate non-Keynesian results as well as the traditional Keynesian results. Secondly, and more generally, whether or not a policy tool is efficient may depend very much on the 'regime' the economy is in. A famous example is the Barro and Grossman fixprice macroeconomic model, with its 'Classical unemployment' and 'Keynesian unemployment' regimes. Finally, it appears that the results of these models are quite sensitive to both the price formation mechanism on each market, as well as on the expectations formation mechanisms on both price and quantities (cf. for example, Benassy (1986), which experiments with various hypotheses).

This quite naturally leads to the need of further theoretical work, and to the necessity of empirically testing these models, an issue to which we now turn.

Disequilibrium Econometrics

In order to estimate microeconomic or macroeconomic disequilibrium models a whole new econometric technology has developed in recent years. Let us consider the very simplest case, that of a single market with a rigid price. The most basic system to estimate is then:

$$X^d = a_d Z_d + \varepsilon_d$$
$$X^s = a_s Z_s + \varepsilon_s$$
$$X = \min\left(X^d, X^s\right)$$

where X^d is quantity demanded, Z_d is the set of variables affecting demand, a_d is the vector of corresponding parameters and ε_d is a demand disturbance term (and symmetrically on the supply side). The market is assumed for the moment to function efficiently so that transaction X is the minimum of demand and supply X^d and X^s. The problem in estimating such a model, as compared with an equilibrium model, where by assumption

$$X = X^d = X^s$$

is that only X is observed, not X^d or X^s. Techniques for dealing with these problems are reviewed in Quandt (1982). Of course this is the simplest possible model, and numerous extensions are now considered: (1) The prices may be flexible, either within the period of estimation, or between successive periods. The price equation must then be estimated simultaneously with the demand–supply system. (2) Since some applications are made on macroeconomic markets, the 'minimum' condition may not be satisfied, and is replaced by an explicit procedure of aggregation of submarkets. (3) Finally models with several markets in disequilibrium have been estimated, notably at the macroeconomic level.

Concluding Remarks

The development of disequilibrium analysis has clearly led to an enlargement and synthesis of both traditional microeconomics and macroeconomics.

Usual microeconomic theory in the market clearing tradition has been generalized in a number of directions: the study of the functioning of non-clearing markets and the formation of quantity signals, a theory of demand and supply

responding to these quantity signals as well as to price signals, the integration of quantity expectations into microeconomic theory. This line of analysis further includes a theory of price making by agents internal to the system which also bridges the gap with the traditional theories of imperfect competition.

As for the corresponding macroeconomic models, they turn out to be a very useful synthetical tool, as they cover all possible disequilibrium configurations. They are more general than either traditional Keynesian macromodels, which considered only excess supply states, or than 'new classical' macromodels which postulate market clearing at all times. They are of course the natural tool to study problems such as involuntary unemployment.

Still richer developments lie ahead with further developments in the theories of price and wage formation in markets without an auctioneer. The methodology outlined here will permit us to derive the micro and macro consequences, as well as the consequences in terms of economic policy prescriptions. Much is also to be expected of the development of the associated econometric methods, which should allow us to choose the most relevant hypotheses, and to characterize specific historical episodes.

See Also

- ▶ Equilibrium: An Expectational Concept
- ▶ Rationed Equilibria
- ▶ Temporary Equilibrium
- ▶ Uncertainty and General Equilibrium

Bibliography

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

Arrow, K.J., S. Karlin, and H. Scarf. 1958. *Studies in the mathematical theory of inventory and production*. Stanford: Stanford University Press.

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

Barro, R.J., and H.I. Grossman. 1976. *Money, employment and inflation*. Cambridge: Cambridge University Press.

Bellman, R. 1957. *Dynamic programming*. Princeton: Princeton University Press.

Benassy, J.P. 1975. Neo-Keynesian disequilibrium theory in a monetary economy. *Review of Economic Studies* 42: 502–523.

Benassy, J.P. 1976. The disequilibrium approach to monopolistic price setting and general monopolistic equilibrium. *Review of Economic Studies* 43: 69–81.

Benassy, J.P. 1977a. A Neo–Keynesian model of price and quantity determination in disequilibrium. In *Equilibrium and disequilibrium in economic theory*, ed. G. Schwodiauer. Boston: D. Reidel Publishing Company.

Benassy, J.P. 1977b. On quantity signals and the foundations of effective demand theory. *Scandinavian Journal of Economics* 79: 147–168.

Benassy, J.P. 1982. *The economics of market disequilibrium*. New York: Academic Press.

Benassy, J.P. 1986. *Macroeconomics: An introduction to the Non-Walrasian approach*. New York: Academic Press.

Bushaw, D.W., and R. Clower. 1957. *Introduction to mathematical economics*. Homewood: Richard D. Irwin.

Chamberlin, E.H. 1933. *The theory of monopolistic competition*. Cambridge, MA: Harvard University Press.

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

Cuddington, J.T., P.O. Johansson, and K.G. Lofgren. 1984. *Disequilibrium macroeconomics in open economies*. Oxford: Basil Blackwell.

Drèze, J.H. 1975. Existence of an exchange equilibrium under price rigidities. *International Economic Review* 16: 301–320.

Grossman, H.I. 1972. A choice-theoretic model of an income investment accelerator. *American Economic Review* 62: 630–641.

Hahn, F.H., and T. Negishi. 1962. A theorem on non tatonnement stability. *Econometrica* 30: 463–469.

Hansen, B. 1951. *A study in the theory of inflation*. London: Allen & Unwin.

Hicks, J.R. 1965. *Capital and growth*. London: Oxford University Press.

Hildenbrand, K., and W. Hildenbrand. 1978. On Keynesian equilibria with unemployment and quantity rationing. *Journal of Economic Theory* 18: 255–277.

Keynes, J.M. 1936. *The general theory of money, interest and employment*. New York: Harcourt Brace.

Keynes, J.M. 1937. Alternative theories of the rate of interest. *Economic Journal* 47: 241–252.

Leijonhufvud, A. 1968. *On Keynesian economics and the economics of Keynes*. Oxford: Oxford University Press.

Machlup, F. 1958. Equilibrium and disequilibrium: Misplaced concreteness and disguised politics. *Economic Journal* 68: 1–24.

Malinvaud, E. 1977. *The theory of unemployment reconsidered*. Oxford: Basil Blackwell.

Muellbauer, J., and R. Portes. 1978. Macroeconomic models with quantity rationing. *Economic Journal* 88: 788–821.
Neary, J.P., and J.E. Stiglitz. 1983. Towards a reconstruction of Keynesian economics: Expectations and constrained equilibria. *Quarterly Journal of Economics* 98(Supplement): 199–228.
Negishi, T. 1961. Monopolistic competition and general equilibrium. *Review of Economic Studies* 28: 196–201.
Negishi, T. 1979. *Microeconomic foundations of Keynesian macroeconomics*. Amsterdam: North-Holland.
Patinkin, D. 1956. *Money, interest and prices*. New York: Row, Peterson & Co.; 2nd ed. New York: Harper & Row, 1965.
Portes, R. 1981. Macroeconomic equilibrium and disequilibrium in centrally planned economies. *Economic Inquiry* 19: 559–578.
Quandt, R.E. 1982. Econometric disequilibrium models. *Econometric Review* 1: 1–63.
Robinson, J. 1933. *The economics of imperfect competition*. London: Macmillan.
Samuelson, P.A. 1947. *Foundations of economic analysis*. Cambridge, MA: Harvard University Press.
Tobin, J., and H.S. Houthakker. 1950. The effects of rationing on demand elasticities. *Review of Economic Studies* 18: 140–153.
Triffin, R. 1940. *Monopolistic competition and general equilibrium theory*. Cambridge, MA: Harvard University Press.
Younès, Y. 1975. On the role of money in the process of exchange and the existence of a non-Walrasian equilibrium. *Review of Economic Studies* 42: 489–501.

Disguised Unemployment

Amit Bhaduri

Marx set out the notion that a 'reserve army' of unemployed labour is more or less continuously maintained in the course of capitalistic development. In the initial phases, this reserve army may be created through the destruction of the pre-capitalistic modes of production while, in later phases, a systematic bias in favour of labour-displacing innovations could serve the same purpose. This entails a broad vision of capitalistic development under extremely elastic supply conditions for labour where the actual level of wage employment is usually demand-determined. This means that the supply of labour tends to adjust to its demand through various routes such as, higher participation rate (e.g. as more married women join the labour force or the average schooling period is shortened), interregional and international migration of labour etc., all this taking place against the background of continuous induced innovations. Under these circumstances, it is not very useful to think of a 'natural' rate of growth, set by the growth of labour force and of labour productivity, as the maximum feasible growth rate of a capitalist economy (Marglin 1984 pp. 103–8).

The elastic nature of the labour supply and its adjustability to the level of demand entail the existence of open or disguised unemployment as an untapped reservoir of labour in the normal course of capitalist development. However, such disguised unemployment although real is a somewhat amorphous phenomenon in an advanced capitalist economy for two distinct reasons. First, under normal circumstances, many potential entrants (e.g. married women, late school-leavers, young people on the farm) may not actually even try to enter the labour market unless demand is seen to be high with all sorts of job vacancies exceeding their corresponding numbers in registered unemployment. Second, the economic nationalism in the richer countries often takes the form of strictly regulating the migration of 'guest-workers' so that open unemployment in the (potentially) labour-exporting countries, rather than disguised unemployment in the advanced capitalist countries, becomes the normal pattern. And yet, prolonged stagnation in economic conditions in an advanced capitalist country may make this phenomenon of disguised unemployment more visible, as the redundant workers either seek various forms of self-employment with virtually no invested capital or try to sell their labour services directly as porters, odd-jobmen, domestic servants, farm-hands etc. (Robinson 1956, pp. 157–8). Their earnings in these peripheral jobs would then become the 'reservation price' of this marginalized labour force. When unemployment dole and social security set a higher reservation price, some of this unemployment may come out in the open instead of being disguised. In this sense, it is probable that the

growth of the welfare state may openly register as unemployed some who would have been otherwise unemployed in a disguised fashion earlier. And, the reverse could happen if the social security measures are cut by the government.

The existence of such disguised unemployment on a significant scale is usually accommodated by a secondary or informal labour market mostly in the service sector. This is much more easily visible in the phenomenon of massive migration to urban centres from rural areas in many developing countries. While all such migrants from rural areas aspire to limited job opportunities in organized industries located in urban areas, only a small fraction among them are actually able to find proper jobs at any given point of time. The rest spend their time, waiting in search of appropriate jobs. In the meantime, they somehow manage to disguise their unemployment either by self-employing themselves with tiny amounts of invested capital (e.g. polishing shoes, cleaning cars etc.) or by selling their labour services directly in odd jobs or even, simply taking recourse to the support of the elaborate kinship system in more traditional societies e.g. by living off better-placed relatives and migrant workers from their home areas. Thus, the phenomenon of disguised unemployment in the urban areas of many developing countries becomes closely linked with the massive migration from rural areas during the course of industrialization.

A distinguishing feature of disguised unemployment in such an informal sector is the irregular and often long hours of work per day. This is evident enough in the case of most self-employed persons in the informal sector; but even those who are employed on a wage-labour basis usually have highly flexible wage contracts in many respects (e.g. domestic servants, odd-jobmen etc.). Partly the explanation lies in the lower unionization of this sector. However, a deeper explanation lies in the fact that most self-employed persons as well as workers paid at the piece-rate have to work extended hours per day simply to make a livelihood. But this also could have a limited advantage for some of them insofar as the entire family can participate in the work (e.g. traditional carpet making, weaving and other types of artisan work are often done by many members of the family working together). In this context, we have to make a sharp distinction between labour-service and the labourer providing such service: the same amount of labour service (say, 18 hours per day) may be spread out over several family members working as labourers (say, three). In some cases, each family member (labourer) may on an average have a lighter work load (of only six hours) per day compared to an average worker in the organized industry. This brings us to a somewhat different analytical dimension of disguised unemployment: some persons may be unemployed in a disguised manner not only in the sense of having a very low earning rate i.e. *income-wise* unemployment but also in the sense of relatively light work-intensity per day, i.e. *time-disposition-wise* unemployment. And, unless one believes in the neoclassical proposition that income necessarily reflects the marginal product, one would have to devise, a third (and separate) criterion of disguised unemployment in terms of abnormally low *productivity* of labour. However, given the structure of reward in a capitalist economy, one needs to be careful in applying these concepts. Thus, an 'important person' belonging to the board of directors of several large corporations, may be making a well-above-average income by attending only a couple of board meetings per month. Such a person may very well be considered to be disguised unemployed by the time-disposition criterion and even perhaps by the labour-productivity criterion although, he cannot, by any means, be considered unemployed, disguised or not, by the income criterion! Also recall in this context that 'unproductive labour' was a common category used in the classical tradition of political economy and, all those engaged in unproductive labour (e.g. 'priests, prostitutes and professors' according to a picturesque phrase employed by Rosa Luxemburg) could be considered to be disguised unemployed by the productivity criterion.

In the normal organization of factory work under the capitalist system, the threefold distinction between income-wise, time-wise and productivity-wide disguised unemployment may not be particularly relevant. Thus, an unemployed

industrial worker is both income– and time-wise unemployed and of course, he does not have much a chance to be productive either. However, such a distinction can be highly relevant in the context of traditional, family-based agriculture, especially for characterizing such phenomena as rural poverty or the existence of surplus labour. Consider for example a typical rural woman in the poorest strata: in addition to all her other work inside and outside the house, she may have to spend long hours collecting wood for fuel and carrying water home from a distance. Although she has exceptionally hard and long working hours every day and must be considered time – and disposition-wise fully employed and certainly productive in every normal sense of the term, in keeping her family going under most difficult circumstances, in all probability she would *not* be classified as 'gainfully employed' by the income criterion. Indeed her case is the opposite of that our 'important person' who has a high income by attending a couple of board meetings every month. It is to be noted that the worst kind of rural poverty is often concentrated among people who are fully employed by the time-disposition criterion, but may be described as disguised unemployed by the income criterion, because of their miserably low earning rate per hour of work. After all, this is what the phrase 'eking out a living' usually means.

There can hardly be any serious doubt that in the backward agriculture of many populous countries (e.g. in South Asia), a high proportion of the population engaged in cultivation have extremely low income and, in this sense suffers from disguised unemployment by the income-criterion. Nevertheless, it is far more problematic to identify what such disguised unemployment by the income criterion implies in terms of either the time-disposal or the productivity criterion. If one were to believe in the ideologically potent neoclassical slogan that all 'factors of production' including labour always tend to get paid according to their marginal product even in pre-capitalist, backward agriculture, then that proportion of population with extremely low income could be said to be rather unproductively engaged in agriculture. Their low income would be the 'evidence' of their low productivity which in turn would imply a corresponding level of disguised unemployment in agriculture. But this would involve implicit theorizing based on the dubious assumption that income (earning) is always positively associated with productivity, even in traditional agriculture.

Such implicit theorizing apart (a sophisticated example of which is the so-called 'efficiency wage' hypothesis e.g. Bliss and Stern 1978) the important question remains as to whether there is any meaningful sense in which one can argue about the existence of significant surplus labour and disguised unemployment in backward agriculture, judged by the productivity criterion. This would imply that some surplus labour can be withdrawn from agriculture without adversely affecting the level of agricultural output. Or, in more textbookish jargon, 'at the margin' labour contributes nothing to output so that, the marginal product of labour is zero in such agriculture. Put in such general terms, the formulation is too fuzzy to be useful. For instance, if by 'margin', one means the *intensive* margin of higher labour input per unit of land, then considerable empirical evidence exists, at least in India, to suggest that the smaller-sized land holdings usually do use family labour more intensively, both in current agricultural operations *and* in direct investment of labour for improving land quality. As a result, the total output, taking all crops together over the year, tends to be higher per unit of land on smaller holdings (Bharadwaj 1974, chs. 2, 3 and 7 provide an excellent account). This tendency towards an *inverse* relation between farm size and productivity per acre in traditional agriculture would tend to cast doubt on the simple-minded proposition that the 'marginal' product of labour is zero, especially if the notion of intensive margin is used.

Without going into such finer points of intensive and extensive margin, Schultz (1964, ch. 4) proposed the 'epidemic test': the 1918–19 influenza epidemic in India killed 6.2% of the 1918 population and 8.3% of the working population in agriculture (the latter according to Schultz's estimate). Schultz found that, although the weather conditions were roughly similar in 1916–17 and in 1919–20, in the latter year agricultural output

was lower by about 3.8%, providing circumstantial evidence that withdrawal of labour from agriculture did affect output level. However, apart from many statistical and conceptual problems (e.g. the relation between acrage change in the sense of extensive margin and output change which is a resultant of both extensive and intensive margin in his macro-level statistical investigation), this 'epidemic test' must be deemed to be over-simplistic despite its apparent ingenuity. At best, it showed that a *random x*% withdrawal of labour from cultivation did affect the acrage and/or output level. But it does in no way establish the impossibility of *selectively* withdrawing *x*% labour through suitable reorganization of agricultural production at the family and regional level (e.g. Sen 1967). And yet, most of the important initial proponents of the 'surplus labour' doctrine had in mind such selective (but not random) withdrawal of labour that may be induced by industrialization and expansion in urban employment opportunities (Nurkse 1953; Lewis 1954). And, once it is recognized that such withdrawal of labour from agriculture can be accompanied by reorganization of labour in the family farm through adjusting the hours of work of the family members staying back on the farm or through higher availability of land per cultivating family, it seems plausible to argue analytically (e.g. Takagi 1978) as well as empirically that, labour can usually be released from agriculture without adversely affecting the level of agricultural output. Indeed, post-revolutionary experiences of agrarian reorganization in China and Vietnam demonstrated the possibility of using surplus labour to improve the quality of land through better drainage and irrigation without significant drop in short-run agricultural output, despite all the serious problems of lack of adequate incentive to private production in cooperative and collective agriculture.

See Also

- ▶ Harris–Todaro Hypothesis
- ▶ Labour Surplus Economies
- ▶ Robinson, Joan Violet (1903–1983)

References

Bharadwaj, K. 1974. *Production conditions in Indian agriculture* (A study based on farm management surveys). Occasional Paper 33, Department of Applied Economics. Cambridge: Cambridge University Press.

Bliss, C., and N. Stern. 1978. Productivity, wages and nutrition. Parts I and II. *Journal of Development Economics* 5(4): 331–398.

Lewis, W.A. 1954. Economic development with unlimited supplies of labour. *Manchester School of Economic and Social Studies* 22: 139–191.

Marglin, S.A. 1984. *Growth, distribution, and prices*. Cambridge, MA: Harvard University Press.

Nurkse, R. 1953. *Problems of capital formation in underdeveloped economies*. Oxford: Clarendon Press.

Robinson, J. 1956. *The accumulation of capital*. London: Macmillan.

Schultz, T.W. 1964. *Transforming traditional agriculture*. New Haven: Yale University Press.

Sen, A.K. 1967. Surplus labour in India: A critique of Schultz's statistical test. *Economic Journal* 77: 154–160.

Takagi, Y. 1978. Surplus labour and disguised unemployment. *Oxford Economic Papers* 30(3): 447–457.

Disintermediation

Charles Goodhart

'Intermediation' generally refers to the interposition of a financial institution in the process of transferring funds between ultimate savers and ultimate borrowers. The forms of services that such financial intermediaries provide, the characteristics of their liabilities and assets, and the rationale for their existence is described elsewhere. For this purpose, we only need to assume that a certain pattern of financial intermediation is given, say by actual historical development, or is theoretically optimal.

Disintermediation is then said to occur when some intervention, usually by government agencies for the purpose of controlling, or regulating, the growth of financial intermediaries, lessens their advantages in the provision of financial services, and drives financial transfers and business into other channels. In some cases the transfers of funds that otherwise would have gone through the

books of financial intermediaries now pass directly from saver to borrower. An example of this is to be found when onerous reserve requirements on banks leads them to raise the margin (the spread) between deposit and lending rates, in order to maintain their profitability, so much that the more credit-worthy borrowers are induced to raise short-term funds directly from savers, for example, in the commercial paper market. Another, more recent, example arises when stringent capital adequacy requirements lead banks to provide funds to borrowers in a form that can be packaged into securities of a kind that can be on-sold to ultimate savers, rather than kept on the books of the banks involved, and thereby need larger capital backing.

Disintermediation not only refers to those instances where financial flows are constrained by intervention to pass more directly from saver to borrower (than in an unconstrained context), but also where such flows pass through different, and generally less efficient, channels than would otherwise be the case. This latter is just as common in practice. For example, where constraints and regulations are imposed on some sub-set of domestic financial institutions, substitute services of a similar kind will become provided by 'fringe' financial institutions that are not so constrained. More generally, in the absence of exchange control, constraints and burdens on the provision of domestic financial services will encourage financial institutions to provide these same services abroad, notably in the international Euro-markets. Indeed, the development of the Euro-markets provides a case study of the power of disintermediation out of more rigorously controlled domestic financial markets into an international milieu not subject to such controls. The likelihood of such disintermediation imposes a limit on the authorities' ability to impose controls and regulations on financial intermediaries. If such controls are to be effective, they presumably force financial intermediaries to behave in a way that they would not voluntarily do, and hence represent a burden on them. There will then be an incentive for the controlled financial intermediary to seek to escape such a burden, for example through disintermediation. This represents a perennial problem for the monetary authorities. Logically, it might seem to lead to a tendency for the authorities to be forced to extremes, either to prevent disintermediation altogether by extending the ambit of controls to all forms and kinds of financial intermediation, or alternatively to allow complete laissez-faire within the financial system, despite the dangers of financial instability that might ensue. In practice, however, the authorities try to seek a compromise in the form of regulations sufficiently well-designed to maintain monetary control and financial stability, without being sufficiently burdensome to cause large-scale disintermediation. This is not, however, an easy exercise and requires continuous adjustment by the authorities as the financial system evolves.

See Also

- ▶ Financial Intermediaries
- ▶ Monetary Policy

Displacement in the Middle East: Where the Past is Prologue?

Dawn Chatty and Maira Seeley

Abstract

The Middle East is now the major refugee-producing region of the world, as well as the major hosting region, with nearly 63% of the world's refugees (UNHCR statistics 2014). These major population flows and changes going back to the middle of the 19th century have had significant impacts on the development of the region throughout the 20th century and into the 21st.

Keywords

Community; Development; Diaspora; Displacement; Homeland; Identity; Immigration; Iraq; Jordan; Lebanon; Middle East;

Migration; Nationalism; Refugee; Regeneration; Ottoman; Palestine; Syria; Turkey

JEL Classifications
J15; J61; N34

Displacement at the End of the Ottoman Empire

The Middle East has been the focus of centuries, if not millennia, of movements of people. For much of the past 500 years the largely involuntary movement of people was supported by a system of government which encouraged and tolerated variations among people, drawing out differences between neighbours and encouraging the formation of unique identities based on cultural, linguistic or religious grounds. In this heartland of the Ottoman Empire, belonging was not based on physical birthplace alone, but specifically included the social community of origin (Humphreys 1999; Kedourie 1984). Immigrants – forced and voluntary – were readily accepted into the fabric of this multicultural empire and institutionalised mechanisms were set up by the state to assist in their integration (Chatty 2010). The Ottoman Empire upon which such identities were based came to an end with the First World War.

Amid the rubble at the end of the Great War was a startling range of social movement. This included social groups on the Russian-Ottoman border lands, such as the Armenian, Circassian and other Northern Caucasus peoples (Barkey and Von Hagen 1997; Brubaker 1995). Other dispossessions had their origins in the lines drawn on maps by the great Western powers (e.g. the Sykes–Picot agreement) to create new nation states (Bocco et al. 1993; Gelvin 1998; Helms 1981; Morris 1987; Wilkinson 1983). These include the Palestinians, the Kurds, the pastoral Bedouin and a variety of 'stateless peoples'. In some cases, such as those of the Yazidis, the Assyrians and some Armenian groups, migration was closely linked to regional efforts to create a pan-Arab, socialist or Islamic society (Al-Rasheed 1994; Khalidi 1997; Lerner et al. 1958). These refugees, exiles and 'exchangees' found new homes and created new communities without much attention or assistance from the new international order. They established themselves in new soil, but managed their memories so as not to lay down new roots, but rather to strengthen the commonality and trust in their immediate social network. They were creating moral and economic communities with social capital that oiled internal social cohesion (Chatty 2010).

These forced migrant movements also had profound impacts on the economic development of the region in the early 20th century. Armenians fleeing Turkish persecution began to arrive in the Arab world in the early 1900s, moving into Palestine and Transjordan in the early 1910s (Al-Khatib 2000). While economic data from this early period is scarce, these Armenian refugees contributed expertise in trades such as shoemaking, machine mechanics and still photography, as well as financial expertise, supporting the development of local economies (Becker and El-Said 2013). Circassian and Chechen refugees arriving in Palestine, Syria and Jordan during the same period expanded agricultural production, and much of Jordan's developing market economy was initially structured along ethnic lines, with displaced ethnic groups playing a critical role (Becker and El-Said 2013). Refugees from all three ethnic groups also played crucial roles in Jordan's early urban development, particularly in and around Amman.

The transnational links created by forced migration in the Middle East also contributed to a continued re-imagination of geographic and social space in response to population movements. In this region the mass movement of people over the past 150 years makes the attempt to regard the area as a set of homelands or cultural regions bewildering, to say the least. The Assyrian Arabs, once largely found in pre-and post-colonial Iraq, have reappeared in Chicago; the Circassians have centred their diasporic headquarters in New Jersey; and Iraqi refugees and exiles have found new community nodes in London and other major western cities. Remittance links between these new communities and the Middle East also contributed significantly to the economic

development of the region; charitable organisations formed by, for example, the Iraqi diaspora in the USA, including Chaldeans, previously provided critical financial support for Iraqis seeking refuge in Syria and Jordan, as well as education and medical aid (Blayney 2011). The 'here' and the 'there' have become blurred in such translocal or diasporic situations and the cultural certainty of the 'centre' becomes equally unclear. Thus the experience of displacement was not restricted to those who have moved to the periphery, but also affected those in the core.

It is clear that nationalism played an important role in the politics of 'place-making' out of territorial spaces. Thus, the creation of natural links between places and people lay largely with the dominating cultural group which controlled the state. However, contestation or opposition to these natural links was common among the dispossessed, as evidenced by the emergence of ethnic 'counternations' such as the Circassians, the Palestinians, the Kurds and the Armenians. Palestinians, for example, expressed a deeply felt relationship to the 'villages of origin' and the 'land' in general.

In many of the states of the region, Syria being just one case in point, the sense of national unity was created through the struggle for independence (Brandell and Rabo 2003). Beginning in 1920, with the awarding of the League of Nations mandate to the French administration, the territory was divided into a number of states. Through common cause and hostility, the population of the territories rebelled and continue to fight the French policy of 'dividing and ruling' Syria as six separate statelets. After more than a decade of insurrection and conflict the French government agreed to reunite the territory administratively into a single state. The exceptions were the areas that had been attached to Mount Lebanon, to create the new state of Greater Lebanon and the Sanjak of Alexandretta.

Displacement and the Creation of New States

The close link between culture, national identity and territory, which has been so characteristic of European nation states, does not translate as easily to the contemporary states which make up the modern Middle East. In the new states which emerged after the demise of the Ottoman Empire, the violent displacement of people, often through compulsory exchange, was generally accompanied by a variety of state and international assistance, which included the granting of citizenship and housing aid, the provision of land, and sometimes financial packages as well as employment. Thus, for example, Asia Minor Greeks were taken and given space to live by the Greek state. The League of Nations' Refugee Settlement Commission (the effective predecessor of the United Nations High Commission for Refugees – UNHCR), financed such resettlement by high-interest international loans, assisted with land allocations and agricultural start-up packages. Between 1923 and 1930, it set up some 2000 villages which were created at the Greek state's direction in the newly conquered zones from which Muslims had been forced to leave 'voluntarily' (Hirschon 1998; Loizos 1999).

Most of the dispossessed, uprooted and deported, who struggled to build new lives and re-create communities in the early 1920s, however, were not provided with much national or international assistance. They were often left to their own devices to survive and reconstruct their networks and communities. Not having international support was balanced, however, by being in the midst of supportive social environments made up of discrete communities of people who had migrated into the region decades earlier and who shared common beliefs about their identities based on ideas of religion and, also, ethnicity (Barth 1969; Eriksen 1993). Thus when the Muslims from Crete arrived on the Levantine and Asia Minor coasts, they expected to be resettled on abandoned properties. On their arrival, however, they often found that the formerly Greek Orthodox-owned lands and houses which should have been available to them had been appropriated by local people or government officials (Loizos 1999, p. 245). But their more widely flung social networks along both coasts meant that they were able to tap into a supportive environment made up of similarly discrete

communities who had arrived and settled in these territories a century earlier.

They 'healed' each other and built new communities based on trust, exchange and mutuality. They, too, consciously retained a separate identity from the rest of their surroundings and thus actively sought to mark themselves out as an unassimilated minority. The Cretans Muslims in Turkey re-created their past by retaining certain selected key elements of their culture while other parts diminished in importance (cf. Hirschon 1998).

The Muslim Circassians and related peoples, the Armenians and Eastern Christian peoples, the Palestinians and the Kurds represented a significant range of the ethno-religious communities that were dispossessed, uprooted and eventually, largely through their own efforts, re-established in the Arab Middle East. These four communities are elaborated in Chatty (2010), bringing the voice of the forced migration to the fore. This work uniquely extends our understanding of displacement and dispossession in the modern Middle East beyond the Palestinian case, which has rightly dominated contemporary scholarship. The work pulls in the nearly 3 million 'others' (mainly Muslim and Christian forced migrants of the 19th and 20th centuries) and sets the Circassians and Chechnyans, the Armenians and other Christian groups as well as the Kurds on a level playing field with Palestinian refugees.

These four cases give us a deeper and more complex understanding of the meanings of home and homeland, myth and myth-making, community regeneration, economic development and resilience, as well as the local rejection of diaspora and transnationalism to reaffirm the process of social and economic integration without cultural assimilation in new physical spaces (Chatty 2010). It addresses the unique roles that each of these forced migrant communities have played in the modern development of, for example, commercial monopolies in particular trades (jewellery, tailoring, textiles and photography), as well as in the security services (Circassian and Armenians in Syrian gendarmerie and Circassians and Chechnyans in the Royal Household protocol and security services of Jordan).

The contributions of these displaced ethno-religious communities to the development of their host countries have been significant. As noted above, the early 20th century influxes of Circassians, Chechens and Armenians were instrumental in economic and urban development in the Levant, especially in previously isolated Transjordan. The displacement of Palestinians in 1948 and after also impacted the region in several critical ways. The majority of Palestinians seeking refuge during and after the 1948 war settled in Lebanon, Syria, Transjordan, Egypt and Iraq, and the impact of Palestinians on the development of Amman, Jordan's capital, provides insight into the effects of Palestinian displacement. By 1950, 500,000 Palestinian refugees were in Jordanian-controlled territory, and additional refugees fled to Amman in the 1950s (Hanania 2014). These mass arrivals prompted urban development in the capital and the construction of new roads and municipal infrastructure. Palestinian refugees also contributed significantly to the formation of a new middle class of professionals and skilled workers in Amman (Hanania 2014). Many refugees brought higher levels of education and technical skills that fuelled economic and social development in the capital, and the arrival of wealthy Palestinian families drove demand for imported and manufactured goods, as well as investment (Hanania 2014). Additional refugees arrived during and after the 1967 war, contributing further to Amman's expansion. Remittances from Palestinians employed in the Gulf in later decades were also an important factor in the development of Jordan's public and private sectors, as a source of start-up funding (Chatelard 2010). After the 1990 Gulf War, most of the educated professional Palestinians who had helped to build the economy of Kuwait in the 1960s, 70s and 80s were expelled and 'returned' to Jordan (many had grown up in Kuwait and had never visited Jordan before). They invested their savings largely into property, fuelling a construction spurt that witnessed an exponential period of growth in Amman in the decade that followed.

However, the influx of refugees (many of whom had little education or formal technical training) also led to a sharp rise in unemployment

in Jordan, with the majority of Palestinians living in refugee camps and reliant on humanitarian aid, as well as heavy pressure on Amman's existing water resources (Hanania 2014). It is also critical to note that the development impact of Palestinian refugees differed in other host countries, such as Lebanon (Chaaban et al. 2010).

21st Century Displacement Within the Region

The 21st century has seen a new wave of dispossession in the Middle East, so large and so sudden as to threaten the economic and political stability of many countries in the region. Lebanon, Iraq, Syria and Jordan have seen waves of dispossessed and displaced enter their countries over the past five years, which have dwarfed the refugee loads experienced in Europe in previous decades.

Critics of the 2003 Iraq war have tended to focus on the cost in money and lives rather than on the catastrophic consequences for Iraq. One consequence in particular deserves more attention than it has received: the plight of Iraq's 4 million refugees, most of whom have remained in the region. Crucially, Iraqis' recent refuge in the neighbouring countries of Syria, Jordan and Lebanon rapidly became a protracted crisis, notwithstanding the tolerance of their hosts. Unwilling to return and largely unable to emigrate further west or north, Iraq's refugees in the Middle East remain in a perilous situation.

In the aftermath of the invasion of Iraq in March 2003, the western powers prepared for 1 million Iraqi 'refugees' to flee their country. Camps were duly set up to receive those who might try to escape the conflict. However, six months after the fall of the Iraqi regime, few Iraqis actually had fled their country. The international aid regime had miscalculated the Iraqi people's response to the invasion; the empty emergency camps were dismantled and pre-positioned food and equipment were removed.

Three years later, in 2006, the West was caught off-guard as hundreds of thousands of Iraqis fled their homes to escape the deadly sectarian violence which had escalated with the al-Askari mosque bombing in Samarra in the February of that year. That single event became the iconic image of sectarian violence and the 'unmixing' of people which followed. Nearly 4 million Iraqis fled their homes in 2006 and 2007, with 1–1.5 million crossing national borders into Syria and Jordan. The UNHCR and affiliated NGOs raced to set up reception centres and to provide emergency aid. In both Syria and Jordan, Iraqis were not regarded legally as refugees by the host governments, partially because neither country was a signatory to the 1951 United Nations Convention relating to the Status of Refugees.

Many of the Iraqis seeking asylum were from the educated, professional middle class. The extent of the extraordinary 'brain drain' from Iraq during this period has been well documented (Sassoon 2009). A number managed to escape with savings, which helped to ease their transition in exile. Migrations during previous decades meant that some Iraqi social networks were already in place in the host countries. The residual cultural memory of the '*millet*' system of the Ottoman Empire, which gave minority/religious communities a limited amount of power to regulate their own affairs, meant that Iraqi arrivals in these cities were generally tolerated, if not actively comforted. Also, memory of the Pan-Arab aspirations in the region meant that Iraqis were seen as temporary guests and 'Arab brothers'.

The Iraqi displacement crisis has reached a critical stage. International interest in Iraq is declining. Yet the lack of security, continuing civil conflict and economic uncertainty makes it unlikely that a mass Iraqi return will occur. More likely, Iraqi refugees will remain in neighbouring states under increasingly difficult circumstances. As their savings diminish and their circular movements into and out of Iraq to make money become more precarious, it is likely that irregular and long-distance migrations will occur in larger numbers.

Iraqi displacement has had a number of effects on host countries' development. As noted above, movements in and out of Iraq have both enabled some of the displaced to continue to transfer income generated within Iraq to neighbouring

countries, while family members are based in Syria, Lebanon and Jordan (Crisp et al. 2009). Many Iraqi refugees in these countries have come from middle class origins, with some bringing considerable investment to their countries of refuge, fuelling economic development. Most of the 25,000 Iraqis who have gained residence rights in Jordan since the beginning of the Iraqi conflict in 2003 have done so through investment (Chatelard 2010). In Jordan, a majority of displaced Iraqis surveyed by UNHCR in 2009 were professionals and 35% held a university degree (Crisp et al. 2009).

However, very few Iraqis in the three host countries are able to access employment, and many suffer from high levels of need. As of 2008, 42% of Iraqis in Jordan depended on remittances from Iraq for survival, while 24% of Iraqis in Syria depended on remittances from abroad (Harper 2008). In Syria, the arrival of refugees caused a dramatic increase in the price of basic necessities and rents, as well as strain on public services in education and healthcare, lowering the quality of services provided (Al-Miqdad 2007).

The speed with which Syria disintegrated into violent armed conflict after 2011 shocked the world; it has also left the humanitarian aid regime in turmoil as agencies struggled to react effectively to the massive displacement crisis. By 2015 there were more than 5 million Syrians seeking refuge in neighbouring countries, as well as another 6 million internally displaced in the country. The international aid regime has attempted to provide assistance to refugees who register with the UN if they are deemed needy. Perhaps only 5–10% of these millions have been able to access food vouchers and basic survival kits. For the most part, refugees in the neighbouring countries are not permitted to work, making their reliance on the UN particularly significant.

Each country bordering on Syria has responded differently to this complex emergency: Turkey rushed to set up its own refugee camps for the most vulnerable groups, but generally permitted self-settlement; Lebanon refused to allow the international humanitarian aid regime to set up formal refugee camps; and Jordan facilitated the creation of a UN refugee camp near its border with Syria. Turkey, Lebanon and Jordan have all granted refugees "guest" status under domestic legislation. And although Turkey has signed the 1951 Convention, it has reserved its interpretation of the Convention to apply only to Europeans seeking refuge/asylum in Turkey. UN estimates are that over 80% of the Syrian refugee flow across international borders is self-settling in cities, towns and villages where they have social networks. Despite a general rejection of encampment among those fleeing, still some 15–20% of the Syrian refugee population is in camps.

Each of these states has established a variety of temporary measures to deal with the crisis, which has reached proportions far outstripping the displacement crisis at the end of the Second World War. With refugees from Syria now estimated at nearly 3 million in Turkey and 1.5 million in Lebanon and officially nearly 700,000 in Jordan, as well as nearly 1 million now in Europe, the crisis threatens not only the security of the hosting nations in the region but also the unity of the European Union. Syrians who have sought refuge in the neighboring states are largely not permitted to work (although Jordan has recently sought to increase the accessibility of work permits for Syrian refugees). Many are unable to access adequate education, food and healthcare for their families.

The Syrian refugee crisis has also had wide-ranging effects on the socioeconomic development of neighbouring countries in the region, who have received the vast majority of refugees from the conflict. In Lebanon, increasing demand for public services has both strained public finances and lowered the quality of services provided (World Bank 2013). The Lebanese Ministry of Health and Social Affairs reported a 40% increase in use of its health and social programmes (World Bank 2013). This strain on social safety nets and services is ultimately pushing tens of thousands of Lebanese below the poverty line and may be increasing the youth unemployment rate in a country already experiencing significant poverty (World Bank 2013). It is important to note, however, that high unemployment has long been present in the Middle East and that current unemployment may not have been caused directly

by locals' competition with Syrian refugees (IRC 2016).

Overall, competition from Syrian refugees has driven down wages in the informal sector and negatively affected working conditions in host countries (IRC 2016). In Jordan, overall unemployment has risen from 14.2% to 22.1% since the beginning of the refugee crisis, with more significant increases among youth and less educated sectors of the population, and there is some evidence that Syrians may have replaced Jordanians in specific sectors, such as construction, wholesale and retail (Stave and Hillesund 2015). Child labour is more prevalent among refugee families, whose children have lower school enrolment rates than Jordanian children: while 95% of Jordanian children are still in school at 17, less than 40% of Syrian children are still in school at age of 15 (Stave and Hillesund 2015). As in Lebanon, the refugee influx in Jordan has strained social and public services, particularly in education, deteriorating the quality of services provided (REACH 2014).

The presence of Syrian refugees has, however, brought benefits for host countries that may have a positive impact on their long-term development. In Jordan, direct investment by Syrians has stimulated industry and created employment opportunities for both Jordanians and Syrians (IRC 2016). In Turkey, 26% of newly registered businesses in 2014 were either Syrian-owned or possessed Syrian capital (Del Carpio and Wagner 2015). The increase in humanitarian aid flows to host countries has also boosted local economies: each US$1 spent on humanitarian assistance had a multiplier value of 1.6 for the Lebanese economy (UNHCR and UNDP 2015). In addition, each US$1 of cash assistance spent by Syrian refugees in Lebanon generated US$2.13 of Lebanon's GDP (IRC 2016). Lebanon, Jordan and Turkey have all proven relatively economically resilient during the refugee crisis (IRC 2016).

Conclusion

Over the past 150 years the Middle East has provided refuge and asylum to numerous groups of people dispossessed of their property as a result of the upheaval leading to and including the end of empire and ensuing neo-colonial enterprises endorsed by the League of Nations. The Middle East has provided comfort and relief both on an individual basis and also for social groups. Perhaps as a residual trait of the tolerance which the Ottoman empire had enshrined in its *millet* system towards multi-ethnic and plural society, the states to emerge from the Arab Ottoman provinces all tolerated, if not actively, the development of these minority cultures.

Only in the mid-20th century did a different instrument for managing and ordering the displaced and disposed emerged – the refugee camp. Here, a system of control and standardised routine emerged as the principal tool for managing large numbers of displaced and refugee populations around the world. In the Middle East, the United Nations Relief and Works Agency, established in 1949, was set up to deal with nearly 1 million Palestinians displaced by the 1947–48 War. Here, the basics of life, food, shelter, healthcare and primary education were provided by the Agency, but the interstitial nature of the lives of the individual refugees was not addressed (Brand 1988; Farsoun and Zacharia 1997; Peteet 1991; Rosenfeld 2004).

For the earlier wave of involuntary migrants of the Middle East, return to the homelands of origin was a hope, a nostalgic dream or a unifying myth. Those early Muslim refugees of the 19th and early 20th century knew they could not go back. They had to create their homelands in new spaces. None of the populations exchanged after the 1923 Treaty of Lausanne had any ambiguity about their condition. The liminality might have been physical, but there was no question of their future. They had to create a new community, both imagined and moral, in which new ties or kinship and trade could emerge. The Kurds, perhaps more than any other group, held out for a return and alternated between a realistic hope and a nostalgic dream. Their homeland remains divided between four modern states.

Many Palestinian refugees live within a hundred miles of their original villages and urban neighbourhoods. Some can even see the lights of

their hometowns and settlements at night. Some Armenians have travelled back to visit the homeland – both in Turkey and in the Republic of Armenia. So, too, have the Circassians and other Caucasians. A few Kurds, recent migrants to Syria, have managed to smuggle themselves across the border, sometimes on the backs of Peshmergas fighters, to visit their mountainous places of birth. The effort to reverse the misfortune of displacement and dispossession and to *em-place* themselves has become a strategy for survival and its success is a measure of the resilience of the forced migrant as exhibited by the new communities established by Circassians, Armenians, Palestinians and Kurds in the Arab Middle East.

How successful forced migrants are in re-creating and re-placing themselves depends on the nature of the displacement and dispossession itself. The way people experience movement to a new place and the extent to which this is a shocking and disruptive experience is determined by the conditions under which they move and whether they can extend their notions of territorial attachment to new areas not necessarily adjacent to each other. Thus the Cretan Muslims were able to re-create their identity in several new locations outside of Crete, on the northern coast of Lebanon and Syria as well as on an island off the coast of Izmir in Turkey. For most forced migrants, however, the move is generally conducted in more traumatic conditions. The task of re-creating a place, a home or a neighbourhood, of 'producing a locality', is dominated by the effort to re-establish some continuity with the past places of origin. This work of continuity maintenance and management of memory is clearly articulated in the writings of Hirschon (2001), Parkin (1999), Malkki (1995), Loizos (1999) and Chatty (2010).

The nature of post-Ottoman Arab society – as separate from its politics – has been such that it has tolerated and acknowledged multiple layers of belonging in the struggle to make new places in the world. Although not physically displaced, the peoples of the Arab provinces of the post-Ottoman Empire have spent most of the 20th century creating new identities, and *em-placing* themselves in a new social order. This process of re-placing and re-creation has had a variety of impacts on the development of Middle Eastern nations absorbing waves of refugees, from the first arrivals of Circassians, Chechens and Armenians in the early 20th century, to the current Syrian refugee crisis. Despite the challenges of integrating large and often destitute populations into still-developing regional economies, the contributions of early forced migrants to the economic and urban development of Jordan, Lebanon and Syria, as well as the contributions of professionally skilled Palestinians to their host countries' development, both in the region as well as in the Arabian Gulf, demonstrate the strong link between forced migration and both local and regional development. More recent arrivals of Iraqi and Syrian refugees have had complex impacts on the development of host countries, many of which are still emerging. Despite the significant strains on public finances and services that these refugee influxes have created, as well as social tensions, there are indications that the presence of Iraqi and, in particular, Syrian refugees may positively impact their host countries' long-term development.

Ethnic minority communities in the Middle East have found ways to economically, physically and socially integrate themselves in their new surroundings, but at the same time resist the natural phenomena of assimilation over the long term. Patronage and real as well as 'fictive' kinship networks are powerful positive forces; so too are the religious and charitable associations which these groups set up to help those less fortunate in their communities.

Whether the current wave of dispossessed from Syria can weather the storms of dislocation with similar support and equanimity to that of their forefathers remains to be seen. Much will depend upon the way in which international humanitarian emergency assistance can unfold and develop into concerted measures to educate and assist the displaced in finding sustainable livelihoods. An educated population has agency and will contribute to the development of its host state. A current refugee population with no access to education or employment remains vulnerable and passive and a drain on the national economy. It remains to be

seen whether lessons from the late Ottoman reforms will be learned regarding the integration of refugees and other forced migrants, recognising their potential contributions to the long-term development of host countries.

See Also

- ▶ International Coordination in Asylum Provision
- ▶ Labour Markets in the Arab World

Bibliography

Al-Khatib, M.A. 2000. Language shift among the Armenians of Jordan. *International Journal of the Sociology of Language* 152: 153–177.

Al-Miqdad, F. 2007. Iraqi refugees in Syria. *Forced Migration Review Special Issue* (June 2007): 19–20. Available at: http://www.fmreview.org/sites/fmr/files/FMRdownloads/en/syria/dahi.pdf. Accessed 20 July 2016.

Al-Rasheed, M. 1994. The myth of return: Iraqi Arab and Assyrian refugees in London. *Journal of Refugee Studies* 7(2/3): 199–219.

Barkey, K., and M. Von Hagen, ed. 1997. *After Empire: Multiethnic societies and nation-building: The Soviet Union and the Russian, Ottoman, and Habsburg Empires*. Boulder: Westview.

Barth, F., ed. 1969. *Ethnic groups and boundaries: The social organization of culture difference*. Oslo: Scandinavian University Press.

Becker, K., and H. El-Said. 2013. *Management and international business issues in Jordan*. London: Routledge.

Blayney, C. 2011. *Iraqi American diasporic philanthropic remittances to Iraqi refugees in Jordan: past projects and potential for future partnerships*. American University in Cairo, John D. Gerhart Center for Philanthropy and Civic Engagement. Available at: http://dar.aucegypt.edu/handle/10526/4294. Accessed 19 July 2016.

Bocco, R., R. Jaubert, and F. Métral, ed. 1993. *Steppes d'Arabies, États, Pasteurs, Agriculteurs et Commerçants: le Devenir des Zones Sèches*. Paris: Presses Universitaires de France.

Brand, L.A. 1988. *Palestinians in the Arab World: Institution Building and the Search for State*. New York: Columbia University Press.

Brandell, I., and A. Rabo. 2003. Nations and nationalism: Dangers and virtues of transgressing disciplines. *Orientalia Suecana* LI–LII: 35–46.

Brubaker, R. 1995. Aftermaths of Empire and the unmixing of peoples: Historical and comparative perspectives. *Ethnic and Racial Studies* 18(2): 189–218.

Chaaban, J., H. Ghattas, R. Habib, S. Hanafi, N. Sahyoun, N. Salti, K. Seyfert, and N. Naamani. 2010. *Socio-economic survey of Palestinian refugees in Lebanon*. Beirut: American University of Beirut and the United Nations Relief and Works Agency for Palestinian Refugees in the Near East (UNRWA).

Chatelard, G. 2010. *Jordan: A refugee haven*. Migration Policy Institute. Available at: http://www.migrationpolicy.org/article/jordan-refugee-haven. Accessed 19 July 2016.

Chatty, D. 2010. *Dispossession and displacement in the modern Middle East*. Cambridge: Cambridge University Press.

Crisp, J., J. Janz, J. Riera, and S. Samy. 2009. *Surviving in the city: A review of UNHCR's operation for Iraqi refugees in urban areas of Jordan, Lebanon and Syria*. Geneva: Policy Development and Evaluation Service, UNHCR . Available at: http://www.unhcr.org/4a69ad639.html. Accessed 19 July 2016.

Del Carpio, X.V., and M.C. Wagner. 2015. The impact of Syrian refugees on the Turkish labor market. Policy Research working paper no. WPS 7402. World Bank Group, Washington, DC. Available at: http://documents.worldbank.org/curated/en/2015/08/24946337/impact-syrians-refugees-turkish-labor-market. Accessed 20 July 2016.

Eriksen, T.H. 1993. *Ethnicity and nationalism: Anthropological perspectives*. London: Pluto.

Farsoun, S.K., and C.E. Zacharia. 1997. *Palestine and the Palestinians*. Westview: Boulder.

Gelvin, J.L. 1998. *Divided Loyalties: Nationalism and mass politics in Syria at the close of empire*. Berkeley: University of California Press.

Hanania, M.D. 2014. The economic impact of the Palestinian Refugee Crisis on the development of Amman, 1947–1958. *British Journal of Middle Eastern Studies* 41(4): 461–482.

Harper, A. 2008. Iraq's refugees: Ignored and unwanted. *International Review of the Red Cross* 90: 869 (March): 169–190. Available at: https://www.icrc.org/eng/assets/files/other/irrc-869_harper.pdf. Accessed 20 July 2016.

Helms, C. 1981. *The cohesion of Saudi Arabia*. London: Croom Helm.

Hirschon, R. 1998. *Heirs of the Greek Catastrophe: The social life of Asia minor refugees in Piraeus*. Oxford: Berghahn.

Hirschon, R. 2001. *Surpassing Nostalgia: Personhood and the experience of displacement* (Colson Lecture). Oxford: unpublished.

Humphreys, R.S. 1999. *Between memory and desire: The middle east in a troubled age*. Egypt: University of California Press.

International Rescue Committee (IRC). 2016. Economic impacts of Syrian refugees: Existing research review and key takeaways. IRC Policy Brief No. 1 (January 2016). Available at: https://www.rescue.org/sites/default/files/document/465/ircpolicybriefeconomicimpactsofsyrianrefugees.pdf. Accessed 20 July 2016.

Kedourie, E. 1984. Minorities and majorities in the Middle East. *European Journal of Sociology* 25(2): 276–282.
Khalidi, R. 1997. *Palestinian identity: The construction of modern national consciousness*. New York: Columbia University Press.
Lerner, D., L.W. Pevsner, and D. Riesman. 1958. The passing of traditional society: Modernizing the middle east. In *Free Press*. New York: Collier-Macmillan.
Loizos, P. 1999. Ottoman half-lives: Long term perspectives on particular forced migrations. *Journal of Refugee Studies* 12(3): 237–263.
Malkki, L.H. 1995. *Purity and exile: Violence, memory, and national cosmology among Hutu refugees in Tanzania*. Chicago: University of Chicago Press.
Morris, B. 1987. *The birth of the Palestinian refugee problem, 1947–1949*. Cambridge: Cambridge University Press.
Parkin, D. 1999. Mementoes as transitional objects in human displacement. *Journal of Material Culture* 4(3): 303–320.
Peteet, J.M. 1991. *Gender in crisis: Women and the Palestinian resistance movement*. New York: Columbia University Press.
REACH. 2014. Evaluating the effect of the Syrian refugee crisis on stability and resilience in Jordanian host communities: Preliminary impact assessment. REACH. Available at: https://data.unhcr.org/syrianrefugees/download.php?id=11108. Accessed 20 July 2016
Rosenfeld, M. 2004. *Confronting the Occupation: Work, Education and Political Activism of Palestinian Families in a Refugee Camp*. Stanford: Stanford University Press.
Sassoon, J. 2009. *The Iraqi refugees: The new crisis in the middle east*. London: I.B.Tauris.
Stave, S. E. and Hillesund, S. 2015. Impact of Syrian refugees on the Jordanian labour market. International Labour Organization (ILO). Available at: http://www.ilo.org/wcmsp5/groups/public/—arabstates/—ro-beirut/documents/publication/wcms_364162.pdf. Accessed 20 July 2016.
UNHCR. 2014. *UNHCR statistical yearbook*. 14th ed. Geneva: UNHCR.
United Nations High Commissioner for Refugees (UNHCR) and United Nations Development Program (UNDP). 2015. Impact of humanitarian aid on the Lebanese economy. UNHCR and UNDP. Available at: http://reliefweb.int/sites/reliefweb.int/files/resources/Impact%20of%20Humanitarian%20Aid-UNDP-UNHCR.PDF. Accessed 20 July 2016.
Wilkinson, J. 1983. Traditional concepts of territory in South East Arabia. *The Geographical Journal* 149(3): 201–315.
World Bank. 2013. *Lebanon: Economic and social impact assessment of the Syrian conflict*. Washington, DC: World Bank. Available at: http://documents.worldbank.org/curated/en/2013/09/18292074/lebanon-economic-social-impact-assessment-syrian-conflict. Accessed 20 July 2016.

Dispute Resolution

Amy Farmer and Paul Pecorino

Abstract

The high cost of disputes creates an incentive for parties to disputes to settle. In civil litigation and arbitration, settlement failure may arise from asymmetric information or optimism. Devices to induce settlement include voluntary disclosure and mandatory discovery. The effects of these are considered, as are the English rule (whereby the loser at trial pays the reasonable legal costs of the winner), the use of contingency fees, and the operation of conventional arbitration and final offer arbitration. Researchers continue to propose new arbitration mechanisms in the hope of improving the dispute resolution process.

Keywords

Asymmetric information; Bargaining; Signalling; Screening; Arbitration; Final offer arbitration; Contingency fees; English rule; Fee-shifting; Optimism; Self-serving bias

JEL Classification

C7

Disputes may arise in a variety of settings, including labour negotiations, civil disputes and family conflict. If individuals fail to reach an agreement, there exist a variety of mechanisms for resolving the dispute. These include civil litigation, arbitration and, in labour relations, the strike. Resolving disputes in these ways is costly, thereby creating a contract zone within which both parties strictly prefer to settle. Given the high cost of disputes, considerable research has been devoted to understanding why settlement sometimes fails to occur and how different mechanisms affect the dispute

rate. Here we focus on dispute resolution in the context of civil litigation and arbitration.

Why Settlement Fails

The dominant rational choice explanation for settlement failure is asymmetric information. An alternative explanation, not consistent with the assumption of full rationality, is optimism. If agents have symmetric information and beliefs about the expected outcome of a dispute, theory suggests a settlement will occur. However, if one party has private information about the expected outcome of the dispute, settlement failure can occur. Similarly, if one or both parties to the dispute are subject to optimism, a contract zone may fail to exist.

There are two basic models in the asymmetric information literature, which make different assumptions about the structure of information. When the uninformed party makes the offer, we have a screening model which was developed by Bebchuk (1984). When the informed party makes the offer we have a signalling model developed by Reinganum and Wilde (1986). Both models' predictions are consistent with the existence of costly disputes in equilibrium.

To explore the intuition behind these models, consider a civil dispute in which the failure of negotiations would result in a trial. Suppose a plaintiff known to be harmed has private information concerning the damages she has incurred and that this information would be revealed at trial. Further, suppose the plaintiff is one of two types: a weak type with a low expected payoff at trial or a strong type with a high expected payoff. A risk-neutral plaintiff will accept a settlement offer if and only if it equals or exceeds her expected net payoff from trial. The defendant knows the probability that he is facing a weak or strong plaintiff but not the plaintiff's exact type. In a screening model, the uninformed defendant makes an offer to the plaintiff. He will choose between a low (screening) offer that only weak plaintiffs would accept and a high (pooling) offer that both types would accept. If he makes the low offer, then a strong plaintiff would proceed to trial. The screening offer is more likely to be optimal for the defendant when there is a high prevalence of weak plaintiffs, when court costs are low, and when the difference in expected trial awards for the two plaintiff types is large.

If the informed plaintiff is allowed to make the offer, this is called the signalling game. While these games generally have multiple equilibria, the D1 refinement (Cho and Kreps 1987) has been employed to focus on a separating equilibrium in which the weak plaintiff submits a low demand to the defendant, while the strong plaintiff submits a high demand. Under D1, it is assumed that an out of equilibrium offer is made by the plaintiff willing to make that offer for the largest set of acceptance probabilities. In equilibrium, the high demand must be rejected with a sufficiently high probability so as to discourage the weak plaintiff from also making this demand. These rejections lead to a positive probability of trial with the strong plaintiff.

While we used a two-type model to motivate the discussion, the Bebchuk and Reinganum and Wilde models employ a continuum of types whose distribution is known by the uninformed party. These models have been extended in numerous ways by allowing for two-sided information asymmetries (Schweizer 1989; Daughety and Reinganum 1994) and multiple offers (Spier 1992) among other extensions. While the effects of policy variables (such as cost shifting at trial) are often sensitive to the modelling details, the prediction that asymmetric information can result in costly disputes is quite robust. Excellent surveys of the literature are provided by Spier (1998) and Daughety (1999).

The empirical studies by McConnell (1989), Conlin (1999) and Osborne (1999) support the model of asymmetric information.

The optimism or self-serving bias explanation for settlement failure relies on bargainers who have potentially inaccurate beliefs about the expected outcome at trial. For example, the plaintiff's belief about the probability she will prevail at trial may exceed the defendant's belief about this same probability. If these differences in beliefs are not based on differences in information, then we are in the realm of the optimism

model. Versions of this model have been employed by Landes (1971), Posner (1973), Shavell (1982), and Priest and Klein (1984). Optimism violates rationality, but Bar-Gill (2002) finds that cautious optimism can allow the optimistic party to obtain a larger portion of the joint surplus from settlement. As a result, cautious optimism can persist in an evolutionary setting. Babcock and Loewenstein (1997) survey an experimental literature documenting the existence of a self-serving bias which leads players in the role of a plaintiff to expect a greater payoff at trial than the defendant, even though both are exposed to the same set of facts. When players are exposed to the facts of the case before being assigned their role as plaintiff or defendant, the self-serving bias tends to disappear.

Waldfogel (1998) and Farmer et al. (2004) find empirical evidence that is consistent with the optimism model. Note that the optimism and asymmetric information explanations are not mutually exclusive. It is possible that each factor is responsible for some proportion of observed disputes.

Mandatory Discovery and Voluntary Disclosure

If asymmetric information causes disputes, it is logical to ask whether voluntary disclosures and mandatory discovery can eliminate these asymmetries. In a screening model where credible disclosure is costless, Shavell (1989) shows that plaintiffs with strong cases will reveal enough information to ensure that all cases settle. Plaintiffs who do not reveal their information (those with weak cases) receive a pooling offer that all accept. However, the work of Sobel (1989) shows that this result is not robust to the introduction of positive costs of disclosure. He also shows that a costless (to the plaintiff) discovery procedure will lead to greater settlement. Farmer and Pecorino (2005) consider costly discovery and disclosure in both the signalling and the screening games. Costly disclosures may be made in the signalling game but not the screening game, while costly discovery may be invoked in the screening game but not the signalling game. If the cost of these procedures is not too high, the combination of the two will lead to a great deal of information transmission and a large reduction in the dispute rate.

Why then do disputes persist? Perhaps, as Shavell (1989) suggests, private information has strategic value if withheld until trial. Hay (1995) develops a model in which an initial informational asymmetry on the merits of the case is resolved by mandatory discovery, but by the time this occurs a new asymmetry – namely, the extent of attorney preparation – has emerged. This second asymmetry leads to trials in the equilibrium of the model. Hay notes that the extent of attorney preparation is not subject to discovery. This is also true of preferences. Farmer and Pecorino (1994) show that asymmetric information on risk preferences can lead to trial, and that this information is neither subject to discovery nor easy to credibly transmit. As a result, this type of asymmetry may tend to persist in the face of mandatory discovery and opportunities for voluntary disclosure.

Other Institutional Features

There is a voluminous literature which examines how a variety of institutional features affect settlement in civil litigation. What follows is a much abbreviated discussion of a large and complex literature. One difficulty in addressing this question is that even a single institution is likely to have multiple effects on the litigation process. Thus, a single institution may have conflicting effects on the dispute rate and may also have important influences on other aspects of the litigation process.

A classic example of this difficulty is reflected in the analysis of the English rule under which the loser at trial pays the reasonable legal costs of the winner. If the probability of a finding for the plaintiff at trial is private information, then fee shifting at trial reduces settlement rates by, in effect, spreading out the distribution of player types (Bebchuk 1984). If players are optimistic in their assessments of the probability that the plaintiff will prevail, then fee shifting will aggravate this optimism and reduce the probability that a contract zone will exist (Shavell 1982).

This prediction – that fee shifting will increase the probability of trial – is made with expenditure at trial held constant. It is well established that the fee shifting at trial will increase expenditure (Braeutigam et al. 1984). If the expenditure effect is strong enough, it can result in fewer (but more costly) disputes (Hause 1989). The English rule also affects the mix of cases which are filed. It discourages cases where there are large stakes but a low probability of success, and encourages low stakes cases with a high probability of success (Shavell 1982).

Many of the theoretical predictions on fee shifting at trial appear to be borne out in the data (see Hughes and Snyder 1998).

Under a contingency fee, the plaintiff's lawyer receives a percentage of the judgment at trial if the plaintiff wins the case and nothing if she loses. The effects of contingency fees on the litigation process are very complex and wide ranging (see Rubinfeld and Scotchmer 1998, for a survey). However, one effect of contingency fees on settlement is clear: if the attorney controls the settlement decision, he will have an excessive incentive to settle the case relative to the interests of his client. The reason is that the attorney bears most of the costs of a trial but is paid only a fraction of the award. On the other hand, if the client controls the case, she may have an excessive incentive to reject a settlement offer and bring the case to trial. (This is particularly true if the contingency percentage is not lower for cases which settle early.)

When a single defendant faces multiple plaintiffs in sequence, some interesting issues regarding settlement arise. Spier (2003a, b) and Daughety and Reinganum (2004) have analysed the use of most favoured nation (MFN) clauses in the context of repeat litigation. Suppose a plaintiff settles early under MFN. If another plaintiff later settles for more, the early settlement is adjusted upward. An MFN clause can be a mechanism whereby the defendant commits to not raising his offer to plaintiffs who settle later in the process. While there is some ambiguity of the effects of MFN on settlement rates and the overall dispute costs (see especially Spier 2003a), the general thrust of these papers suggests that MFN clauses are efficiency enhancing in the sense that they will reduce the expected dispute costs associated with litigation.

Arbitration

Under conventional arbitration (CA), the arbitrator is free to impose her preferred settlement on the bargaining parties. Under final offer arbitration (FOA), each party to the dispute submits an offer to the arbitrator who must pick one of the submitted offers. While there is some evidence that submitted offers affect the outcome in CA (Farber and Bazerman 1986), for the purpose of the following discussion we assume that they do not. From a modelling standpoint, this makes CA look exactly like a simple version of civil litigation. Under FOA, the submitted offers clearly affect the outcome, a feature which has important implications for dispute resolution.

Consider the two-type version of the screening model where the plaintiff can have a strong or a weak case. In CA, the defendant will either make an offer that only a weak plaintiff will accept or a pooling offer that both types will accept. If all negotiation takes place prior to the submission of offers to the arbitrator, then under FOA it is possible that the defendant will make an offer that neither type will accept, resulting in a 100% dispute rate (Farmer and Pecorino 2003). This can occur because the sequentially rational offer submitted to the arbitrator influences the acceptable settlement prior to arbitration. The lack of early settlement allows the defendant to commit to an offer which is optimal against the entire distribution of plaintiff types. Farmer and Pecorino (2003) also show (in contrast to CA) that costless voluntary disclosure never takes place when FOA is the dispute resolution mechanism. The reason is that information has strategic value in this game. Both of the impediments to settlement discussed above disappear if bargaining is permitted after offers are submitted to the arbitrator but prior to the arbitration hearing.

While not totally conclusive on this point, the results of Farmer and Pecorino (1998) also suggest that allowing for bargaining after offers are submitted to the arbitrator can increase settlement

for reasons different from those discussed above. Because a submitted offer affects the outcome of arbitration, it will tend to reflect private information. This may in turn promote settlement. Taken together, the results on FOA suggest that the effects of this institution on settlement are sensitive to whether or not bargaining occurs in the face of offers submitted to the arbitrator. In major league baseball, a prominent use of FOA, a good deal of bargaining and settlement occurs after offers have been submitted to the arbitrator.

FOA was proposed by Stephens (1966) and has since become an important alternative to CA. Researchers continue to propose new arbitration mechanisms in the hope of improving the dispute resolution process. Combined arbitration (Brams and Merrill 1986) is a mixture of FOA and CA. Other proposed mechanisms include tri-offer arbitration (Ashenfelter et al. 1992) and amended final offer arbitration (Zeng 2003).

See Also

- ▶ Epistemic Game Theory: An Overview
- ▶ Epistemic Game Theory: Incomplete Information

Bibliography

Ashenfelter, O., J. Currie, H. Farber, and M. Spiegel. 1992. An experimental comparison of dispute rates in alternative arbitration systems. *Econometrica* 60: 1407–1433.

Babcock, L., and G. Loewenstein. 1997. Explaining bargaining impasse: The role of self-serving biases. *Journal of Economic Perspectives* 11: 109–126.

Bar-Gill, O. 2002. *The success and survival of cautious optimism: Legal rule, and endogenous perceptions in pre-trial settlement negotiations*. Public law working paper No. 35. Cambridge, MA: Harvard Law School.

Bebchuk, L. 1984. Litigation and settlement under imperfect information. *RAND Journal of Economics* 15: 404–415.

Braeutigam, R., B. Owen, and J. Panzar. 1984. An economic analysis of alternative fee shifting systems. *Law and Contemporary Problems* 47: 173–185.

Brams, S., and S. Merrill. 1986. Binding versus final-offer arbitration: A combination is best. *Management Science* 32: 1346–1355.

Cho, I., and D. Kreps. 1987. Signaling games and stable equilibria. *Quarterly Journal of Economics* 102: 179–222.

Conlin, M. 1999. Empirical test of a separating equilibrium in National Football League contract negotiations. *RAND Journal of Economics* 30: 289–304.

Daughety, A. 1999. Settlement. In *Encyclopedia of law and economics*, vol. 5, ed. B. Bouckaert and G. de Geest. Cheltenham: Edward Elgar.

Daughety, A., and J. Reinganum. 1994. Settlement negotiations with two-sided asymmetric information: Model duality, information distribution, and efficiency. *International Review of Law and Economics* 14: 283–298.

Daughety, A., and J. Reinganum. 2004. Exploiting future settlements: A signaling model of most-favored-nation clauses in settlement bargaining. *RAND Journal of Economics* 35: 467–485.

Farber, H., and M. Bazerman. 1986. The general basis of arbitrator behavior: An empirical analysis of conventional and final offer arbitration. *Econometrica* 54: 819–854.

Farmer, A., and P. Pecorino. 1994. Pretrial negotiations with asymmetric information on risk preferences. *International Review of Law and Economics* 14: 273–281.

Farmer, A., and P. Pecorino. 1998. Bargaining with informative offers: An analysis of final offer arbitration. *Journal of Legal Studies* 27: 415–432.

Farmer, A., and P. Pecorino. 2003. Bargaining with voluntary transmission of private information: Does the use of final offer arbitration impede settlement? *Journal of Law, Economics, and Organization* 19: 64–82.

Farmer, A., and P. Pecorino. 2005. Civil litigation with mandatory discovery and voluntary transmission of private information. *Journal of Legal Studies* 34: 137–159.

Farmer, A., P. Pecorino, and V. Stango. 2004. The causes of bargaining failure: Evidence from major league baseball. *Journal of Law and Economics* 47: 543–568.

Hause, J. 1989. Indemnity, settlement, and litigation, or I'll be suing you. *Journal of Legal Studies* 18: 157–179.

Hay, B. 1995. Effort, information, settlement, trial. *Journal of Legal Studies* 24: 29–62.

Hughes, J.W., and E.A. Snyder. 1998. Allocation of litigation costs: American and English rules. In *The new Palgrave dictionary of economics and the law*, vol. 1, ed. P. Newman. London: Macmillan.

Landes, W. 1971. An economic analysis of the courts. *Journal of Law and Economics* 14: 61–107.

McConnell, S. 1989. Strikes, wages and private information. *American Economic Review* 79: 801–815.

Osborne, E. 1999. Who should be worried about asymmetric information in litigation? *International Review of Law and Economics* 19: 399–409.

Posner, R. 1973. An economic approach to legal procedure and judicial administration. *Journal of Legal Studies* 2: 399–458.

Priest, G., and B. Klein. 1984. The selection of disputes for arbitration. *Journal of Legal Studies* 13: 215–243.

Reinganum, J., and L. Wilde. 1986. Settlement, litigation, and the allocation of litigation costs. *RAND Journal of Economics* 17: 557–566.

Rubinfeld, D., and S. Scotchmer. 1998. Contingent fees. In *The new Palgrave dictionary of economics and the law*, vol. 1, ed. P. Newman. London: Macmillan.

Schweizer, U. 1989. Litigation and settlement under two-sided incomplete information. *Review of Economic Studies* 56: 163–178.
Shavell, S. 1982. Suit, settlement, and trial: A theoretical analysis under alternative methods for the allocation of legal costs. *Journal of Legal Studies* 11: 55–82.
Shavell, S. 1989. Sharing of information prior to settlement or litigation. *RAND Journal of Economics* 20: 183–195.
Sobel, J. 1989. An analysis of discovery rules. *Law and Contemporary Problems* 52: 133–159.
Spier, K. 1992. The dynamics of pretrial negotiation. *Review of Economic Studies* 59: 93–108.
Spier, K. 1998. Settlement of litigation. In *The new Palgrave dictionary of economics and the law*, vol. 3, ed. P. Newman. London: Macmillan.
Spier, K. 2003a. 'Tied to the mast': Most-favored-nation clauses in settlement contracts. *Journal of Legal Studies* 32: 91–120.
Spier, K. 2003b. The use of 'most-favored-nation' clauses in settlement of litigation. *Rand Journal of Economics* 34: 78–95.
Stephens, C. 1966. Is compulsory arbitration compatible with bargaining? *Industrial Relations* 5(1): 38–52.
Waldfogel, J. 1998. Reconciling asymmetric information and divergent expectations theories of litigation. *Journal of Law and Economics* 41: 451–476.
Zeng, D. 2003. An amendment to final offer arbitration. *Mathematical Social Sciences* 46(1): 9–19.

Distortions

T. N. Srinivasan

The voluminous literature on distortions, including a masterly survey by Jagdish Bhagwati (1971), contains no formal definition of the term distortion. The analysis often proceeds in terms of specific examples. Bhagwati analyses distortions in the context of foreign trade policies and the welfare of home consumers. He characterizes distortions as departures from the equality of the marginal rate of transformation of one commodity into some other through foreign trade (the so-called foreign rate of transformation) with transformation through domestic production (the domestic rate of transformation) and with the marginal rate of substitution in the consumption of the same pair of commodities by each consumer. Also, the failure to achieve aggregate production efficiency, in the sense of not producing on the boundary of the set of production possibilities given available resources and technology, is deemed a distortion.

Given non-interdependent consumer preferences the above equalities and production efficiency are indeed necessary conditions (leaving aside corner optima) for a feasible allocation to be Pareto Optimal. Any combination of production, consumption, and foreign trade vectors such that there is no positive excess demand for any commodity is feasible. Pareto Optimality ensures that no other feasible allocation can make at least one consumer better off without making some other consumer worse off. These conditions are also sufficient if it is assumed that the aggregate production set and individual preferences are convex. Further, if each consumer has a positive endowment of every commodity, any such Pareto Optimal allocation can be sustained as a competitive equilibrium provided redistribution of endowments among consumers or of their incomes through lump sum transfers in feasible. This is in essence the second fundamental theorem of neoclassical welfare economics.

Bhagwati, and others analysing distortionary taxation (Atkinson and Stiglitz 1980), factor market distortions (Magee 1976) etc., all seem to take (implicitly) the second fundamental theorem as the point of departure. This suggests the following definition: *a distortion exists in an economy in which lump sum transfers or their equivalents are feasible redistribution instruments, if some Pareto Optimal allocations from the point of view of consumers of that economy cannot be characterized as competitive equilibria.* This is consistent with the argument of Arrow (1964) that

> the best developed part of the theory (of externalities) related only to a single problem: the statement of a set of conditions, as weak as possible, which insure that a competitive equilibrium exists and is Pareto-efficient. Then the denial of any of these hypotheses is presumably a sufficient condition for considering resort to nonmarket channels of resource allocation – usually thought of as government expenditures, taxes, and subsidies.

A distortion can clearly arise in situations when some of the premises of the theorem are violated. Obviously, if lump sum income transfers are infeasible, even if all the other premises are satisfied some Pareto Optima cannot be characterized as competitive equilibria. Hence any redistribution achieved through other instruments may be Pareto dominated by an equilibrium achievable with lump sum transfers. Violations of other premises will in general call for the use of additional policy instruments besides lump sum transfers. If externalities or increasing returns in production lead to non-convexity of the aggregate production set, in general, a set of Pigouvian taxes and subsidies are the needed additional instruments. If price-taking producers and consumers do not perceive a country's market power in its foreign trade then taxes on foreign trade are needed.

The preceding discussion implicitly views the distortions as structural features of the economy or in Bhagwati's terminology, as endogenous and the policies as 'first best' responses that assure Pareto Optimality. However, if the same policy instruments are used in the absence of distortions or at levels that are not 'first best optimal' in their presence, the resulting equilibrium will be Pareto dominated by another achievable equilibrium by refraining from their use in the former case and by using them at first best optimal levels in the latter. Bhagwati characterizes such inappropriate use of policy as autonomous policy imposed distortion.

Policymakers may have other objectives besides consumer welfare. Johnson (1965) termed such social concerns non-economic objectives. The literature that followed (Bhagwati and Srinivasan 1969) addressed two policy questions. The first derives the 'first best' policy that achieves a given non-economic objective, with the least cost in terms of consumer welfare. One feasible policy for achieving the non-economic objective has a higher cost than another, if the equilibrium associated with the former is Pareto dominated by an equilibrium achievable using the latter and making lump sum transfers between consumers as needed. The second question is the ranking of alternative feasible policies starting from the 'first best' to 'second best', 'third best' etc. in terms of their cost in achieving the non-economic objective.

Governments may wish to raise the output (for instance, for reasons of national defence) of some industry above its level in a laissez-faire competitive equilibrium through policy intervention. A production subsidy (or its equivalent) to that industry is the appropriate first best policy for achieving the production (non-economic) objective. Such a subsidy would be non-optimal, and hence a distortion, in the absence of the objective. Bhagwati characterizes such interventions as policy imposed instrumental distortions, the word instrumental signifying that the policy is an instrument for achieving non-economic objectives. Trade tariffs and quotas, consumption taxes and subsidies, wage subsidy and similar factor use taxes or subsidies turn out to be the first best policy instruments for achieving suitably specified non-economic objectives. Each such policy could also be a feasible policy for achieving non-economic objectives for which it is not the first best.

The impacts, measured in alternative ways, of particular policies or processes in the presence of a distortion rather than the optimal policy response to it have been analysed. An example is the impact of an import tariff in an economy with a distortion in the labour market in the form of a minimum wage above its market clearing level. The literature on immiserizing growth (Bhagwati 1968) and its offshoots analyse the impact of the processes factor accumulation, technical change, external capital inflow, etc. on an economy with a tariff distortion. This diverse literature establishes two important common propositions. First, given an existing distortion, the impact of policies other than the first best or of processes could be in a direction opposite to that they would have taken had there been no distortion or had the distortion been addressed with a first best policy. For example, the accumulation of capital which would have increased consumer welfare had there been an optimal tariff could be welfare-worsening in its absence. The shadow price of a factor to be used in social cost benefit analysis is a small open economy with a distortionary tariff can be negative. Thus the withdrawal of that factor from its existing employment for use in a project instead of adding to the project's cost increases its social value! (Srinivasan and Bhagwati 1978). An

implication of this is that some of the production activities in an economy are subtracting rather than adding value at shadow prices, while obviously their value added at market prices is positive.

The second proposition shows that policies other than the first best, even if distortionary, can increase welfare. Thus given an existing distortion, introduction of another can improve welfare. In the so called Harris-Todaro (1970) economy a distortionary minimum wage is enforced in urban manufacturing activity. Rural workers migrate to urban areas as long as their expected wage (taking into account the probability of being unemployed) exceeds the rural wage. If the first best policy of a wage subsidy to both sectors is not feasible, an output subsidy to agriculture can improve welfare compared to the laissez-faire equilibrium. The second proposition is an illustration of the general theorem of the second best: 'if there is introduced into a general equilibrium system a constraint which prevents the attainment of one of the Paretian conditions, the other Paretian conditions, though still attainable, are in general, not desirable' (Lipsey and Lancaster 1956). The constraint of this theorem is the equivalent of a distortion and violating other attainable Paretian conditions is equivalent to introducing other distortions. The theory of the second best is rigorously analysed by Guesnerie (1979, 1980).

The literature on rent seeking (Krueger 1974) and directly unproductive profitseeking (DUP) activities (Bhagwati 1982) has highlighted another aspect of distortions. A distortion by raising the demand price of a commodity above its relevant supply price creates a rent that may trigger a competition for acquiring it. For example, an import quota (tariff) by raising the domestic price above the import price could trigger a competition for quota rents (tariff revenues), thereby diverting resources from production. However, such a diversion takes place in the context of an existing distortion (an inappropriate quota or tariff) and as such, paradoxically, it can improve consumer welfare if it succeeds in reducing the welfare loss associated with the distortion more than the welfare loss it creates in reducing resources available for production. It has also been shown that the welfare ranking of policies that achieve a given non-economic objective can be reversed once seeking activities triggered by such policies are taken into account.

To sum up, a distortion by definition creates a welfare loss; first best optimal policies could often be devised to offset this loss; if for some reason, first best policies are infeasible, other welfare-improving policies may exist and sometimes, they can be ranked as 'second best', 'third best' etc.; however, such policies, can have effects in directions opposite to those they would have had in the absence of a distortion; distortions have implications for social cost-benefit analysis; finally distortions can trigger rent-seeking activities.

See Also

- ▶ Optimal Tariffs
- ▶ Pareto Efficiency
- ▶ Pareto Distribution
- ▶ Second Best
- ▶ Taxes and Subsidies

Bibliography

Arrow, K.J. 1964/1970. Political and economic evaluation of social effects and externalities. In *Analysis of public output*, ed. J. Margolis. New York: National Bureau for Economic Research.

Atkinson, A.B.., and J.E. Stiglitz. 1980. *Lectures on public economics*. New York: McGraw Hill.

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

Bhagwati, J.N. 1971. The generalized theory of distortions and welfare. Chapter 12 in *Trade, balance of payments and growth: Papers in international economics in honor of Charles P. Kindleberger*, ed. J.N. Bhagwati, R.W. Jones, R. Mundell, and J. Vanek. Amsterdam: North-Holland.

Bhagwati, J.N. 1982. Directly unproductive profit – Seeking (DUP) activities. *Journal of Political Economy* 90: 988–1002.

Bhagwati, J.N., and T.N. Srinivasan. 1969. Optimal intervention to achieve non-economic objectives. *Review of Economic Studies* 36: 27–38.

Guesnerie, R. 1979. General statements on second best Pareto optimality. *Journal of Mathematical Economics* 6: 169–194.

Guesnerie, R. 1980. Second-best policy rules in Boiteux tradition. *Journal of Public Economics* 13: 51–58.
Harris, J.R., and M.P. Todaro. 1970. Migration, unemployment and development: A two sector analysis. *American Economic Review* 60: 126–142.
Johnson, H.G. 1965. Optimal trade intervention in the presence of domestic distortions. Chapter 11 in *Trade, growth and the balance of payments*, ed. R.E. Caves, H.G. Johnson, and P.B. Kenen. Amsterdam: North-Holland.
Krueger, A. 1974. The political economy of the rent-seeking society. *American Economic Review* 64: 291–303.
Lipsey, R.G., and K. Lancaster. 1956. The general theory of the second-best. *Review of Economic Studies* 24: 11–32.
Magee, S.P. 1976. *International trade and distortions in factor markets*. New York: Marcell Dekker.
Srinivasan, T.N., and J.N. Bhagwati. 1978. Shadow prices for project selection in the presence of distortions: Effective rates of protection and domestic resource costs. *Journal of Political Economy* 86: 97–116.

Distributed Lags

Philip Hans Franses

Abstract

This article reviews various aspects of distributed lag models. Specific attention is paid to the interpretation of model parameters.

Keywords

Almon lags; ARMA models; Cointegration; Distributed lags; Error correction models; Koyck transformation; Maximum likelihood; Multicollinearity; Nonlinear least squares; Ordinary least squares; Vector autoregressions

JEL Classifications

C22

Distributed lag models correlate a single dependent variable with its own lags and with current and lagged values of one or more explanatory variables. Examples concern the current and dynamic correlations between output and investment and between sales and advertising. Distributed lag models typically assume that the explanatory variable is exogenous. (In case of doubt, one usually resorts to vector autoregressive models where two or more variables can be endogenous; see Sims 1980.)

This article highlights a few aspects of distributed lag models. The two main aspects are representation and interpretation. Useful extended surveys appear in Dhrymes (1971), Griliches (1967) and Hendry et al. (1984).

Representation

Consider a dependent variable y_t and, for ease of notation, a single explanatory variable x_t. Indicator t runs from 1 to n and it can concern seconds, hours, days or even years. A general (autoregressive) distributed lag model is given by

$$y_t = \mu + \alpha_1 y_{t-1} + \ldots + \alpha_p y_{t-p} + \beta_0 x_t + \beta_1 x_{t-1} + \ldots + \beta_m x_{t-m} + \varepsilon_t, \quad (1)$$

where p and m can take any positive integer value, and where it is usually assumed that ε_t is an uncorrelated variable with mean zero and variance σ^2. (Part of the literature assumes the label distributed lags model for the case where $p = 0$ and $m = \infty$. Below we will see that such a model is often approximated by a model as in (1).)

As the model contains the lagged dependent variables, it is called an autoregressive distributed lag model with orders p and m, in short ADL(p, m). The model allows for delayed effects of x_t, as β_0 can be 0, and it also allows for time gaps in these effects when some β parameters are zero and others are not.

Reducing the Number of Parameters

Basically, given fixed and finite values of p and m, the parameters in (1) can be consistently estimated with ordinary least squares (OLS). (Typically one uses information criteria as those of Akaike or Schwarz to choose the relevant values of p and m in practice.) In practice, p and m can be large, and in theory even as large as ∞. This can be

inconvenient, for two reasons. First, the variables y_t and x_t each can be strongly autocorrelated, and then the regression in (1) suffers from multicollinearity. Second, with many parameters in a model there might be many values to evaluate and interpret.

To reduce the number of parameters and to facilitate interpretation, one can impose restrictions. Early suggestions are the Almon and Shiller lag structures, where the parameters are made functions of i, $i = 0, 1, 2, \ldots, m$ (see Almon 1965, and Shiller 1973), and the so-called Koyck transformation (see Koyck 1954).

Almon and Shiller Transformations

Consider the version of (1) with $p = 0$ and $m = m$ and set μ at 0 for convenience, that is, consider

$$y_t = \beta_0 x_t + \beta_1 x_{t-1} + \ldots + \beta_m x_{t-m} + \varepsilon_t. \quad (2)$$

Almon (1965) proposes to reduce the number of parameters by assuming the approximation

$$\beta_i = \alpha_0 + \alpha_1 i + \alpha_2 i^2 + \ldots + \alpha_q i^q. \quad (3)$$

with $q > m$. This makes the sequence of β_i parameters a polynomial and hence a smooth function without possibly implausible spikes.

Working out the Almon lags, one can derive that the structure implies that

$$\beta_{i+1} - 2\beta_i + \beta_{i-1} = \gamma_i, \quad (4)$$

where γ_i is a function of α_i values. Shiller (1973) considers this as too restrictive and he proposes to assume that

$$\beta_{i+1} - 2\beta_i + \beta_{i-1} \sim N(0, \zeta^2), \quad (5)$$

for $i = 1, 2, \ldots, m - 1$.

Koyck Transformation

The Koyck model can be interpreted as a model which includes adaptive expectations. Suppose that

$$y_t = \alpha + \beta x_t^* + \varepsilon_t, \quad (6)$$

where x_t^* denotes the expected value of x_t, an expectation formed at $t - 1$. When the adaptive expectations schedule is assumed, like

$$x_t^* = \lambda x_{t-1}^* + (1 - \lambda) x_t, \quad (7)$$

with again $|\lambda| < 1$, then substituting (7) into (6) gives

$$y_t = \alpha(1 - \lambda) + \lambda y_{t-1} + \beta(1 - \lambda) x_t + \varepsilon_t - \lambda \varepsilon_{t-1}. \quad (8)$$

The short-run effect of x_t on y_t is $\beta(1 - \lambda)$, while the long-run effect is $\frac{\beta(1-\lambda)}{1-\lambda} = \beta$, as could be expected given (6).

Consider the case were m equals ∞, and where all α parameters are set to zero. When it is further assumed that $\beta_j = \beta_0 \lambda^{j-1}$, with $|\lambda| < 1$ for j is 1,2,..., then (1) becomes

$$y_t = \mu + \beta_0 x_t + \beta_0 \lambda x_{t-1} + \beta_0 \lambda^2 x_{t-2} + \ldots + \varepsilon_t. \quad (9)$$

Subtracting λy_{t-1} from this expression gives

$$y_t = (1 - \lambda)\mu + \lambda y_{t-1} + \beta_0 x_t + \varepsilon_t - \lambda \varepsilon_t - 1, \quad (10)$$

which is again (8). This Koyck transformation leads to a rather simple model with a moving average (MA) error term. The appropriate estimation method is maximum likelihood, as it is described in, for example, Hamilton (1994, p. 132) for general ARMA models. Note that the parameter λ appears in the autoregressive part and in the MA part.

Restructuring the Model

An alternative way to reduce the number of parameters, also in order to facilitate interpretation, is to restructure the model.

To overcome multicollinearity, one can rewrite model (1) in the so-called error correction format. This format combines levels and differences of levels, which is convenient as these are usually much less correlated than the levels themselves,

and hence multicollinearity will be much less of a problem. An additional feature of the error correction format is that it provides an immediate look at key parameters such as the total effect, the current effect, and the speed at which the total effect is accomplished.

With Δ_j denoted as the j-th order differencing filter, that is, $\Delta_j y_t = y_t - y_{t-j}$, an error correction representation for (1) reads as

$$\begin{aligned}\Delta_1 y_t = \mu &+ \left(\sum_{j=1}^{p}\alpha_j - 1\right)\left[y_{t-1} - \frac{\sum_{i=0}^{m}\beta_i}{1-\sum_{j=1}^{p}\alpha_j}x_{t-1}\right] \\ &+ \beta_0\Delta_1 x_t - \sum_{i=2}^{m}\beta_i\Delta_{i-1}x_{t-1} \\ &- \sum_{j=2}^{p}\alpha_j\Delta_{i-1}y_{t-1} + \varepsilon_t,\end{aligned} \tag{11}$$

where lagged levels are suitably combined into differenced variables such that at each lag a higher-order differenced variable appears. This representation even further reduces chances of having multicollinearity. Note that the model can also be written in terms of lagged levels and first differences only, that is as

$$\begin{aligned}\Delta_1 y_t = \mu &+ \left(\sum_{j=1}^{p}\alpha_j - 1\right)\left[y_{t-1} - \frac{\sum_{i=0}^{m}\beta_i}{1-\sum_{j=1}^{p}\alpha_j}x_{t-1}\right] \\ &+ \beta_0\Delta_1 x_t - \sum_{i=1}^{m}\gamma_i\Delta_1 x_{t-i} \\ &+ \sum_{j=1}^{p}\theta_j\Delta_{i-1}y_1 + \varepsilon_t\end{aligned} \tag{12}$$

With the use of (11), all but two parameters (that is, α_1 and β_1) can be directly estimated by using OLS, while $\hat{\alpha}_1$ and $\hat{\beta}_1$ straightforwardly follow from applying OLS to (1). Note that model (11) can also be written such that the levels (now at $t-1$) enter at $t-2$ or, say, $t-p$.

Interpretation

We now turn to the interpretation of distributed lag models.

Long-Run and Short-Run Effects

The error correction model in (11) provides immediate estimates of current and dynamic effects. (Fok et al. 2006, show that when the series y_t and x_t have a unit root and are cointegrated, as defined by Engle and Granger 1987, one should speak of the long-run effect, while when the series are stationary there is a total or cumulative effect. For the latter, see also Hendry et al. 1984.) The current effect is β_0 and the long-run or total effect is

$$\frac{\sum_{i=0}^{m}\beta_i}{1-\sum_{j=1}^{p}\alpha_j}. \tag{13}$$

Note that the long-run effect can be larger or smaller than the short-run effect, depending on the values of the parameters. The parameters in the error correction model, when written as

$$\begin{aligned}\Delta_1 y_t = \mu + \rho[y_{t-1} - \gamma x_{t-1}] + \beta_0\Delta_1 x_t \\ - \sum_{i=2}^{m}\beta_i\Delta_{i-1}x_{t-1} \\ - \sum_{j=2}^{p}\alpha_j\Delta_{j-1}y_{t-1} + \varepsilon_t,\end{aligned} \tag{14}$$

can be estimated using non-linear least squares. This method provides direct estimates of the long-run effect γ and its associated standard error.

Duration Interval

As well as the long-run and short-run effects, one may also be interested in the speed with which the effect of x_t decays over time. To be able to compute this so-called duration interval, one needs explicit expressions for $\frac{\partial y_{t+k}}{\partial x_t}$ for all values of k running from 1 to, potentially, ∞. Given the expression in (1), these expressions are easily derived as

D

$$\frac{\partial y_t}{\partial x_t} = \beta_0$$
$$\frac{\partial y_{t+1}}{\partial x_t} = \beta_1 + \alpha_1 \frac{\partial y_t}{\partial x_t}$$
$$\frac{\partial y_{t+2}}{\partial x_t} = \beta_2 + \alpha_1 \frac{\partial y_{t+1}}{\partial x_t} + \alpha_2 \frac{\partial y_t}{\partial x_t}$$
$$\vdots$$
$$\frac{\partial y_{t+k}}{\partial x_t} = \beta_k| + \sum_{j=1}^{k} \alpha_j \frac{\partial y_{t+(k-j)}}{\partial x_t}$$

where it should be noted that $\alpha_k = 0$ for $k > p$, and that $\beta_k = 0$ for $k > m$. Hence, the final form of a distributed lag model (see Harvey 1990), is

$$y_t = \sum_{i=0}^{\infty} \delta_i x_{t-i} + error, \tag{15}$$

where

$$\delta_i = \frac{\partial y_{t+i}}{\partial x_t}. \tag{16}$$

With these δ_i, one can derive all kinds of summary effects (like mean and median, or half lives of shocks) of x_t on y_t.

When δ_i decays monotonically, it is useful to define the decay factor by

$$p_k = \frac{\frac{\partial y_t}{\partial x_t} - \frac{\partial y_{t+k}}{\partial x_t}}{\frac{\partial y_t}{\partial x_t}}$$

This can be computed only for discrete values of k as there are only discrete time intervals. This decay factor is a function of the model parameters. Through interpolation, one can decide on the time k it takes for the decay factor to be equal to some value of p, which typically is equal to 0.95 or 0.90. This estimated time k is then called the p per cent duration interval. This measure is frequently used in advertising research (see Clarke 1976; Leeflang et al. 2000; Franses and Vroomen 2006).

Final Issues

Distributed lag models continue to be a standard empirical approach. When the models are applied, there are at least two further issues that one needs to address, that is, next to selecting p and m and a useful transformation. The first concerns the statistical analysis of the model. For example, if y_t and x_t are not stationary, one needs to rely on cointegration techniques that involve non-standard asymptotic theory. The theory that is most relevant here is formulated in Boswijk (1995). Also, in the case of the Koyck model, one faces the so-called Davies (1987) problem. Under the null hypothesis that $\beta_0 = 0$, the model collapses to $y_t = \varepsilon_t$, and hence λ is not identified then.

The second issue concerns aggregation over time. It may be that y_t and x_t are not available at the same sampling frequency. For example, television commercials last for 30 seconds and recur each hour, say, while sales data are available only at the weekly level. Tellis and Franses (2006) have a few recent results, but more work is needed.

Bibliography

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

Boswijk, H. 1995. Efficient Inference on cointegration parameters in structural error–correction models. *Journal of Econometrics* 69: 133–158.

Clarke, D. 1976. Econometric measurement of the duration of advertising effect on sales. *Journal of Marketing Research* 8: 345–357.

Davies, R. 1987. Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* 64: 247–254.

Dhrymes, P. 1971. *Distributed lags: Problems of estimation and formulation*. San Francisco: Holden-Day.

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

Fok, D., C. Horvath, R. Paap, and P. Franses. 2006. A hierarchical Bayes error correction model to explain dynamic effects of price changes. *Journal of Marketing Research* 43: 443–461.

Franses, P., and B. Vroomen. 2006. Estimating confidence bounds for advertising effect duration intervals. *Journal of Advertising* 35(Summer): 33–37.

Griliches, Z. 1967. Distributed lags: A survey. *Econometrica* 35: 16–49.

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

Hansen, B. 1996. Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica* 64: 413–430.

Harvey, A. 1990. *The econometric analysis of time series*. London: Philip Allan.
Hendry, D., A. Pagan, and J. Sargan. 1984. Dynamic specification. In *Handbook of econometrics*, ed. Z. Griliches and M. Intriligator, Vol. 2. Amsterdam: North-Holland.
Koyck, L. 1954. *Distributed lags and investment analysis*. Amsterdam: North-Holland.
Leeflang, P., D. Wittink, M. Wedel, and P. Naert. 2000. *Building models for marketing decisions*. Boston: Kluwer.
Shiller, R. 1973. A distributed lag estimator derived from smoothness priors. *Econometrica* 41: 775–788.
Sims, C. 1980. Macroeconomics and reality. *Econometrica* 48: 1–48.
Tellis, G., and P. Franses. 2006. The optimal data interval for econometric models of advertising. *Marketing Science* 23: 217–229.

Distribution Theories: Keynesian

Mauro Baranzini

As Kaldor has pointed out, Keynes was never interested in the problem of distribution of income as such; the determination of its level was his main concern: 'One may nevertheless christen a particular theory of distribution as 'Keynesian' if it can be shown to be an application of the specifically Keynesian apparatus of thought' (Kaldor 1956, p. 94). Since the middle Fifties a large number of neo-or post-Keynesian models of economic growth and income distribution have appeared, originating mainly in the University of Cambridge. Post-Keynesian distribution theory now occupies an undisputed place in most macroeconomic textbooks. These models have been labelled as 'post-Keynesian' since savings passively adjust to the externally given full-employment investment, via redistribution of income between wages and profits and/or among social classes. This contrasts with the pre-Keynesian or neo-classical framework, where investment is governed by saving, and where the production function and marginal productivity theory play a crucial role in determining income distribution. The 'post-Keynesian' model also differs from the static Keynesian scheme, where changes in the level, rather than in the distribution, of income ensure equality between saving and investment.

The Role of Social Classes

In order to study income distribution in classical, post-Keynesian and neo-Ricardian theories it is relevant to define what kind of relationships exist between property and income earners. Property rights are fundamental determinants of distribution if the production process requires some form of cooperation from individuals having the power of 'withdrawing' certain essential inputs. As in most classical theories, social classes remain crucial for post-Keynesian theories, and their distinctive feature is given by their saving and consumption behaviour. In more sophisticated theories (Pasinetti) the assumption of 'separate appropriation' of each production factor is no longer in the foreground, and workers' income is made up by wages and profits on accumulated savings. In a general sense in post-Keynesian theories different rates of saving are associated with different economic or social classes. And the distribution of income among classes will be such as to yield an overall saving equal to the desired level of full-employment investment. We shall start by considering the origin of Keynesian income distribution theories.

The Hharrod–Domar Dilemma

When in the late Thirties and in the Forties the first macro-economic models of economic growth were developed, the theory of income distribution was caught in an impasse, represented by the well-known Harrod–Domar equilibrium condition $s = g(K/Y)$, where s is the aggregate saving ratio, g the natural rate of growth (which can include 'labour saving' technical progress), and K/Y the capital/output ratio. If these three variables are all given, then it is unlikely that the Harrod–Domar condition can be satisfied. Hence, in order to have a model in which the

possibility of steady growth is assured, it is necessary to relax one or another of the assumptions. The equality between s and $g(K/Y)$ can be obtained by: (a) flexibility in K/Y (also referred to as the technology assumption); (b) flexibility in s (saving assumption); (c) flexibility in g (labour-market and/or labour-supply assumption).

The above three cases can, of course, be combined in various ways as, for instance, in Samuelson and Modigliani's (1966) model, where (a) and (b) apply simultaneously.

Two Different Ways of Answering the Harrod–Domar Dilemma

Solution (a) above was adopted by the neoclassical or marginalist school:

> Instead of there being fixed coefficients in production there may exist a production function offering a continuum of alternative techniques, each involving different capital– labour ratios; ... The consequence is that the capital–output ratio v is adjustable, instead of being fixed, and this provides a way in which s/v and n may be brought into equality (Hahn and Matthews 1964, p. 785).

The second answer to the Harrod–Domar dilemma, that is, the assumption of a flexible aggregate saving ratio, was primarily adopted by the (neo-) Keynesian or Cambridge School. Of course there are many ways in which one can give flexibility to s ; but the one which has played the major role is the hypothesis of a two-class society (namely workers and capitalists, or consumers and entrepreneurs), each with a different (constant) propensity to save. In this way there always exists a distribution of income between the two classes which produces precisely that saving ratio that will equal the value $g(K/Y)$, so satisfying the Harrod–Domar equilibrium condition. The validity of this approach is reinforced by the fact that the assumption of a uniform aggregate saving ratio ignores all possible differences in saving (and consumption) behaviour between, for instance, different classes of income receivers, or categories of income or even different sectors of the economy. Moreover the problem of aggregating savings might give rise to particular and unknown difficulties, so that it may be safer to consider it in a disaggregate way, as the neo-Keynesian model does. Thirdly, this assumption also receives empirical support from the observed high rates of saving out of corporate profits and lower rates out of labour income. Considering a full-employment long-run equilibrium growth model with a capitalis' class (whose income is derived entirely from capital) and a workers' class (whose income is derived from wages and accumulated savings), both with constant propensities to save, the Cambridge economists were in a position to (1) provide a solution to the Harrod–Domar dilemma (by specifying an aggregate saving ratio s which equals $g(K/Y)$, where g and K/Y are both exogenously given); (2) determine the long-run equilibrium value of the rate of profits, the distribution of income between profits and wages, and the distribution of disposable income between the two classes; (3) allow the existence of an income residual, namely the wages, consistent with the assumption of a relationship between the savings of that class of individuals (the capitalists) who are in the position to control the process of production and the patterns of capital accumulation; and (4) give some insights into the process of accumulation of capital by specifying the equilibrium capital shares of the two classes. This range of results is obtained within a fairly simple framework and on the basis of relatively few assumptions, much less 'hybrid, opposite and extreme' than those of the marginalist model.

Kaldor's Theory of Distribution

Kaldor's distribution theory plays a fundamental role in the Cambridge or post-Keynesian theories of income distribution. His original analysis appeared first in the *Review of Economic Studies*, 1956 and, in a slightly different form in *Essays on Value and Distribution* in 1960. Kaldor considers a one-sector growing economy in which there are two classes: one whose income is derived entirely from capital (the capitalists, who are not wage-earners) and a second one which derives its income uniquely from wages (the workers). At each of these two groups he attaches a fixed

propensity to save, sc and sw respectively, higher for the capitalists and lower for the workers. Kaldor's model, as well as all other neo- or post-Keynesian models, is based on the assumption of long-run, full-employment equilibrium.

Assuming that national income (Y) is divided into wages (W) and profits (P) and a situation of steady growth, where all variables grow at the same rate g and where all ratios among macroeconomic variables remain constant, Kaldor derives explicit formulae for the overall rate of profits and share of profits in national income. Additionally, by making the 'classical' assumption that sw (the propensity to save of the workers) is zero, he obtains the following two simple relationships: $P/K = g/sc$ and $P/Y = gK/s_c$. The first solution shows that the equilibrium rate of profits depends only on the exogenously given rate of growth (g) and on the constant propensity to save of the capitalists' class. The second solution shows that the long-run share of profits in national income is determined by the rate of growth, the capital/output ratio (K) and the propensity to save of the capitalists (all exogenously given).

Pasinetti's Theorem

As we have seen, Kaldor's saving function considers, essentially, two types of income. What happens if we assume that the saving propensities differ, not according to classes of income, but according to classes of individuals (a more realistic assumption, referring to the weak definition of social class discussed above)? It is at this point that the basic contribution of Pasinetti may be brought in, where he assumes that saving propensities differ by class (assumed inter-generationally stable), rather than by type of income. His contribution, in his own words

> has come from the discovery of a fundamental relation (passed unnoticed in the whole of previous economic literature) which links profits to savings through the ownership of the capital stock. This relation simply follows from the institutional principle that profits are distributed in proportion to the ownership of capital and that the ownership of capital derives from accumulated savings (Pasinetti 1974, p. 127)

In this way the workers' class is allowed to own a share of the total capital stock, from which it derives an interest income. By solving the model Pasinetti obtains an explicit value for the rate of profits and share of profits in national income; in particular the former turns out to be $P/K = g/sc$, which has been defined as Pasinetti's Theorem, or 'Cambridge equation' (as a matter of fact it should be defined as the 'New Cambridge equation', since the original one had been found by Kaldor). Pasinetti's analytical results are similar to those obtained by Kaldor; there is, however, a fundamental difference, since Pasinetti's solutions have been obtained without making any assumption whatsoever on the propensity to save of the workers, which may assume positive values indeed. These results are undoubtedly of importance and establish a direct and simple relationship between the rate of profits and the rate of growth, through the interaction only of the capitalists' propensity to save. More precisely, the value of the rate of profits shows that on the long-run equilibrium growth path, the propensity to save of the workers, through influencing the distribution of income between capitalists and workers, does not influence the distribution of income between profits and wages.

Implications of the 'New Cambridge Equation'

The first thing that can be stressed is that the rate of profits and the share of profits in national income both vary inversely with sc. Hence, all other things being equal, the less the capitalists save and the greater is their return on capital (but with a smaller share of the capital stock). Exactly the opposite is true for the workers: the more they spend, the less they will receive for their future consumption (through a reduced share of the capital stock).

Secondly, as Pasinetti himself points out in the original exposition and more recently (Pasinetti 1974, Ch. VI) the irrelevance of workers' propensity to save gives the neo-Keynesian growth model much more generality than it appears at first sight. It is not necessary to make any hypothesis whatever on the aggregate saving behaviour

of the workers for the simple reason that both the rate of profits and the distribution of income are determined independently of the propensity to save of the workers. Therefore the workers could be divided into any number of sub-categories we wanted. Again, the particular saving behaviour of any sub-category of workers would influence the distribution of income among the various sub-classes of workers and, of course, between the workers and capitalists, but the distribution of income between wages and profits would not be affected at all, given the constancy of the capitalists' propensity to save.

Third and finally, the 'New Cambridge equation' shows and uncovers, for the first time in modern economic theories, the 'absolute strategic importance' of the saving behaviour of just one group of individuals (the capitalists) for the determination of the most vital relationships of the model. On the other hand the saving behaviour of the other class (or sub-classes) has nearly no power at all: they can save as much as they want, and of course receive an interest on it, but they will not influence the distribution of income between profits and wages. Moreover the share of wages in national income is a residual, once the share of profits (a function of the capital/output ratio, the rate of growth and the propensity to save of the capitalists) has been determined. The concept of residual of the classical economists is to be found again: but while for Ricardo the residual was represented by profits, in post-Keynesian models wages are a residual, once profits have been determined.

The Marginalists' Reply to the 'Cambridge Equation'

The results obtained by the neo-Keynesian economists did of course attract the attention of the neoclassical economists, who defined the Cambridge equation of income distribution as a 'paradox'. Their reaction was not surprising, since the Cambridge equation makes the whole 'well-behaved' production function framework irrelevant. With the aim of defending the theory of marginal productivity of capital, Meade (1963, 1966) and Samuelson and Modigliani (1966) set out to find a condition for which Kaldor-Pasinetti's Theorem would be prevented from operating, by arguing that when the propensity to save of the workers is exactly equal to the propensity to save of the capitalists times the profits share, then the capitalists cannot in equilibrium survive in the system and their propensity to save cannot determine the rate of profits. In such a situation all equilibrium savings of the system would be provided by the workers only, and the two-class system would become a single-class model where the marginalist scheme could be applied again to determine income distribution. But, as the ensuing debate has shown, such a situation is very unlikely to happen in the real world and, more importantly, it does represent a 'knife-edge' solution since in equilibrium it applies only when $s_w = s$. To devise one 'knife-edge' in order to answer another one (the Harrod–Domar's) may not represent the best counter-argument.

Other Criticisms of the 'Cambridge Equation'

The most common criticisms of post-Keynesian income distribution models (cf., for instance, Bliss 1975, Ch. 6; Samuelson and Modigliani 1966) seem to concentrate on: (a) the assumption of the equality, in the long-run, between the rate of profits earned by the capitalists and the rate of interest earned by the workers on their accumulated savings; (b) the constancy of the propensity to save of the two classes, exogenously given and hence independent of other variables as, for instance, the rate of interest or the rate of population growth; and (c) the assumption and identification of individuals who retain their class identity forever, that is, of classes which are intergenerationally stable.

Let us consider these points in some detail. In the late Sixties and early Seventies several authors have suggested that if one were to assume a differentiated rate of return for workers' and capitalists' savings, the Cambridge equation would no longer apply. As a matter of fact, as Pasinetti

(1974, pp. 139–41) has formally proved, this may not be true: the hypothesis of a differentiated interest rate comes to reinforce his analysis, since 'A rate of interest lower than the rate of profit has the same effect of a higher propensity to save of the capitalists, as it redistributes income in favour of the class that owns the physical capital stock.' The second criticism was put forward by economists who thought that the introduction of the life-cycle hypothesis on savings into the two-class model (where individuals would make their saving plans on the basis of the level of the interest rate and of other life-cycle parameters) would make the equilibrium interest rate a function of all parameters of the model. The assumption of the life-cycle hypothesis is of course not strictly compatible with the neo-Keynesian framework, where investment is independent of savings; nonetheless it has been shown that even in the context of a two-class life-cycle model, as long as there exists a class of 'pure' capitalists, the equilibrium interest rate is a function of the behavioural parameters of the capitalists only. The third main criticism concerns the assumption of intergenerationally stable classes; one would expect that the relaxation of this assumption would invalidate the relevance of the Cambridge theorem. But it is not really so, as few authors seem to conclude.

Vaughan (1979), for instance, in his analysis obtains a third solution for the interest rate, which approaches Pasinetti's solution when the net transference of individuals between classes is low, as it may be the case over the very long run which constitutes the framework of these models.

Conclusions

Summing up we may say that post-Keynesian theories place themselves half-way between classical and marginalist theories of income distribution, since on the one hand they reject the strong version of the social-class theory of distribution postulated by classical economists, where each income share meets a strong 'claim' associated with the property of an essential input (labour, capital or land). Instead post-Keynesian theories put forward a much more flexible concept of social class, characterized by a given saving and consumption behaviour (for Pasinetti the workers may even be divided into sub-classes). But on the other hand post-Keynesian theories differentiate themselves from the models of competitive economics where individuals react only with respect to the markets on which they have little effect. Post-Keynesian theories do moreover allow for elements of monopoly power, and retain the concepts of residual income and circularity of the production process contemplated by classical economists. It may well be that their extension to include certain elements of the life-cycle theory of saving and consumption behaviour will give them some micro-foundations.

See Also

- ▶ Kaldor, Nicholas (1908–1986)
- ▶ Widow's Cruse

References

Baranzini, M. 1975. The Pasinetti and the anti-Pasinetti theorems: A reconciliation. *Oxford Economic Papers* 27: 470–473.

Baranzini, M., and R. Scazzieri. 1986. Knowledge in economics: A framework. In *Foundations of economics: Structures of inquiry and economic theory*, ed. M. Baranzini, and R. Scazzieri. Oxford/New York: Basil Blackwell.

Bliss, C.J. 1975. *Capital theory and the distribution of income*. Amsterdam: North-Holland.

Hahn, F.H., and R.C.O. Matthews. 1964. The theory of economic growth: A survey. *Economic Journal* 74: 779–902.

Harcourt, G.C. 1972. *Some Cambridge controversies in the theory of capital*. Cambridge: Cambridge University Press.

Kaldor, N. 1956. Alternative theories of distribution. *Review of Economic Studies* 23(2): 83–100.

Kaldor, N. 1960. *Essays on value and distribution*. London: Duckworth.

Meade, J.E. 1963. The rate of profit in a growing economy. *Economic Journal* 73: 665–674.

Meade, J.E. 1966. The outcome of the Pasinetti process: A note. *Economic Journal* 76: 161–165.

Pasinetti, L.L. 1962. The rate of profit and income distribution in relation to the rate of economic growth. *Review of Economic Studies* 29: 267–279.

Pasinetti, L.L. 1974. *Growth and income distribution, essays in economic theory*. Cambridge: Cambridge University Press.
Pasinetti, L.L. 1981. *Structural change and economic growth. A theoretical essay on the dynamics of the wealth of nations*. Cambridge: Cambridge University Press.
Robinson, J. 1962. *Essays in the theory of economic growth*. London: Macmillan.
Samuelson, P.A., and F. Modigliani. 1966. The Pasinetti paradox in neoclassical and more general models. *Review of Economic Studies* 33: 269–302.
Vaughan, R.N. 1979. Class behaviour and the distribution of wealth. *Review of Economic Studies* 46: 447–465.

Distribution Theories: Marxian

David M. Gordon

It is hard to imagine a more important topic within Marxian economics than the distribution of income and the means of production among the principal classes in capitalist economies. For example: (1) The share of profits (or, inversely, the share of wages) constitutes one important component of the rate of profit. (2) The rate of profit operates as a fundamental determinant of the pace of investment and, therefore, of accumulation. (3) The rate of accumulation serves as a kind of life-force invigorating capitalist economies over time – regulating their growth and development, and the wealth of their participants. (4) Distribution, production and accumulation are thus fundamentally interconnected, forming the foundation of lives and livelihoods in capitalist societies.

In this respect, indeed, Marx himself regarded 'distribution relations' as part of the core of the capitalist economy. Criticizing those who ventured an 'initial, but still handicapped, criticism of bourgeois economy' by seeking to distinguish between the level of priority of production and distribution, Marx affirmed that both production relations and distribution relations are part of the 'material foundations and social forms' of any given historical epoch. Distribution relations and production relations are 'essentially coincident', he argued, since 'both share the same historically transitory character'. (Marx 1894, pp. 883, 878).

And yet, despite these reasonably self-evident theoretical connections, the analysis of distribution has remained substantially underdeveloped in the historical evolution of Marxian economics. While such classic issues as crisis theory, the transformation problem and the usefulness of the labour theory of value have been intensively and vigorously reviewed, the determination of distribution patterns over time and cross-sectionally has been elided in synthetic treatments of Marxian analytics and largely ignored in more focused scholarly investigations.

More recent developments in Marxian economics, fortunately, have finally begun to overcome this traditional reticence. This essay provides a brief review of traditional attention – or, more accurately, *in*attention – to the problem of distribution and then surveys some promising recent cultivations of this historically fallow terrain.

Terms of Analysis

Before beginning that review, however, it will be useful to clarify the defining boundaries of this topic.

It is probably most useful to begin with the role of distribution in the determination of profitability, that central fulcrum of economic behaviour. A familiar accounting identity reminds us that the rate of profit of the individual firm, r, can be expressed as the product of the share of profits in firm value-added, s_r, the ratio of output to utilized capital stock, y_u, and the ratio of utilized to owned capital stock, K^*, or

$$r \equiv s_r \cdot y_u \cdot k^*, \tag{1}$$

where

$$r \equiv \Pi / K_0;\ s_r \equiv \Pi / y;\ y_u = Y / K_u;\ k^* \equiv K_u / K_0; \tag{2}$$

and Π is firm profits, K_0 is the value of the firm's owned capital stock, Y is firm value-added, and K_u

is the portion of the owned capital stock which is currently utilized. In the aggregate, abstracting from variation among firms for such purposes, the same accounting identity applies.

In this accounting identity, distribution relations primarily affect the level of and changes in s_r, the share of profits in firm revenue. Factors affecting the rate of capital accumulation and the productivity of the means of production primarily affect y_u. Secular trends in the robustness of aggregate demand and its fluctuations over the business cycle have their most direct impact on k^*.

At this first level of approximation, then, analysis of distribution relations among the principal classes of a capitalist economy can begin with a focus on the determinants of s_r. Such analyses would immediately concern themselves with the wage share, s_w, as well, since $s_w \equiv (1 - s_r)$.

This is, of course, only a first level of approximation. At a second level of investigation, we must deal with three further refinements of focus.

1. Accounting Eq. 1 is formulated in revenue terms, not in value terms, so it does not yet encompass the Marxian concern with the value-theoretic determinations of economic relations. But this additional consideration requires simply that we add an analysis of the *rate of exploitation* (or the rate of surplus value), ε, to the definition of our task, since conventional Marxian value analytics establish a straightforward transformation between the profit share and the rate of exploitation. In one simple formulation, for example, the rate of exploitation is equal to the ratio of profits ($\prod$) to wages (W) weighted by the capital–labour ratio (k_1), or $\varepsilon \equiv k_L \cdot (\Pi\ W)$. (See Marglin 1984, pp. 57–60 and 191–192, for a useful elaboration of these relations of equivalence.)
2. The first level of approximation, represented by Eqs. 1 and 2, also allows for the existence of only two classes in capitalist economies, abstracting from all other relevant economic groupings or subsidiary classes. At a second level of approximation, therefore, we must also consider the existence of and determination of the shares of any other categories of economic agents, beyond our starting groups of capitalists and workers, which may seem relevant or necessary for our analyses.
3. A share of revenue need not necessarily translate into an exactly equivalent share of real income, since the prices confronting workers and capitalists may not exactly parallel each other over time. The relative purchasing power of their revenues received, and therefore the distribution of income, may consequently vary as a result of changes in the relative prices of capital goods and wage goods as well. It is conceivably useful, therefore, to decompose the profit share in Eq. 1 into two terms, one involving a ratio of 'real' profits to real income and the other a ratio of capital-goods prices to an index of (weighted) output prices. (See Weisskopf 1979, for useful elaboration of this kind of decomposition.)

A final consideration seems critical for defining the scope of our analysis. It is taken for granted within the Marxian tradition that a given class's share of revenues is conditioned, at the most basic level, by the extent of its power over the means of production. And yet, over time, a given class's relative control of the means of production will be responsive to systematic changes in its share of revenues. It is not at all inappropriate, therefore, to treat the class distribution of revenues and the class distribution of control over the means of production as interdependent and mutually-determining over the long term. We may therefore define our task most broadly, in this respect, as *the analysis of the determination of class (and group) shares of revenue (and therefore of income) and of the class distribution of relative control over the means of production.*

Marx was himself clear on the importance of defining the analysis of distribution in both of these two senses. 'It may be said . . .' he wrote at the end of Volume III of *Capital* (1894, p. 879), 'that capital itself . . . already presupposes a distribution: the expropriation of the labourer from the conditions of labour [and] the concentration of these conditions in the hands of a minority of individuals . . .' This underlying dimension of distribution 'differs altogether', he continued, 'from

what is understood by distribution relations . . . [as] the various titles to that portion of the product which goes into individual consumption'. This does not in any way suggest, he insisted, that distribution in this former sense does not involve 'distribution relations' or should somehow remain peripheral to our analysis:

> The aforementioned distribution relations, on the contrary, are the basis of special social functions performed within the production relations by certain of their agents . . . . They imbue the conditions of production themselves and their representatives with a specific social quality. They determine the entire character and the entire movement of production.

Traditional Analysis

Inherited approaches to the problem of distribution are most easily viewed through three somewhat separable lenses: the growth-theoretic perspective, crisis-theoretic hypotheses of a rising profit share, and antipodal crisis theories based on a falling profit share.

Long-Term Trajectories

Marxian economics has not always found it congenial to reflect upon the long-term growth paths of capitalist economies, since such perspectives are tainted in some minds by associations with concepts like 'stability' and 'equilibrium'. It is nonetheless possible to extract from traditional Marxian analyses a clear approach to the logic of determination of 'steady-state' tendencies – provided this exercise is understood, in Marglin's words (1984, p. 52), 'as a subset of Marxian theory and not as an attempt to represent the whole'.

It seems reasonably clear, in that context, that distribution relations are exogenously given to the traditional model, determined *outside* the set of basic interactions which jointly establish 'equilibrium' rates of growth and rates of profit. Historical conditions, not directly subject to internal economic analysis, establish a 'customary' wage. Existing levels of productiveness, also exogenous to the system, determine the level of output per hour and therefore, given the wage, the profit share as a residual. The behavioural hypothesis that capitalists save all profits combines with the determination of consumption by customary wage levels to create the conditions for a feasible and warranted steady-state combination of profit rates and growth rates. Marglin concludes (1984, p. 62): 'In contrast with the inherited neoclassical approach, in which resource allocation determines income distribution, causality here runs from [exogenously-determined] distribution to growth.'

There is, of course, nothing intrinsically wrong with these assumptions about directions of causality. Treating distribution as exogenous to the internal operations of the capitalist economy has simply meant that Marxian economists have tended to elide the factors determining distribution, setting them aside as consequences of 'historical and moral elements' and the 'technical' conditions of production.

Hypotheses of a Rising Profit Share

Distribution has played a somewhat more explicit role in analyses of tendencies toward economic crisis. One group of theories has built upon hypotheses of a secular tendency toward an increasing profit share.

Perhaps the first systematic example of this hypothesis emerges in Lenin's account of imperialism and monopoly capitalism (1917). In its essence, Lenin's argument begins with the relatively simple hypothesis of increasing oligopoly and therefore, 'since monopoly prices are established' (p. 241), of relatively reduced competitive pressures. With the help of financial oligarchies, corporations are able to achieve a continuously rising profit share and therefore to amass 'an enormous "surplus of capital"' (p 212). With this surplus of capital, capitalists are prompted to export capital overseas and, eventually, to reduce efforts at technical improvements. Over time, 'the *tendency* to stagnation and decay, which is characteristic of monopoly, continues to operate . . .' (p. 241; emphasis in the original).

The model begins, therefore, with a strong hypothesis about distribution – presuming a

strong initial tendency under monopoly capitalism towards a rising profit share. And yet, the conditions which would be necessary to derive this as a prevailing long-term tendency are unexplored. There is no real analysis of wages, although prevailing assumptions about competitive labour markets are implicitly incorporated into the model. There is equal taciturnity about the initial determination of real productivity, even though the rate of growth of real productivity must exceed the growth of real wages for the initial condition of a rising profit share and an ultimate 'surplus of capital' to hold. And, despite the international orientation of the analysis, there is no real incorporation of a model of international pricing and exchange which would support the hypothesis of rising profit shares in all the advanced countries.

These elisions are subsequently reproduced in most 20th-century analyses of underconsumption and monopoly capital. The models begin with a premise of growing capitalist power, most frequently from increasing monopoly control over product markets. This power leads to a rising 'surplus' and therefore to a rising profit share. From that set of initial premises, the problems of effective demand and urgent efforts to absorb the surplus follow naturally (Bleaney 1976; Baran and Sweezy 1966). As with Lenin, however, there is remarkably little attention to the conditions which permit this initial increase in the profit share. What about wages? Or labour productivity? Or conditions of international pricing? There is, in general, the simple presumption that conditions have evolved in a such a way as to permit consistent increases in the profit share, but little reflection on the relations which make those conditions possible. Baran and Sweezy admit some of this inattention, particularly to the social relations which would allow real productivity growth to outstrip real wage growth (1966, pp. 8–9):

> We do not claim that directing attention to the generation and absorption of surplus gives a complete picture of this or any other society. And we are particularly conscious of the fact that this approach, as we have used it, has resulted in almost total neglect of a subject which occupies a central place in Marx's study of capitalism: the labour process.

Hypotheses of a Falling Profit Share

For completeness, it is useful to consider the alternative hypothesis of a falling profit share, although attention to this possibility has only emerged within Marxian analysis more recently, primarily in the post-World War II era.

This hypothesis has relatively simple analytic foundations. For whatever reasons, working-class power may increase sufficiently to allow wages to rise more rapidly than labour productivity and therefore to result in a persistent increase in the wage share of revenues.

The hypothesis follows most naturally in a cyclical context and bears close connections to Marx's own analysis of cyclical dynamics in Chapter XXV of Vol. I of *Capital* (1867). In the short run, rapid expansion may lead to tight labour markets, increasing workers' bargaining power and resulting in a rising wage share. (Boddy and Crotty (1975) provide a useful development of this cyclical model in relatively traditional terms.)

The hypothesis needs further grounding in order to serve as the basis for a theory of economic crisis, however. The forces which lead to tight labour markets in short-term expansions could plausibly result in comparably slack labour markets during short-term contractions and therefore to a recovery of the profit share. In order properly to ground a theory of secular crisis upon this hypothesis of a falling profit share – and therefore fully to develop a 'profit squeeze' theory of economic crisis – one must show why cyclical contractions do not restore the profit share and, other things equal, the rate of profit. This requires analyses of conditions which permit rising worker power – even in the age of oligopolistic competition – from one business cycle to the next. Until the mid-1970s, Glyn and Sutcliffe (1972) were the principal Marxian economists to have formally developed such an analysis, and in their case primarily for the sole case of England.

Even in their case, however, the analytic requirements for the secular version of the 'profit squeeze' theory of crisis are not fully developed. What are the explicit conditions of labour market competition which explain particular patterns of wage growth? Under what conditions in the

organization of production and the promotion of technical change would real productivity growth fail to keep pace with real wage growth? What are the conditions of international economic linkages which would or would not support tendencies towards a falling profit share? A further problem involves the closeness of the relationship between profits and surplus value; Shaikh (1978) reviews some of the problems with casual assumptions about this connection.

Kalecki and Mandel as Connecting Writers

We can find in the work of Michal Kalecki and Ernest Mandel some early instances of the kinds of concerns which have fuelled more recent explorations.

Particularly in his later essays, Kalecki identifies but does not yet develop some of the lines of inquiry which would be necessary for a more advanced analysis of distribution. In 'Class Struggle and Distribution of National Income' (1971), Kalecki refines the analysis of the relationship between wages and the profit share, noting that analyses of the conditions of product market competition are necessary 'to arrive at any reasonable conclusion on the impact of bargaining for wages on the distribution of income' (p. 159); that trade union power is likely, *ceteris paribus*, to reduce the level of the mark-up; and that, in general,

> class struggle as reflected in trade-union bargaining may affect the distribution of national income but in a much more sophisticated fashion than expressed by the crude doctrine: when wages are raised, profits fall pro tanto. (p. 163)

In 'Trend and Business Cycle' (1968), Kalecki develops what he regards as a more satisfactory analysis of the relationship between short-and longer-term determinants of investment and therefore, *a fortiori*, the conditions which are likely to affect movements in the profit share over time.

Both of these analyses are entirely preliminary, however, since they constitute more of a programme for further work than a report on completed analyses. In particular, Kalecki notes that most of his analysis hangs on a handful of coefficients which he takes as given for his purposes, including the level of labour productivity, the share of gross profits flowing into capitalist consumption, capitalists' propensities to invest, and the rate of embodied technical progress. 'To my mind', he concluded, 'future inquiry . . . should be directed . . . towards treating . . . the coefficients used in our equations . . . as slowly changing variables rooted in past development of the system' (p. 183). The real problem, in short, is not to assume the central parameters of the determination of profits and investment but rather to derive them from determinant structural and historical analysis.

Mandel serves as a transitional figure in a different way. Although much of Mandel's analysis is hard to pin down precisely, he has nonetheless helped highlight the importance of an integration between formal Marxian analytics and structural/historical analysis. In *Late Capitalism* (1972), in particular, he suggests the rich possibilities for analysis of the particular conditions which might or might not give rise to variations in the rate of surplus value. There is much to learn, he urged (p. 183):

> Late capitalism is a great school for the proletariat, teaching it to concern itself not only with the immediate apportionment of newly created value between wages and profits, but with all questions of economic policy and development, and particularly with all questions revolving on the organization of labour, the process of production and the exercise of political power.

Recent Explorations

As this review is being written, a rich range of Marxian work on distribution in advanced capitalist societies has recently been completed or is currently under way. Since much of it is still in progress and unpublished, full references are difficult and probably inappropriate for an enduring encyclopedia. This final section will therefore concentrate on a synthetic review of the kinds of explorations which have recently been undertaken and the promising possibilities which have begun to emerge.

Changing Power Relations

One central problem in traditional Marxian analysis, which the examples of Kalecki and Mandel as connecting figures help to highlight, was the reluctance to forge determinate linkages between formal analytic categories, on one side, and the structure of and changes in power relations, on the other. Many appear to have felt either that these two loci of investigation operated at different levels of logical abstraction or that power relations, with all the social complexity of phenomena like the class struggle, could not be rendered analytically or studied empirically in any kind of formal or rigorous fashion. One is left with analyses, to quote Harris (1978, p. 166), which remain 'essentially ad hoc and tentative'.

Recent work has begun to overcome these hesitations. It has pursued careful and analytically determinate investigations of the relationship between power relations and, among other variables, the profit share. Attention has been focused primarily on three different dimensions of power relations: capital–labour relations, global linkages, and contests over state policy and practice.

Capital–Labour Relations

It has been recognized since Marx that class struggle over wages could conceivably affect distribution. But the formal linkage of conditions of class struggle to the determination of wage and profit shares has been hampered by the impression that levels and rates of change of productivity are determined orthogonally – by technical conditions and the pace of investment – and therefore that the two kinds of concerns could not somehow be combined into a single, inclusive, determinate analysis of changes in the profit share itself.

This problem appears to have been overcome. In recent work, particularly by Weisskopf, Bowles and Gordon (1983), a 'social model of productivity growth' has formally linked factors affecting capital–labour relations with the more traditional analyses. Several hypotheses about factors affecting the level of labour intensity in production have been both elaborated mathematically and tested empirically. This 'social model' appears to provide a robust explanation of variations in rates of productivity in the United States in the decades following World War II.

One crucial insight in that work is also beginning to invigorate Marxian wage analysis. Traditional perspectives on wage determination, building upon the 'reserve army' effect, focused on the relationship between wage bargaining and the threat of unemployment. As capitalist societies have developed, however, the threat of unemployment has been tempered by the availability of various components of what is typically called the 'social wage' – such as unemployment insurance and income maintenance expenditures. This has prompted the development of a more inclusive measure of the threat to workers of job dismissal: an index of 'the cost of job loss'. It calculates the expected income loss resulting from job termination, usually calculated as a percentage of the expected annual income if still employed, and incorporates estimates of the average wage in employment, expected unemployment duration, available income-replacing benefits, and available non-income-replacing benefits (which workers receive whether employed or not). (For provisional definition and measurement, see Weisskopf et al. 1983). Building upon these insights, it is likely that we will soon see much more fully developed and sophisticated analyses both of the determinants of wage growth and of the relationship between wage growth and labour demand.

Taken together, these new hypotheses about wage change and productivity growth themselves combine to provide the possibility of much more advanced hypotheses about determinants of changes in the profit share. Given that it is formally true that the rate of change of the real profit share is equal to the rate of change of real productivity minus the rate of change of real wages, analytic determinations of changes in the class distribution of revenues can now properly reflect both 'social' and 'technical' determinations.

Global Power

As noted above, another elision in traditional Marxian analyses of distribution has involved international connections. Traditional analyses have either assumed perfect competition, an awkward first approximation, or have tended, following models of monopoly capitalism, to assume a constant or rising price mark-up. But in an open economy, neither assumption seems useful, even as a first approximation, because of the likelihood of secular changes in a given economy's relations with other suppliers and buyers in global markets. And these changes are quite likely to affect the distribution of revenues, since they are bound to affect either relative input prices or the mark-up and through either path potentially to influence the real profit share.

Analyses of international linkages have lagged behind studies of capital–labour relations, but some promising initial exploration are under way. Two principal avenues of approach seem to be emerging. One seeks explicitly to model the effects of changes in the level and variability of the terms of trade on domestic productivity and profitability. The other aims at understanding and eventually modelling the effects of changing conditions of international power and, in particular, the effects of the internationalization of capital and growing multinational corporate leverage over domestic labour. (Bluestone and Harrison (1982) provide a useful early account of some of these latter effects for the US.) This kind of work is still in its early stages but seems increasingly essential in a more and more interdependent economy.

State Policy and Practice

The state can obviously have important effects on the private distribution of income among classes, both through tax policies and through the effects of expenditures on the costs of production and the relative bargaining power of the respective classes. Work on these connections has not yet moved beyond its early stages. Gough (1979) reviews the paths of likely effect on both the tax and expenditure side. Bowles and Gintis (1982) provide one provisional study of the effects of state policies on the profit share in the United States. And some of the studies of capital–labour relations discussed above are beginning to shed important light on the effects of 'social wage' expenditures on private-sector wage and productivity determination.

Combined Effects

These three dimensions of power relations need not be quarantined in separate cells of analytic isolation. It is possible to derive an inclusive model of their combined effects which retains a focus on the power relationships incumbent in each. Bowles et al. (1986) provide one such model of the determination of the profit rate; it includes factors affecting labour intensity, relative international power, and relationships with the state. Applied econometrically, the model appears to provide the most robust account available of variations in the rate of profit in the US in the postwar era. Although the study focuses on the rate of profit as a dependent variable, its approach could also permit more focused analysis of the profit share as a potentially separable component of profitability.

Comparative Analysis

It seems equally important, finally, to advance our understanding of the factors which explain cross-sectional variations in the levels and time patterns of the class distribution of revenues and income. This task must inevitably come rather late in the game, since it largely presupposes the availability of existing models of distribution which work for at least one country or groups of countries on their own terms. At the time of writing, some promising initial studies of cross-national variations in the determination of profit rates and shares are just under way. The best existing review of the political economic history upon which such studies

must build is the excellent comparative analysis provided by Armstrong et al. (1984).

One, Two . . . Many Classes?

One final analytic task remains. Almost all recent studies of distribution have accepted the traditional preoccupation with a two-class model of capitalist economies – focusing almost exclusively on the single pair of opposing magnitudes, the profit share and the wage share. It is important at least to consider the possibility that a more variegated categorization of individuals would be fruitful, even for traditional Marxian problematics. What about managers? The petty bourgeoisie? Financiers? Different strata of the working class?

Empirical analyses aimed in this direction have lagged in large part because of continuing uncertainty and conflict over the appropriate definition of group boundaries and their inter-relationships. Two main approaches appear to have emerged as the principal lines of inquiry within the Marxian perspective.

One approach seeks to derive a more complex mapping of primary and 'intermediate' or 'subsumed' classes from the method and essential categories of traditional Marxian analysis. Sharp debates nearly overwhelmed these efforts in the mid-to late-1970s, but it is conceivable that a relatively widespread agreement on the terms of analysis may be emerging in the mid-to late-1980s. Almost all of these analyses presuppose the usefulness of a single category of 'productive workers' and seek to distinguish, as carefully as possible, among various groups of intermediate agents and non-productive workers whose incomes largely draw upon realized surplus value. Wright (1978) offers one useful early review of the possibilities and problems in this approach, while Resnick and Wolff (1985) present an interesting recent treatment.

A second approach, usually encompassed under the general heading of 'segmentation theory', has paid primary attention to the importance of various divisions within the working class. Different analyses of labour segmentation have emerged in studies of various countries, and it is not at all clear that a single uniform model of labour segmentation in advanced capitalist formations can or should emerge. These studies nonetheless suggest the promise and importance of studying (a) the effects of different structures of production and labour on the opportunities and realized incomes of individual members of the working class; and (b) the potential impact of systematically structured divisions within the working class on the wage share of the class as a whole. Gordon et al. (1982) provide one important analysis of segmentation for the United States; Wilkinson (1981) offers one useful early compilation of comparative studies; while Bowles and Gintis (1977) provide a formal analytic integration of segmentation analysis within the value-theoretic context of more traditional Marxian theory.

These two approaches are potentially complementary, not conflicting, since the former concentrates largely on the group distribution of realized surplus value while the latter primarily explores the group distribution of variable capital. They have not yet been properly vetted, compared, and integrated, however, so we still await a complete and satisfactory theoretical and empirical account of the distribution of revenues among all the relevant categories of individuals in capitalist economies.

See Also

- ▶ Marxian Value Analysis
- ▶ Surplus Approach to Value and Distribution
- ▶ Surplus Value

Bibliography

Armstrong, P., A. Glyn, and J. Harrison. 1984. *Capitalism since World War II*. London: Fontana.

Baran, P.A., and P.M. Sweezy. 1966. *Monopoly capitalism*. New York: Monthly Review Press.

Bleaney, M. 1976. *Underconsumption theories*. New York: International Publishers.

Bluestone, B., and B. Harrison. 1982. *The deindustrialization of America*. New York: Basic Books.

Boddy, R., and J. Crotty. 1975. Class conflict and macro policy: The political business cycle. *Review of Radical Political Economics*.
Bowles, S., and H. Gintis. 1977. The Marxian theory of value and heterogeneous labour: A critique and reformulation. *Cambridge Journal of Economics* 1(2): 173–192.
Bowles, S., and H. Gintis. 1982. The crisis of liberal democratic capitalism: The case of the U.S. *Politics and Society*.
Bowles, S., D.M. Gordon, and T.E. Weisskopf. 1986. Power and profits: The social structure of accumulation and the profitability of the postwar U.S. economy. *Review of Radical Political Economics*.
Glyn, A., and B. Sutcliffe. 1972. *British capitalism, workers and the profits squeeze*. Harmondsworth: Penguin.
Gordon, D.M., R. Edwards, and M. Reich. 1982. *Segmented work, divided workers*. New York: Cambridge University Press.
Gough, I. 1979. *The political economy of the welfare state*. London: Macmillan.
Harris, D.J. 1978. *Capital accumulation and income distribution*. Stanford: Stanford University Press.
Kalecki, M. 1968. Trend and business cycle. In *Selected essays on the dynamics of the capitalist economy, 1933–1970*, ed. M. Kalecki. Cambridge: Cambridge University Press, 1971.
Kalecki, M. 1971. Class struggle and distribution of national income. In *Selected essasys*, ed. M. Kalecki. op. cit.
Lenin, V.I. 1917. Imperialism, the highest stage of capitalism. In *Selected works*, one-volume ed. New York: International Publishers, 1971.
Mandel, E. 1972. *Late capitalism*, English ed. Trans. Joris De Bres, London: New Left Books, 1975.
Marglin, S. 1984. *Growth, distribution, and prices*. Cambridge, MA: Harvard University Press.
Marx, K. 1867. *Capital*, vol. I. New York: International Publishers, 1967.
Marx, K. 1894. *Capital*, vol. III. New York: International Publishers, 1967.
Resnick, S.A., and R.D. Wolff. 1985. A Marxian reconceptualization of income and its distribution. In *Rethinking Marxism*, ed. S.A. Resnick and R.D. Wolff. Brooklyn: Autonomedia.
Shaikh, A. 1978. An introduction to the history of crisis theories. In *U.S. capitalism in crisis*. New York: Union for Radical Political Economics.
Weisskopf, T.E. 1979. Marxian crisis theory and the rate of profit in the postwar U.S. economy. *Cambridge Journal of Economics* 3(4): 341–378.
Weisskopf, T.E., S. Bowles, and D.M. Gordon. 1983. Hearts and minds: A social model of U.S. productivity growth. *Brookings Papers on Economic Activity*, No. 2.
Wilkinson, F. (ed.). 1981. *The dynamics of labour market segmentation*. London: Academic Press.
Wright, E.O. 1978. *Class, crisis, and the state*. London: New Left Books.

Distribution Theories: Neoclassical

Christopher Bliss

Whenever a theory becomes involved in controversy the question of what constitutes that theory itself becomes a contentious issue, and the neoclassical theory of distribution is no exception to that general rule. Some have seen marginal productivity as an essential feature of neoclassical theory. Others have regarded the aggregation of capital or an aggregate production function (even a function of the Cobb–Douglas form) as essential. Neoclassical distribution theory is viewed as general equilibrium theory by many but Friedman has defended the 'Marshallian' or partial equilibrium approach.

The truth is that any body of ideas widely maintained for a long time inevitably develops and transforms itself, absorbs some ideas, discards others, and fathers traditions and sub-traditions. As the neoclassical theory of distribution has been the predominant view in the leading countries for the development of economics for over 100 years, it is not surprising that it conformed to this pattern and expressed itself in diverse even contradictory voices. Many, whether or not they like neoclassical theory, hold that one voice represents the true message, but neoclassical theory, like christian doctrine, may stand on certain fundamentals but is not and could not be monolithic.

It is important to distinguish between 'neoclassical theory' on the one hand and the history of the development of that theory on the other. Both are valid subjects for study but a scientific assessment of the theory should address itself to the best modern statements. This principle has not always been respected, particularly in the heat of controversy, and some maintain that the theory went wrong from the start, and that if one could only go back to where the vital mistakes were made everything would become clearer. (For an extensive development and discussion of this line of argument, see Baranzini and Scazzieri 1986.)

However, the development of economic theory is not like a complicated calculation in which every step is supported by every earlier step. As with any other discipline, the logical standing of a theory and the history of the development of that theory are distinct entities.

By way of illustration of the last point, consider the way in which the theory developed in its early stages. The 'neoclassical' movement, whose leading members may be taken to include Böhm-Bawerk, Edgeworth, Gossen, Jevons, Marshall, Menger, Walras, Wieser and Wicksell, did not begin with a theory of distribution but quite neglected that side of the economic problem. By focusing on marginal utility and the demand for given resources in a barter economy, the neoclassical economists were able to develop a powerful and flexible method, the marginal principle, so impressive that it has often been taken to define their approach. The so-called 'psychic' notion of marginal utility represented the refinement, no more, of the old idea of 'value in use'. However with its help the neoclassicals eventually succeeded in clarifying, as Smith, Ricardo and Mill had all failed to clarify, how value in use, value in exchange and cost of production could coexist. Only the Austrians with their concept of 'imputation' hung on to the idea that utilities were in some sense primary and other values derived.

Put in unashamedly modern terms, the central neoclassical idea is that the pricing of goods and the pricing of factors of production are governed by common principles, mainly the forces of supply and demand generated by agents who maximize their objectives. From the perspective of the history of the development of the theory the definition is anachronistic. Economics did not develop and refine the notion of a factor of production or the concept of maximizing an objective and later arrive at the neoclassical theory of distribution. Rather the two processes took place in tandem. Despite the lip service to classical ideas paid by some members of the neoclassical school, notably Marshall, neoclassical is a misnomer. The neoclassicals were not revivalists of classical economic ideas, an Oxford Movement of classical political economy. They were revolutionaries.

The Distribution of Income

The theories with which we are concerned are designed to explain the levels of payment to the various factors of production – rents, wage rates, and rates of profit – and by extension the shares of the various factors in the total product. That is to say that they are concerned with the *functional* distribution of income.

We shall not discuss the distribution of personal or household incomes, sometimes called the *size distribution*. The size distribution of household incomes takes the form of a function relating the level of income and the number of units receiving that income. It is true that given the distribution of the ownership of factors among units, strictly the quantities supplied to the various markets, and given also the rates at which those factors are remunerated, the size distribution may be derived. However, except in the short run, the interrelationship between the functional and size distributions is more complicated. This is mainly so because the quantities of factors which may be accumulated by individual units, land and capital, and even the quantity of labour, respond to rates of return to the various factors. Pasinetti (1962) presents a model which unusually takes this interrelationship into account. For a discussion of the Pasinetti model and some of the criticisms which it has attracted, see Marglin (1984, pp. 324–8). On the distribution of personal income and wealth, see Atkinson (1975).

Factors of Production

It is not surprising that the concept of a factor of production plays a leading role in neoclassical theory because it lends itself to the view that the inputs used in production stand to each other in a relation of symmetry, governed by common principles. This is not to say that no differences between the conditions applying to factors are admitted. The symmetry is most marked in the treatment of the demand for factors, while on the supply side important differences are recognized.

The membership of the trinity of land, labour and capital, which have always been taken to be factors of production, goes back to the classical writers, and an additional factor called 'entrepreneurship' is widely recognized by neoclassical and classical writers alike. The development of the theory along formal lines has tended until recently to suppress the role of the entrepreneur and to make the firm into a rather lifeless object. However lately the increasing employment of economic theory in industrial economics has given rise to some richer treatments of the firm.

The employment of the concept of a factor of production has been criticized. It has been argued that labour in particular does not submit itself to the laws of supply and demand like any other input. The introduction of distinctive features of the various factors and their markets tends to undermine the simple symmetry of pure theory. Some have detected apologetics in the designation of capital as a factor of production. On this see the discussion of ideology below.

Marginal Productivity and the Determination of Factor Prices

Do marginal productivities determine factor rewards? This apparently straightforward question conceals conceptual complications and, depending on the context to which the question is applied, either 'no' or 'yes' may be defended as reasonable answers. Robertson (1931) argued that the wage rate 'measures' the marginal productivity of labour. The reference is to the demand curve for labour, which is the schedule of the marginal productivities of various quantities of labour. Robertson was reminding his readers that the wage rate in a competitive market is determined by the intersection of the demand curve and the supply curve – both blades of the scissors cut the paper. If marginal productivities are values determined by the equilibrium solution as much as are wages and prices, talk of one determining the other is misplaced. The same point applies when the marginal product of capital and the return to capital are under consideration.

In certain contexts however it is reasonable to see marginal productivity as the determinant and the payment to the factor as determined. Consider the claim that managers of large enterprises are paid very large salaries because the marginal value productivity of a good manager amounts to a great deal of money. Supposing this argument correct, the high marginal productivity is a general feature which does not depend upon solving out the whole equilibrium. Contrast this with the case of a micro unit, say a farm, facing a given wage rate for labour and able to vary the quantity of labour employed. For that exercise the wage rate is given and the marginal product is determined by it.

A Simple Neoclassical Model

In this section we examine a static model. Growth and capital will be considered below. The idea is to construct a model in which factor prices will drive everything else, including goods prices through cost functions. This requires special assumptions but makes for a model which can be easily presented and which suffices to illustrate some points about the neoclassical model of distribution. For a much more thorough review of neoclassical models, see Ferguson (1969).

We assume factors and goods to be distinct and that factors are not directly consumed. Let there be F factors available in given quantities, and G goods producible from those factors, F and G need not be equal and there may be more goods than factors, or less or the same number. The production function for the ith good is:

$$v^i = f^i\left(x^i\right) \quad (i = 1, \ldots, G); \tag{1}$$

where v^i is the output of the i th good and x^i is a vector of factor inputs to the production of the ith good. $f^i(\,)$ is a concave constant returns production function. The cost function shows the unit cost of producing good i given factor prices. Factor prices are a vector w and the unit cost of the i th good is $C^i(w)$, where $C^i(w)$ is the solution to the programme:

$$\min_{x^i} wx^i; \tag{2}$$

subject to:

$$f^{i}\left(w^{i}\right) \geq 1. \tag{3}$$

We denote the prices of goods by vector $c(w)$, where the i th element of $c(w)$ is $C^{i}(w)$. There are H households. Let the h th household own factors x^{h}, in which case its income will be $w \cdot x^{h}$. All the household's income is assumed spent on goods and the vector of goods demanded by household h is denoted z^{h} and is given by the h th household's demand function:

$$z^{h} = z^{h}\left[c(w), w \cdot x^{h}\right] \quad (h = 1, \ldots, H). \tag{4}$$

Now note that factor prices w imply demands for factors as may be seen by the following line of reasoning. Given w, we have household incomes $w \cdot x^{h}$ and goods prices $c(w)$. Hence we have total demands for goods:

$$\sum_{h} z^{h}\left[c(w), w \cdot x^{h}\right] = z. \tag{5}$$

The amount of factor j used in the production of good i is the partial derivative of $C^{i}(w)$ with respect to w^{j}, denoted c^{i}_{j}. The matrix of these coefficients, denoted C, depends on w only. Hence demand for factors is $C \cdot z$, supply is $\sum x^{h}$, and we have shown that excess demands for factors are a function of factor prices.

To prove the existence of factor prices such that factor demands and supplies are equal (strictly such that there is excess demand for no factor), one has to establish the continuity of the relationship between factor prices and excess demands for factors, and then employ a fixed point theorem (see Arrow and Hahn, 1971, ch. 5).

We note some salient features of this model. First, prices of factors are determined by the supply and demand for those factors although demands for factors are derived demands depending on their employment to produce goods. Secondly, both the technology of production and tastes influence the solution for factor prices. Thirdly, factor prices measure the marginal products of factors, a property which is ensured by the process of cost minimization. However there is clearly no sense in which marginal products are prior to prices.

More and Less General Models

The model of the previous section is designed to illustrate the manner in which the determination of the distribution of income may be viewed as the outcome of a general equilibrium of supplies and demands for factors of production. The model is less general than the standard general equilibrium model. It exhibits, for example, constant returns to scale production functions, no joint production and no direct consumption of factor services. Also, goods are not used as inputs to the production of goods. The introduction of those features would undermine the model's neatness without introducing fundamentally new principles.

More striking results are produced when the model is made still more specialised. The factor input coefficients may be treated as constants independent of w. In this fixed coefficient case the marginal product of a factor in producing a good is undefined. In an extreme case there is only one factor, usually labour, with the result that relative goods prices are independent of demand. (For a discussion of this non-substitution result and its extension to an economy which uses fixed capital, see Bliss, 1975, ch. 11.) Models of this kind typically introduce the use of goods as intermediate inputs to the production of goods. However so long as there is no genuine joint production (the term genuine joint production is used to distinguish the production of final demands jointly from the notional joint production that arises when fixed capital goods are treated as one of the products of the productive process.), the inputs used to produce final output may all be reduced to the quantities of the factor incorporated in them.

The model of Sraffa (1960), sometimes known as the neo-Ricardian model, will be seen to be a version of this model, but including an elegant extension to fixed capital goods. Hahn (1982) has argued against the claim that the neo-Ricardian approach leads to new insights by

pointing out that the model is a special case of the general equilibrium model.

The Problem with Capital

The introduction of capital into the theory of distribution raises two issues which should be distinguished, even though they are not entirely unrelated. One is the aggregation of capital, the other is the nature of the supply of capital in the long run.

Although many expositions of the theory have been expressed in terms of an aggregate called capital, and there have even been attempts to formally underwrite this approach, it is now generally recognized that there is no rigorous method of aggregating a heterogeneous collection of capital goods. (The most famous attack on the use of aggregate capital is Robinson, 1953–4); see also Champernowne, 1953–4; Harcourt, 1972 and Marglin, 1984, ch. 12.) In this respect capital stands on a par with other types of input, labour for example. Highly aggregated models should therefore be seen as simple devices for illustrating how a type of model functions and not as descriptions of the world. Unfortunately, some writers who emphasize the problems of aggregating capital are quite cavalier when it comes to discussing the aggregation of labour or output. Formally however there is little difference between the cases.

With many distinct capital goods, demands for inputs are demands for the services of particular capital goods. However the supply of capital in the long run is the supply of saving, which may translate itself as required into particular capital services. Hence a long-run neoclassical theory of distribution depends on a model of long-run saving, a point which deserves emphasis.

We show how the solution for the quantities of capital goods and equilibrium prices may be obtained in a simple constant returns growth model. Let there be N goods, and let the quantities of them which make up the capital stock used by one unit of labour be represented by the elements of a vector x. Let consumption be proportional to a vector c_o, and γ the rate of growth of the labour force. Let y be the total stock of goods available next period for consumption and as inputs to next period's production. The production function corresponding to a unit labour input is:

$$F(y,x) = 0. \tag{6}$$

In steady state growth with a per capita consumption of αc_o, y will be $\alpha c_o + (1+\gamma)x$. Hence:

$$F[\alpha c_o + (1+\gamma)x, x] = 0 \tag{7}$$

Given a particular per capita consumption αc_o, (7) may be satisfied by various values of x, but only one of these will be the efficient and equilibrium value. To see this let $V(x^1)$ be the maximum value of β such that βc_o is a sustainable per capita consumption starting with a capital stock x^1. If x^1 is the steady state composition of the capital stock for consumption βc_o then $x^2{=}x^1$ must solve:

$$\max_{x^2} \alpha \tag{8}$$

subject to:

$$F\left[(1+\lambda)x^2 + \alpha c_o, x^1\right] \geq 0; \tag{9}$$

and

$$\mathrm{V}\left(x^2\right) \geq \beta c_o. \tag{10}$$

Let the Lagrange multipliers attaching to the constraints (9) and (10) be respectively μ and η and let $F_i(i = y, x)$ denote the vector of partial derivatives of F with respect to the output and input vectors. The necessary conditions for a solution to (8)–(10) are:

$$1 + \mu c_o \cdot F_y = 0; \tag{11}$$

and

$$\mu F_y(1+\lambda) + \eta V_x = 0. \tag{12}$$

Equation (12) states that the marginal rates of substitution between outputs of the various goods shall be equal to the marginal rates of substitution

between those same goods as inputs to the long-term provision of future consumption; compare Dorfman et al. (1958, ch. 12). This condition reduces the degrees of freedom enjoyed by the steady state capital stock to one. That last degree of freedom depends on the level of steady state consumption the determination of which requires a saving condition.

Theory and Ideology

According to its Marxist critics, neoclassical distribution theory is irredeemably apologetic in character, and it is indeed the case that some economists in the past saw the theory, and in particular the concept of marginal productivity, as throwing a relatively favourable light on capitalism. When the justification for the earnings of capital owning rentiers was being questioned, the notion that capital earns no more than its 'contribution' to production was not unwelcome in the salons. The idea that the rich are rewarded according to the marginal productivity of their 'waiting' sounded better still.

It can need a positive effort to see that all this is strictly irrelevant to the scientific standing of neoclassical theory. No one supposes that Newton's mechanics should be dismissed because its author saw in it the justification of a hierarchical organization of social life. A play on the overtones of words such as 'earning' or 'waiting' to justify the distribution of income should be similarly disregarded. Of course the neoclassical theory of distribution can be used to analyse the effects of policy, including policies to redistribute income. In a perfect world the conclusions which emerged from such investigations would be independent of the political stance of the investigator. We do not live in a perfect world, but the fact that the scientific ideal is never fully attainable should not lead us to conclude that economics can know nothing but self-serving apologetics.

See Also

- ▶ Adding-up Problem
- ▶ Clark, John Bates (1847–1938)
- ▶ Marginal Productivity Theory
- ▶ Wicksteed, Philip Henry (1844–1927)

References

Arrow, K.J., and F.H. Hahn. 1971. *General Competitive Analysis*. Amsterdam: North-Holland.

Atkinson, A.B.. 1975. *The economics of inequality*. Oxford: Clarendon Press.

Baranzini, M., and R. Scazzieri. 1986. *The foundations of economic knowledge*. Oxford: Basil Blackwell.

Bliss, C.J. 1975. *Capital theory and the distribution of income*. Amsterdam: North-Holland.

Champernowne, D.G. 1953. 4The production function and the theory of capital: A comment. *Review of Economic Studies* 21(2): 112–135.

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

Ferguson, C.E. 1969. *The neoclassical theory of production and distribution*. Cambridge: Cambridge University Press.

Hahn, F.H. 1982. The neo-Ricardians. *Cambridge Journal of Economics* 6(4): 353–374.

Harcourt, G.C. 1972. *Some Cambridge controversies in the theory of capital*. Cambridge: Cambridge University Press.

Marglin, S.A. 1984. *Growth distribution and prices*. Cambridge, MA: Harvard University Press.

Pasinetti, L.L. 1962. Rate of profit and income distribution in relation to the rate of economic growth. *Review of Economic Studies* 29: 267–279.

Robertson, D.H. 1931. Wage grumbles. In *Economic fragments*, ed. D.H. Robertson. London: P.S. King & Son.

Robinson, J.V. 1953. 4The production function and the theory of capital. *Review of Economic Studies* 21(2): 81–106.

Sraffa, P. 1960. *The production of commodities by means of commodities*. Cambridge: Cambridge University Press.

Distribution, Ethics of

J. B. Clark

The primary fact of economics is the production of wealth. The division of the product among those who create it is secondary in logical order and, in a sense, in importance. Yet the most important subject of thought connected with social

economy is distribution. If the term be used broadly enough it designates all of the economic process that presents moral problems for solution. On the settlement of the ethical questions concerning the division of the social income depends not only the peace of society but the fruitfulness of industry. It is a striking fact that Ricardo, whose studies carried economic science forward in the direction of the truth concerning distribution, but stopped short of that goal, and so strengthened the hands of social agitators, realized the paramount importance of the subject on which his thought was chiefly concentrated: 'To determine the laws which regulate this distribution', he says in his preface, 'is the principal problem in political economy.'

Scientific errors concerning the law of distribution react more harmfully on production than do errors of doctrine concerning production itself. Among self-asserting people, industry loses fruitfulness whenever the belief is widely diffused that products are shared according to an unjust principle. If it were a general conviction that social evolution is in the direction of iniquity – that distribution already robs the workers and will rob them more hereafter – no force could prevent a violent overturning of the social order.

Industry has its fruits and its sacrifices; it creates useful things at the cost of working and waiting. Where production is carried on in a collective way, both the products and the burdens of the process have to shared by different classes of men according to some principle. The apportionment that has to be made is not only of products, which represent positive values, but of sacrifices, which may be treated as negative values of a 'subjective' kind. While the term distribution, as currently used, designates only the apportionment of the positive values, or products, it is capable of being used in a more complete sense, and made to include the apportionment of the negative ones also. It would then include all of economic science that involves moral problems.

Both parts of this twofold distributive process must in any case be studied if the ethical questions connected with industry are to be solved. There is no independent standard of justice in the distribution of products only. What a man ought to get out of the collective income of mankind depends on how much he or some one who represents him has sacrificed in helping to create it. The apportionment of the positive values referred to is inseparably connected with that of the negative values. Political economy must tell us how both products and burdens are actually shared, and ethics must tell us how both of them ought to be shared, if the existing plan of social industry is to be morally tested.

Political economy has not as yet furnished a theory of the actual distribution of positive values, or products of industry, that has met with general acceptance. It has scarcely attempted to furnish a theory of the distribution of the negative values. Ethical science has not furnished a clear standard of justice in the double apportionment.

Every producer experiences in his own person the double effect of industry; he is first burdened and then rewarded. The net effect of the two influences on the man's well-being may be termed the subjective resultant of production. A complete science of distribution must study the economic resultants in the case of different classes of men. How is a labourer on the whole affected by industry? What is the measure of the net benefit that comes to him from this source? How is a capitalist affected? How do the net effects compare with each other? What tendencies are at work to change the two, both absolutely and relatively? These are economic questions; while the ethical question is what the resultants in the two cases ought to be.

The personal resultant of industry is always a positive quantity. Work yields a net gain; the fruits of it are worth more than they cost. For the most hardly-used classes an industrial life is, by economic tests, more than worth living. The hours of labour in a day are increasingly burdensome as the period of work is prolonged. A man might labour three hours a day with little weariness and no injury. The eighth hour is wearying, and the tenth is more so. There comes a time at which work naturally stops, if the man is free, because working longer would cost more in the way of pain than it would secure in the way of pleasure. Final or marginal labour is that which just pays for the weariness that it costs. The gain that comes through labour offsets the burden that it entails at

the point in the working day at which the burden is greatest. The less onerous labour of the earlier hours affords a net personal gain. If the man is paid by the hour he earns a part of his wages very easily. Intramarginal labour, as we may term it, affords a net subjective gain, what some would call producers' rent.

Though the wages of all hours may be equal by money standards, they are of unequal utility to the man who gets them. His first earnings are spent on necessities, later ones on comforts, and final or marginal ones on things that figure in his estimate as luxuries. The last hour of his labour may ensure to him only the least important thing that he gets at all. It is the minimum benefit secured by an hour's labour that offsets the maximum sacrifice caused by it. There is therefore a second net gain coming to the worker in the spending of his money. As the sixpence or dime that is spent for a luxury benefits the man enough to offset the weariness of final or most fatiguing labour, those that are spent for food, clothing, etc., afford an additional benefit. The man enriches himself whenever he buys a loaf of bread. In general the sacrifices and the benefits of production just offset each other at the point at which the sacrifices are the greatest and the gains are the least. Everywhere except at the margin the gains are greater and the sacrifices are less.

Again the positive resultant of industry is increased by social organization. Anarchy, even if it were peaceful, would increase sacrifices and diminish rewards. Whatever might be true of a sparsely settled world, a crowded world is dependent on the multiplying of productive power that combination brings. All classes are debtors to society. No serious case can be made against the existing social order on the ground that it lessens the gain that labour naturally brings.

The indictments brought against the social order are based on the comparative treatment that society accords to men of different classes. Are the benefits conferred on different ones what they ought to be relatively? Does society proceed capriciously in the allotment of rewards and sacrifices? Do some classes fail to get the proportionate benefit that is properly theirs? Are social tendencies in the direction of equity or away from it? These are the ethical questions to be solved by a comparision of the ideally just distribution with the actual one.

Of the ideals of distribution that have been advanced none has been crude enough to provide for the apportionment of the products of industry and take no account of the burdens. A rule of equal rewards for unequal sacrifices would have no moral support. Ethical studies in this field really have as their object the attainment of a rule for adjusting what we have termed the personal resultants of industry, or a rule that, if followed in practice, would make the net effect of industry on the welfare of different classes equitable. Communistic theories make equality nearly synonymous with equity; but the thing that is to be equalized is seldom mere property or income. If the principle of equality be carried into refinements, so as to bring to one level the net benefits that society confers on all its members, the rule approaches, though it is still far from reaching, the ultimate moral ideal of distribution.

The better socialistic ideals are refinements of the rule of equality. In applying the rule to individuals, inheritance is the first disturbing influence encountered. The law of inheritance is based on a certain solidarity of families. Where it is in force the sacrifices of a parent may accrue to the benefit of a child. What we have termed the resultant of industry in the case of the heir to an estate is not to be measured by adding together positive values, represented by the enjoyments that the property brings, with negative values, represented by the inheritor's own sacrifices. If he be considered apart from his family the values in the case are nearly all positive. A crude levelling of individuals' net gains accruing from industry demands the abolition not only of inheritance, but of gifts from parents to children. Where it is advocated it is in the interest of purely individualistic equality.

The handing over of all capital to the state sweeps away even more completely inequalities of wealth in permanent possession. In theory it might avoid the evil connected with the abolition of inheritance, that, namely of reducing the capital that is necessary if wages are to be sustained at a high rate; since it is conceivable that the state itself might accumulate capital with needed rapidity.

This measure also would, in effect, disregard the solidarity of families and tend to put men on a footing of individualistic equality.

Economic difficulties do not need to be considered in the shaping of a moral ideal. The vesting of all capital in the state would save the student of applied ethics one serious difficulty, that, namely, of determining whether the sacrifice of abstinence is unduly rewarded as compared with that of labour, or, in other words, whether interest is too high as compared with wages. A socialistic state has its moral duty simplified, since it has only to reward different kinds of labour equitably.

A scheme that is too crude to have much support makes the wages and the working hours equal for all. Estimate the wages in money or its equivalent, gauge labour by time only, and bring both to an equality in the case of the whole adult population. Even the rewards are not thus in reality equalized, and the sacrifices are very unequal. In real rewards unmarried men would be favoured and large families would suffer. The real sacrifices incurred would vary according to the nature of the work performed.

An improvement on this scheme provides a stipend for each dependent member of a family, and tries to equalize sacrifices by so reducing the number of hours of labour per day in occupations that are disagreeable or hurtful, as to bring all employments to a certain uniformity of burdensomeness. In the case of very disagreeable work the hours would be reduced to a minimum, while in occupations that are less and less repellent they would be shortened proportionately less. Production would of course suffer by this arrangement, and the ideal that the plan of division presents is that of small but equal pay, with easy work, for all.

Another scheme does not content itself with equalizing what we have termed the personal resultants of industry, but aims to level inequalities of condition that lie at the back of industry itself. Society should do more for the lame and the blind than for those who have all faculties in possession, in order that the ultimate condition of all may be made as nearly equal as is possible. Here is the levelling policy in perhaps its most ambitious mood. It is not the treatment of men by society that is to be equalized, but the treatment of them both by nature and society. The industrial organism is to deal with its members unequally in order that it may somewhat neutralize the partiality of nature.

A rule of division that is often regarded as ethically lower than either of those above specified is that of compensation according to actual production. Give to a man the wealth that he creates, neither more nor less. Every one owns what he brings into existence; let not society wrest or filch from him any part of it. Let it keep itself clear from robbery and fraud.

If workers lived side by side in peaceful anarchy, with no division of labour and no exchanges, each man would get what he created. He would get little, but he would get all that would be his own. Introduce now a social union that multiplies products ten-fold but increases some men's returns only five-fold, and you seem to benefit these men and to rob them at the same time. If in organized industry some of the product that is distinctly attributable to labour itself finds its way into the hands of men who do not create it, the labourer suffers a wrong, even though the share that he still keeps may be larger by reason of the fact of his connection with the men who rob him. Such is the conception of industrial society that exists in many minds. The socialistic indictment against society is that it filches from workers a part of their share of the *extra* product of industry due to organization. Does society, under natural law, take from labour a product that is distinctly attributable to it? This is one of the most important questions in economics. A successful analysis of social production answers it. What needs to be known is what part of the composite result of industry is distinctly due to labour itself. In a land peopled by isolated producers and managing to live in peace, each man would get his own; does exchange vitiate this result? If so, organization proceeds here on an unusual principle; since the complications of society as a rule disguise essential facts of primitive industry, but do not annual them. The presumption is that the man who got his own when he worked alone gets it when he trades with his neighbour on terms of genuine freedom, and that a true analysis

of social relations will show the fact. If so, society tends actually to conform to the rule 'to every man the product that is distinctly attributable to the sacrifices that he or others in his interest have made'. There is common honesty in the distribution that takes place under natural law.

The literature of the subject of economic ethics is not as scanty as it is one-sided. The basis of the socialistic movement is ethical, and much of the literature is designed to prove that society is organized on a plan that systematically wrongs workers in the apportionment of the social income. A defence would naturally aim to show that the law of distribution is not itself iniquitous, however many particular cases of injustice might arise under it. A weak point in the defence is the lack of a clear demonstration of the complete nature of the actual law of distribution, a lack that, as is hoped, may soon be supplied. In the meanwhile statistics are appealed to on both sides to prove, on the one hand, that the actual apportionment of wealth is departing more and more from the ideal standard, and, on the other, that it is tending towards it.

Reprinted from *Palgrave's Dictionary of Political Economy.*

Distribution, Law Of

J. B. Clark

The most important share of the income of society is the one falling to labour. The so-called 'wage fund' theory accounted for the rate at which labourers are paid on the ground that wages come from a fund of capital devoted to the fund and that the rate per man depends on the size of the fund and the number of the claimants. The discovery of the fact that wages come from the product of industry, and not from capital, has made a new theory necessary, and has opened the way to the discovery of a general law of distribution.

The parties in the division of the general product of industry are – (1) those who contribute to production the element labour; (2) those who contribute instruments, or wealth in productive forms; and (3) those who bring labour and productive wealth into co-ordination by hiring both of these agents, and receiving and selling their products. The labour furnished includes the work of management, as well as other kinds of industrial effort; and the productive wealth, as the term is here used, includes land as well as other instruments. The co-ordinating function is, in this enumeration, kept distinct from the other two; the man who performs it is not to be treated in this connection as a labourer or as a capitalist, but as the employer of both labour and capital.

The shares to be accounted for are thus wages, interest, and pure profit, and these shares will include the rent of land and the wages of superintendence. The generic varieties of gain come from putting forth productive effort of some kind, from furnishing productive wealth in some form, and from bringing the effort and the wealth into coordination.

The scientific law of distribution determines what reward shall attach to the performing of one of these functions. It does not gauge the income of a particular man, since a man nearly always performs more than one function. A capitalist usually works, a labourer usually has capital, and an *entrepreneur*, or coordinator of labour and capital, almost invariably owns some productive wealth, and does some directive work. A scientific study aims to discover what determines the gain that attaches to the working, to the saving, and to the coordinating. As a man is a composite functionary, it tells us how much he naturally gets in each of his various capacities.

The nature of the distributive process. Social production is a synthesis of distinguishable elements. Distribution is an analysis; and it reverses the synthetic operation step by step. In organized production one worker does not complete a product from the beginning; if he applies his energy to crude nature and begins the making of something that the wants of society require, he passes the product in an incomplete state to a successor. This man in turn advances the article nearer to completion and hands it over to a third man. The product, when ready for final use, has passed through the

Distribution, Law Of, Table 1 Synthesis resulting in the completed product, clothing.

Subproduct	Resulting from:
1. Elementary utility: wool	Joint result of Capital and Labour.
2. Place utility: transporting	Joint result of C′ and L′.
3. Form-utility: manufacturing	Joint result of C″ and L″.
4. Form-utility: tailoring	Joint result of C‴ and L‴.

hands of a series of workers each of whom has put his touch on it and passed it to his successor.

The process may be represented by Table 1. The garment, when completed, is an aggregate of distinct utilities, and we use the term sub-product to denote the quality imparted to it by each specific group of producers. The sharing of the value that a coat represents among the groups that have performed the specific operations of production is an analytical operation, that follows, in a reverse direction, the steps of the productive synthesis.

The first sub-product in the series is wool. It embodies an 'elementary utility', or one that results from calling a raw material into existence. The merchant's sub-product is only the special utility imparted to the wool by conveying it to his warehouse, assorting it, and dividing it into quantities convenient for purchasers. It is mainly a 'place utility', which is the service-rendering quality that a thing acquires by being taken to the place where it can be used; though in a complete statement it would be necessary to recognize a 'form-utility' due to assorting and dividing. The manufacturer's sub-product is not the cloth, but the 'form-utility' imparted to the wool by transmuting it into cloth. The tailor's sub-product is the further 'form-utility' imparted to the cloth by making a coat of it. Each specific utility is created by the joint action of labour and capital; and each of these agents must have its share of the value embodied in its sub-product.

In order that the action of labour and capital within the sub-groups may be a joint-action at all, it is necessary that a certain coordinating act be done. Some one must hire labour of the right kind, borrow capital and invest it in the proper forms, and cause the two to cooperate. This is the work of the entrepreneur, in an unusually limited sense of the term. This functionary, in his capacity as entrepreneur, is not a capitalist and not a labourer, however frequently it may happen that the man who performs the coordinating function may perform others as well. The coordinator, as such, is not a business manager or superintendent. The performing of this function does not require salaried labour; indeed, after the process is begun, it scarcely requires effort at all. Bargaining operations first divide the total product of industry among the general groups of which society as a whole is composed. How much wealth shall come to the entire group of workers, capitalists, and entrepreneurs who are engaged in the creating of the finished products, woollen garments? That depends on the price for which the garments sell. A myriad of finished products from other groups in the world at large must come, by way of exchange, to minister to the wants of the men in this one group; and the quantity and quality of those products is fixed by the sale of the clothing. This sale, and others like it, perform the first and most generic dividing act that takes place in the process of distribution. It determines the total income of those who contribute to the production of clothing.

What fixes the part of the income of this general group that goes to each of the sub-groups that compose it? Bargains again. Each group must buy the utilities made by those that come earlier in the series, and sell them, with the addition of its own utility, to the group that succeeds it. The manufacturing group buys wool and sells cloth; and what it receives, less what it pays, constitutes the reward of the manufacturing operation. As the first division of the income of society resolves it into rewards of general producing groups, the first subdivision resolves the portion falling to one general group into shares for the sub-groups that constitute it.

A further division is to be effected: it is that of the shares falling to labourers, to capitalists, and to entrepreneurs in each sub-group. Here is the test operation of distribution; in this smallest of fields is created and divided the wealth that rewards each class in industrial society.

The productive operation from the fruit of which labour and capital get their pay is *intra-groupal*; it goes on within the specific industry in which a particular force of men and their quota of capital are engaged. The value that rewards woollen weavers and spinners and the men who furnish them capital is created wholly within the mill, and the sum that is divided between these classes is a sum on which no others have any claim. Yet the fact that labour and capital both migrate freely from group to group, so that workers from any group are able to share in the special gains that may come to the earners in any other, creates a certain solidarity of labour on the one hand, and capital on the other. Give to the wool spinners an advance of wages, and movements of labour will in the end distribute the gain among the whole working class. On the other hand, change the cardinal relations of labour and capital as a whole, and you change them in the end within every sub-group. Labour is in reality *trans-groupal*, and capital is the same. Each is a productive agent, the field of which extends directly across the sub-groups of the diagram. It is the relation of all capital to all labour that determines wages and interest. The law of wages is nothing if not general, and the same is true of the correlative law of interest.

It is a familiar fact that interest and wages tend toward uniformity in different occupations. Men of different productive powers may earn different rewards, even within a single trade; and the labour of management regularly receives more than work of the ordinary kinds. Men differ in the amount of working force that they possess, but men of like power tend to receive uniform wages throughout the series of industrial groups. If wages are high in the woollen mill the young men and women who are about to enter the field seek out this part of it, and by their competition reduce the wages there prevalent to the rate that prevails elsewhere. Interest tends to a similar uniformity; under free competition it tends to keep the same rate in all industries.

With interest has often been vaguely grouped what we have termed pure profit itself; the gross gains loosely attributed to capital tend toward equality. It is, however, in a special way that the element that we have distinguished as pure profit tends toward equality in different industries. Wherever it comes into existence it sets at work forces that tend to sweep it again out of existence. In a way this gain is self-annihilating. The uniform rate toward which pure profit tends – though it never reaches it in all groups at once – is a zero rate. Here indeed, we reach controverted ground, and can claim only to present one theory, not a view that has universal support; but the evidence in favour of the correctness of the view is simple and conclusive. Competition tends to annihilate pure profit. The existence in one sub-group of a gain that is in excess both of interest on all the productive wealth that is there used, and of pay for all labour, is an inducement to the entrepreneurs of the group to hire in the market both capital and labour, and secure the pure profit that their joint industry creates. Let woollen mills pay wages, including salaries, and a double interest on the capital that they use, and the mills will speedily enlarge their capacity. The increase in the product will then reduce the price of it, and ultimately bring the enlargement to an end. Under natural law the sub-groups are in stable equilibrium when, aside from insurance and taxes, each earns wages on all labour, including the labour of management, interest on all capital employed, and nothing more. On this point the testimony of experience confirms the conclusions of theory.

The equilibrium is never in practice perfect. Causes that cannot here be analysed in any fulness cause the element pure profit to continually reappear. Inventions, as applied in particular industries, give to one and another of the sub-groups a gain that is in excess of that which perfectly stable conditions would afford. The occupation of new land creates, in a local way, a pure profit for the earlier comers. Continually appearing in particular parts of the field, and slowly disappearing by reason of competition – such is this element of the social income. If we watch a single sub-group we find the profit at intervals appearing and disappearing; if we watch the industrial field as a whole we find it everywhere present, though not long at the same points. Pure profit depends on a relation between industrial groups. What the manufacturer pays to

the earlier groups in the series above represented, and what he receives from the tailoring group, determine this part of his gain. The actual position of the entrepreneur himself, in the diagram that describes the sub-groups, is on the line that separates his own industry from the following one. He is a purchaser of everything that is produced on the left of that line. In the buying of materials he purchases the products of the earlier sub-groups, and in the paying of wages and interest he virtually buys the sub-product created in the group to which he himself belongs. The entrepreneur of the woollen mill buys wool, and so pays for the sub-products created by wool growers and merchants; and he buys the form-utility created in the woollen mill itself by making bargains with workmen and capitalists, giving them fixed sums, and inducing them to relinquish their claims on the cloth. As the place of a particular workman and of a particular amount of capital is, in the diagram, *intra-groupal*, so that of a particular entrepreneur is *inter-groupal*. Workers and capitalists get their pay from results secured wholly within their own industries, while entrepreneurs get theirs from the fruits of mercantile transactions between earlier groups and later ones. Pure profit does not depend on the relation between capital and labour. Moreover, where this profit exists it is local, it depends on the relations between adjacent groups.

We have shown that there is no law of wages that is merely local. There is no force that gauges the pay of wool-spinning independently of the wages paid in other employments. There is a level toward which all wages tend. There is likewise a level toward which interest in every group tends. What is the law that fixes these levels? What is the general law of wages and interest? Here again we are on ground that is actively contested, and we therefore only indicate the nature of a certain theory without claiming for it a position of general acceptance, and without arguing any points in controversy.

In presenting it we may utilize a Ricardian formula for determining the rent of land. If we apply to a fixed area of land an increasing amount of labour, we get returns that diminish *per capita*. The first man set working on 100 acres creates a certain amount of wealth as the result of the tillage. Adding a second man does not double the crop. Adding a third does not increase by a half the product due to the former two. Each man, as he comes into the field, adds less to the total output of the industry than did any of his predecessors.

This hypothesis makes the men enter the field in a certain order of time, and the one who is the final man is so in a literal sense – he is the last to arrive. Actually putting the men into the field one at a time is not necessary in order to reveal the principle that governs the final productivity of labour. Let the full complement of men occupy the field at once, and there will still be what may be treated as the final increment of labour. Take any man away from the force that tills the field, and the remaining men will gain in *per capita* productivity by reason of his absence. The departure of one man out of a force numbering twenty does not reduce the crop by a twentieth, since the nineteen men remaining work at better advantage by reason of the withdrawal of one. The final productivity of labour is gauged by what would be lost if one man out of the force were to stop working. We may, by way of illustration, actually set the men working one at a time, and find what the last comer creates; or we may set them all working at once and see what would be lost by the departure of one. The conclusion is the same in either case: the final unit of labour is the least productive.

If, now, land were the only form of productive wealth that figured in the case, wages would equal the amount created by this final or twentieth man. That would gauge the amount that the employer would lose through the departure of any one man in the force. It would determine what he could afford to pay to any one. Each man tends to get what he is separately worth.

What would be true in the case of labour applied to land, and using no other capital worth considering, is actually true of labour applied to a fixed amount of general capital, or to a fixed quantum of wealth in all productive forms, including both land and other instruments. For the field of limited extent in the Ricardian illustration substitute a fixed value, expressible in pounds or dollars, and invested in such appliances of every

kind as working the needs of the working community require. If there are a hundred men in the force, the departure of one of them will not reduce the product by 1 per cent. His departure will add somewhat to the productivity of the remaining workers. After he is gone the capital will adapt itself in form to the needs of the ninety-nine, and it will be in a slight degree more ample in quantity per man. Wages are gauged, as in the former case, by the final productivity of labour. What on the whole is lost by the departure of one man fixes the importance to employers of every man. If each man gets what employers would lose by his absence, he gets that he is effectively worth.

This principle in a reversed application fixes the rate of interest. It is the productivity of the final increment of capital, as employed by a fixed labour force, that gauges the pay of each increment. Let there be 100 men using 100 units of capital. Take, now, one unit of capital away, and you will not reduce the product by 1 per cent. The 99 units of capital will have gained in productivity per unit in consequence of the departure of the hundredth. The loss inflicted on the entrepreneur by the withdrawal of the one unit of capital gauges the importance of any single unit. Each unit of capital gets as its compensation what would be lost if one unit of capital were withdrawn. This diminution of the total product due to the departure of the final unit of capital gauges the importance to the entrepreneur of each separate unit. It determines what he will pay for the use of each one. Interest is therefore gauged by the final productivity of capital. Each pound or dollar tends, under natural law, to secure for its owner what, in production, it is separately worth.

Bibliography

Cairnes, J.E. 1874. *Some leading principles of political economy*. London: Macmillan.

Clark, J.B. 1886. *The philosophy of wealth*. Boston: Ginn & Co.

Clark, J.B. 1888. *Capital and its earnings*. Baltimore: American Economic Association.

Clark, J.B., and F.H. Gidding. 1888. *The modern distributive process*. Boston: Ginn & Co.

George, H. 1879. *Progress and poverty*. New York: Appleton.

Longe, F.D. 1886. *A refutation of the wage-fund theory of modern political economy*. London: Longmans, Green.

Thornton, W.T. 1869. *On labour*. London: Macmillan.

von Böhm-Bawerk, E. 1884–9. *Kapital und Kapitalzins*. Innsbruck: Wagner.

Walker, F.A. 1876. *The wages question: A treatise on wages and the wages class*. New York: H. Holt & Co.

Walker, F.A. 1883. *Political economy*. New York: H. Holt & Co.

Wieser, F.F.B. 1889. *Der Natürliche Werth*. Vienna: Hölder.

Distributive Justice

Edmund S. Phelps

Social justice is justice in all of the relationships occurring in society: the treatment of criminals, children and the elderly, domestic animals, rival countries, and so forth. Distributive justice is a narrower concept for which another name is economic justice. It is justice in the economic relationships within society: collaboration in production, trade in consumer goods, and the provision of collective goods. There is typically room for mutual gain from such exchange, especially voluntary exchange, and distributive justice is justice in the arrangements affecting the distribution (and thus generally the total production) of those individual gains among the participants in view of their respective efforts, opportunity costs, and contributions.

In earlier times the discussion of distributive justice tended to focus upon the obligations of the individual toward those with whom he or she had exchanges. So an employer was expected to be just or not to be unjust, and the problem was to demarcate employer injustice. With the rise of governments capable of redistribution and the spread of economic liberalism, the focus shifted to the distributional obligations of the central government. Let enterprises and households pursue their self interests while the government attends to distribution (within the limits of its just powers). Distributive justice is largely about redistributive taxation and subsidies. The latter may take many

forms such as public expenditures for schooling and vocational training (beyond the point justified only by the Pareto principle from the status quo ante) as well as cash subsidies for the employment of labour or low-wage labour (whether paid to employer or employee).

Note that the so-called negative income tax, whatever the claims for or against it as a tool of social justice, does not appear to be an instrument for distributive justice unless restricted somehow to those participating (more than some threshold amount?) in the economy (and thus in the generation of the gains to be (re)distributed). In any case, it will not be discussed here, although some propositions about subsidies apply also to the negative tax.

The suggestion that distributive justice might (at least in principle) require subsidies, not merely tax concessions or tax forgiveness for the working poor, tends to raise the eyebrows of some and accounts for the fact that distributive justice raises the hackles of a few. As long as the Iroquois and the Sioux have no contact, there are no gains to be distributed and distributive justice does not apply; if they are let free to engage in bilateral inter-tribal exchanges, however, the payment of a subsidy to pull up the wage of the lowest earners, who are Sioux, say, would come partly or wholly at the expense of the Iroquois. Now some commentators object to the notion that the Sioux, whose exchanges with the Iroquois are entirely voluntary and all of whom have benefited (or could have), we may suppose, might deserve an additional payment from the Iroquois, perhaps through some supra-tribal authority. Ayn Rand (1973), for example, argues that it is one thing to require of a poor person a fare for riding a bus with empty seats that the other riders can finance out of the benefits they receive from the bus – she has no qualms about such a free ride – and another thing for the poor person to tax the other riders. But she has got the economics wrong in the application of her (actually rather Rawlsian) ethical premise. Up to a point, a subsidy to the poorest-earning group (the Sioux in the above example) would have the others (the Iroquois) still with a net gain – a gain after the tax needed to pay the subsidy. This is because of diminishing returns: When the group of Sioux workers is added to the fixed pool of Iroquois' labour and land, the extra product added by the first arrivals – and, more generally, the average of the extra products added by the succession of Sioux workers – is larger than the extra product resulting from the last of these workers, which is the 'marginal product' of Sioux labour; the Iroquois could afford a subsidy equal to the excess of the average extra product over the marginal product. Correctly applied, then, the Randian objection is to a gain-erasing or, at any rate, a gain-reversing subsidy, not to *any* subsidy whatsoever.

Another objection to the concept of distributive justice and to the admissibility of subsidies argues that if these notions were sound it would make sense, by analogy, to apply them to marriage allocation, to the matching of husbands and wives; since we never hear of such applications the ideas are presumably unsound. Of course, it would strike us as novel and foreign to see a proposal for a tax on marriage with Iroquois men and a subsidy to marriage to Sioux men on the ground that the former were apparently more attractive to women (from either tribe) and the resulting inequality of benefits unjust and demanding correction. But the reasons might be other than the supposed unacceptability of the ideas of distributive justice. Maybe the impracticality of deciding on the taxes and subsidies stands in the way. Perhaps a marriage subsidy would be demeaning while employment subsidies would not, being graduated or even a flat amount per hour. Yet the key observation may be that, although there is economic exchange here and although racial discrimination or racial prejudices could cause real injustices, the Sioux and Iroquois men in this example are not cooperating for mutual gain and so no problem about the just division of such gains can arise; they are competing, or contesting, for partners, not forming partnerships with one another. Thus distributive justice cannot apply here.

The terms offered to the working poor, as already implied, is the locus classicus to which notions of distributive justice have been applied. However, two other arenas in which issues of justice are being fought out should be

mentioned. One of these is the problem of intergenerational justice. It was first addressed in a celebrated paper in 1928 by Frank Ramsey, who adopted as the criterion of optimality the standard associated with utilitarianism – the sum of utilities over time. This conception of intergenerational justice encountered difficulties when in the 1960s it was applied to optimum saving of a society in which the population is to grow without bound, although that odd demographic case may have put utilitarianism to an unfair (and absurd) test. In 1970 John Rawls struggled with the problem of intergenerational justice in a famously problematic section of his, only to conclude that '. . . the difference principle [i.e., Rawls's maximin or, more accurately, leximin principle] does not apply to the savings problem. There is no way for later generations to improve the situation of the least fortunate first generation.' This seems to say that intergenerational justice, if there is such a thing, is not a problem of distributive justice, since there is no cooperation for mutual gain among generations, not even between adjacent ones in the chain. But the premise that the current generation cannot be helped by succeeding generations appears, on the face of it, to be a slip in Rawls's economics. In a closed economy, we can help future generations by providing them with more capital – even in an open economy enjoying perfect capital mobility, we can provide them with social overhead capital that the world capital market would not provide (or not so cheaply) – and, if overlapping with us, they can help us by meeting consumption claims we make through our issue of public debt and pension entitlements. Thus distributive justice does apply here, with a precision fit. What Rawls may be interpreted to mean is that if, being the least fortunate owing to heaven-sent technological discoveries over the future, the present generation were permitted to invest nothing (not even gross of depreciation!) – rather as we can imagine the poorest in the static problem to begin by sullenly asking for equality – the future generations could not bribe the present one to do something in their mutual interest – unlike the static problem in which the rich can explain the benefits of trickle-down. But in fact the next generation *can* bribe the present one with some old-age consumption in return for some investment. It may be conjectured that a maximin-optimal growth path would still exist in a model along the lines of the Phelps-Riley model notwithstanding the introduction of technological progress.

The other arena in which we find a debate over distributive justice is the international trade field. When a giant nation trades with a small number of pygmy countries, not large enough even in the aggregate to influence relative prices in the giant state, the latter receive all the gains from trade and the former gets nothing and loses nothing; this is exactly the Rawlsian maximin solution if perchance the pygmy countries are poorer (in some suitably defined way) than the giant. But if these tiny countries 'spoil the market', worsening their terms of trade in the course of exporting to and importing from the giant, because they are not of negligible size at least in the aggregate, then the Rawlsian solution is not obtained by the free market. The recent North–south problem of which the 'Southern' countries complain can be understood as the tendency of the 'Northern' countries that are already the richest countries, such as the North American and European countries, to retain the gain from trade resulting from the aforementioned change in the terms of trade caused by the 'Southern' countries through their trade with the 'Northern' ones. The 'Southern' countries believe justice to require that the 'Northern' countries arrange to give back that gain through some appropriate international transfer mechanism.

There are able and serious philosophers who would be happy to see distributive justice left to the economists. In fact, the history of philosophy has been seen as a process of divesting itself of a sub-field as soon as it could thrive independently. Likewise, there are economists who would leave the subject to philosophers. But, whichever group receives the lion's share of the contract to work on it, it seems that the economics (as well as philosophy) of the problems being studied is an essential element of the subject. In this sense and for this reason, the necessary cross listing notwithstanding, distributive justice is an important field under economics.

See Also

- ▶ Entitlements
- ▶ Equality
- ▶ Equity
- ▶ Exploitation
- ▶ Justice

Bibliography

Phelps, E.S., and J.G. Riley. 1978. Rawlsian growth: Dynamic programming of capital and wealth for intergeneration 'maximin' justice. *Review of Economic Studies* 45(1): 103–120.

Ramsey, F.P. 1928. A mathematical theory of saving. In *Economic justice*, ed. E.S. Phelps. Harmondsworth: Penguin.

Rand, A. 1973. Government financing in a free society. In *Economic justice*, ed. E.S. Phelps. Harmondsworth: Penguin.

Rawls, J. 1970. *A theory of justice*. Oxford/Cambridge, MA: Oxford University Press/Harvard University Press.

Distributive Politics and Targeted Public Spending

Brian G. Knight

Abstract

This article analyses common pool problems associated with the provision of local public goods by central legislatures. In response to incentives associated with common pool problems, legislators act to maximize spending for their home jurisdiction but to restrain spending elsewhere due to the associated tax costs. The resolution of this conflict between jurisdictions depends in the United States upon the distribution of political power across Congressional delegations. Incumbents are rewarded for delivering federal spending to their jurisdiction through increased voter support.

Keywords

Common pool problems; Distributive politics; Earmarked projects; Lobbying; Local public goods; Proposal power; Targeted public spending

JEL Classifications

H5

While conventional models of political economy, such as the median voter model, focus on the provision of national public goods, most federal spending programmes, such as the US interstate highway system, are more aptly characterized as local in nature. While in the United States the benefits of federal spending are concentrated in specific geographic units, such as states, counties, and Congressional districts, the associated tax costs are, by contrast, geographically dispersed. This common pool feature of federal spending – concentrated spending but dispersed financing – leads to a geographic tug-of-war in which jurisdictions attempt to increase own-jurisdiction spending but to reduce spending elsewhere due to the associated tax costs. This conflict between jurisdictions is reflected most intensely in the budget process within the US Congress, whose members are locally elected and thus naturally respond to these common pool incentives.

In this article, I first summarize evidence suggesting that Congressional representatives are responsive to the common pool incentives associated with concentrated spending but dispersed costs. Having established the empirical saliency of this common pool problem in Congress, I then summarize the literature examining how this conflict is resolved. In particular, I analyse the effects of Congressional delegation characteristics, such as size, ideology, seniority, and committee assignments, on the geographic allocation of federal funds. Finally, I review evidence on the effects of the geographic distribution of federal funds on electoral outcomes.

As described in Knight (2006), common pool problems underpin several theoretical models of the legislative process, such as the universalism model of Weingast, Shepsle, and Johnsen (1981) and the legislative bargaining model of Baron and Ferejohn (1989). Whether or not Congressional delegations respond to these incentives in practice, however, is primarily an empirical question. It may be the case, for example, that political parties, or related Congressional organizations, serve as collective mechanisms through which legislators internalize the tax costs in other jurisdictions associated with own-jurisdiction spending. One of the first papers to directly measure the responsiveness of representatives to common pool problems is by DelRossi and Inman (1999), who examine the geographic distribution of water projects authorized by the Water Resources Development Act of 1986. In particular, the authors compare the size of project requests before and after changes in local matching requirements, which significantly increased the fraction of project costs financed by local governments. As hypothesized, districts experiencing larger increases in matching rates requested significantly less funding for water projects. In a similar vein, Knight (2004b) examines Congressional voting in 1998 over whether to finance a set of transportation projects, which were earmarked for specific Congressional districts and were funded primarily via federal gasoline taxes. As predicted, support for funding was concentrated in those districts receiving more in funding and also in those districts with lower gasoline tax burdens.

How is this geographic battle between jurisdictions resolved? Which states and Congressional districts win and why? Regarding the mere size of delegations, an important feature of the US Congress is its bicameral structure in which each state has an equal number of delegates in the Senate but in which seats are apportioned between states according to population in the House of Representatives. This equality of delegation sizes in the US Senate provides small states with power disproportionate to their population; Senators from California, the largest state, currently have over 60 times as many constituents as do senators from Wyoming, the smallest state. In attempting to measure the magnitude of this small-state bias, Atlas et al. (1995) and Lee (1998) find that small states receive significantly more per capita in aggregate federal spending than do large states. While this finding is certainly provocative, it is difficult to distinguish between the role of Senate representation and other factors, such as population density, that make small states inherently different from larger states. In attempting to address this issue of unobserved differences between small and large states, Knight (2004a) demonstrates that small states receive considerably more per-capita funding in projects earmarked in Senate bills; in House bills, by contrast, small and large states receive similar project spending on a per-capita basis. Knight (2004b) also identifies two theoretical channels underlying this small-state bias in the US Senate. Relative to their population, small states are disproportionately represented on key committees (the proposal power channel) but are also cheaper coalition partners (the vote cost channel) given that they pay a smaller share of federal taxes. Interestingly, both channels are shown to be empirically important and, taken together, explain over 90 per cent of the measured small-state bias. In a related study of the size of delegations, Falk (2006) studies discontinuities in the apportionment of seats in the US House arising from both timing (re-apportionment occurs once every ten years) and rounding issues (delegation sizes must be integers). Using this variation in delegation sizes, he finds that increases in seats per capita lead to statistically significant increases in federal spending per capita.

Delegations of similar sizes, however, may differ significantly in their composition. Key differences between delegations in the degree of political power include majority party affiliation, seniority, and representation on key committees. Regarding majority party affiliation, Levitt and Snyder (1995) find that the Democratic Party used its majority control of Congress to channel

federal funds into Congressional districts with a high percentage of Democratic voters during the period 1984–90. However, they find no evidence that, conditional on the percentage of Democratic voters, districts represented by Democrats received higher federal spending. Levitt and Poterba (1999) report that states with very senior Democratic representatives experienced more rapid economic growth than did other states. However, they find no relationship between the partisan affiliation of delegations and the geography of federal spending, a key hypothesized channel of the measured differences in economic growth. Regarding the role of Congressional committees, Knight (2005) finds that Congressional districts represented on key committees received substantially more funding in projects earmarked in transportation bills authorized in 1991 and 1998. He interprets this result as evidence of the importance of proposal power associated with the committee's ability to set the legislative agenda. De Figueiredo and Silverman (2002) examine interactions between committee representation and lobbying in an empirical examination of earmarked projects for universities. In particular, they find a strong correlation between lobbying outlays by universities and the receipt of federal funding; this link between lobbying and spending, however, is found to be much stronger for those universities located in districts that are represented on key appropriations committees.

We have focused throughout this survey on the determinants of the geographic distribution of federal funds. Politicians, however, have an incentive to put forth the effort to secure project funding only if they perceive that the associated political gains are sufficiently high. While clearly important, measurement of the effects of federal spending on incumbent vote shares is plagued with endogeneity problems. For example, incumbents facing the strongest opposition have the strongest incentives to put forth effort in securing funds. Thus, there may be a downward bias in ordinary least squares (OLS) estimates of the effect of federal spending on incumbent vote shares. As an instrument for district-specific federal spending, Levitt and Snyder (1997) use federal spending outside of the district but within the state. The idea is that other actors, such as Senators or governors, also play a role in the geographic distribution of federal funds. Using this exogenous variation in federal spending, they conclude that an additional $100 per capita in spending translates into an additional two percentage points in incumbent vote shares.

We conclude that common pool problems associated with concentrated project benefits but dispersed costs are reflected not only in the behaviour of Congressional delegations but also in the resulting distribution of federal funds. Who wins and who loses in this geographic battle is determined in part by state size and the political power of delegations. Consistent with these results, evidence suggests that incumbent re-election prospects are significantly enhanced by increases in federal spending.

See Also

- ▶ Campaign Finance, Economics of
- ▶ Fiscal Federalism
- ▶ Intergovernmental Grants
- ▶ Local Public Finance
- ▶ Political Institutions, Economic Approaches to
- ▶ Public Choice

Bibliography

Atlas, C.M., T.W. Gilligan, R.J. Hendershott, and M.A. Zupan. 1995. Slicing the federal government net spending pie: Who wins, who loses, and why. *American Economic Review* 85: 624–629.

Baron, D.P., and J.A. Ferejohn. 1989. Bargaining in legislatures. *American Political Science Review* 83: 1881–1207.

de Figueiredo, J.M. and B.S. Silverman. 2002. Academic earmarks and the returns to lobbying. Working Paper No. 9064. Cambridge, MA: NBER.

Falk, J. 2006. The effects of Congressional district size and representative's tenure on the allocation of federal funds. Working paper, University of California, Berkeley.

DelRossi, A.F., and R.P. Inman. 1999. Changing the price of pork: The impact of local cost sharing on legislators' demands for distributive public goods. *Journal of Public Economics* 71: 247–273.

Knight, B.G. 2004a. Legislative representation, bargaining power, and the distribution of federal funds: Evidence from the U.S. Senate. Working Paper No. 10385. Cambridge, MA: NBER.

Knight, B.G. 2004b. Parochial interests and the centralized provision of local public goods: Evidence from congressional voting on transportation projects. *Journal of Public Economics* 88: 845–866.
Knight, B.G. 2005. Estimating the value of proposal power. *American Economic Review* 95: 1639–1652.
Knight, B.G. 2006. Common tax pool problems in federal systems. In *Democratic constitutional design and public policy analysis and evidence*, ed. R.D. Congleton and B. Swedenborg. Cambridge, MA: MIT Press.
Lee, F.E. 1998. Representation and public policy: The consequences of Senate apportionment for the geographic distribution of federal funds. *Journal of Politics* 60: 34–62.
Levitt, S.D., and J.M. Poterba. 1999. Congressional distributive politics and state economic performance. *Public Choice* 99: 185–216.
Levitt, S.D., and J.M. Snyder. 1995. Political parties and the distribution of federal outlays. *American Journal of Political Science* 39: 958–980.
Levitt, S.D., and J.M. Snyder. 1997. The impact of federal spending on house election outcomes. *Journal of Political Economy* 105: 30–53.
Weingast, B.R., K.A. Shepsle, and C. Johnsen. 1991. The political economy of benefits and costs: A neoclassical approach to distributive politics. *Journal of Political Economy* 89: 642–664.

Diversification of Activities

A. Cosh

Diversification is the process by which the modern corporation extends its activities beyond the products and markets in which it currently operates. It is a major determinant of the structure of modern industrial economies and has important implications for competition and efficiency. Robinson (1958, p. 114) defines diversification as 'the lateral expansion of firms neither in the direction of their existing main products, as with horizontal integration, nor in the direction of supplies and outlets, as with vertical integration, but in the direction of other different, but often broadly similar, activities'. The extent of diversification can be measured in a number of ways, but is hampered by the difficulty of precisely defining the boundaries between different products, markets and industries. It is not a simple task to assess the degree to which a firm spreads its operations over different activities. The more narrowly defined are these activities the greater will be the apparent degree of diversification. These problems are not unique to the measurement of diversification and similar difficulties arise in the measurement of concentration in industry. Indeed the process of diversification itself has played a major part in blurring the distinction between industries and in creating these measurement problems. However, it is clear that diversification must involve the firm in producing new products which are sufficiently different from its existing products to involve the firm in new production or distribution activities. Diversification may therefore involve only a small change of direction, or a dramatic switch into an entirely new line of business. In the literature the former is referred to as related, or narrow spectrum diversification and the latter as unrelated, or broad spectrum diversification.

One possible measurement of the extent of diversification involves identifying the number of industries, or products in which the firm is involved. The other main approach is to measure the proportion of the firm's activity in its core business in comparison with the proportions in its diversified activities. This measure has been refined in a number of ways to take account of the number and importance of these diversified activities (e.g. Berry 1975; Jacquemin and Berry 1979; Utton 1979).

The process of diversification is not a new phenomenon, but the principal empirical studies (Gort 1962; Rumelt 1974; Berry 1975; Utton 1979) have demonstrated a marked increase in the degree of diversification over the past few decades. The studies suggest that diversification tends to be narrow spectrum diversification into similar industries. However, both Gort and Rumelt were able to discern some shift towards broad spectrum diversification. The intensity of narrow spectrum diversification was found to be industry related, but the extent of broad spectrum diversification was independent of the primary industry from which diversification was occurring. Rumelt was also able to identify a growth in importance of acquisitive conglomerates and

there can be little doubt that their importance has grown further since his study. There was general agreement that firms tended to diversify into industries characterized by high research and development intensity and rapid technological change. The industries also tended to be faster growing, but showed no significant differences in terms of profits variability, or the degree of concentration, than industries less popular with diversifying firms. The industries from which higher levels of diversification occurred were not slower growing than other industries, but did tend to be characterized by a higher degree of seller dominance. Such industries might give less scope for firm growth by capturing market share. The more rapid diversifiers tended to be larger firms with higher proportions of scientific and technical employees and this is consistent with the importance of technological industries as diversification choices which was noted above. Finally firms with above average rates of diversification tended to have above average rates in subsequent periods. This may be related to the organizational changes associated with diversification which have been identified by Chandler (1963) and others (e.g. Williamson 1970; Channon 1973). This issue is explored further below.

The growth of firms and the role of diversification in this growth process were elucidated in the pioneering work of Penrose (1959). Penrose identified three explanations for diversification: first, as a response to specific opportunities; second, as a response to specific threats; and third, as a general strategy for growth. The opportunity to diversify arises naturally as a byproduct of the existing activities of the firm. A key area is the research and development activities of the firm. Such activities develop the firm's knowledge of its technology which is unlikely to be product specific. Furthermore whether research is carried out only to improve the firm's existing products, or the develop new products, it is likely to provide new opportunities for diversification. The knowledge of the markets for its existing products and their channels of distribution provide the firm with other opportunities for diversification. Another opportunity for diversification arises from retained earnings from existing activities. The finding that these earnings are invested in diversification rather than, for example, paying dividends, is probably associated with the growth orientation of management and the tax position of shareholders. Thus the normal operations of the firm create both new opportunities for expansion and the availability of unused productive resources to meet these opportunities. The second explanation offered by Penrose concerns the exposure declines in demand for their products. Diversification is a means of spreading risk through reducing the firm's dependence on a few products. The reduction in perceived risk may also reduce the cost of capital to the firm. Diversification may also occur in response to diversification by a competitor. This type of competitive strategy raises the question of the implications of diversification for competition and this issue is examined below. Finally diversification may occur as part of a general policy for growth. This part of Penrose's work has been taken further by Marris (1964). Marris gives diversification a central role in his model of the growth of firms. The management of firms have a strong motivation to seek growth since it confers on them improved status, salary and security. But growth within their existing markets will eventually be limited by the growth of demand for these products and diversification is the means by which this demand constraint may be overcome. It has been argued above that a certain degree of diversification will be both natural and beneficial as opportunities are exploited. However, Marris argues that management will be prepared to press growth, and hence diversification, beyond the level which is optimal for shareholders. The drawback of too rapid a rate of diversification is a higher failure rate of new products due to a lack of managerial, financial, development and marketing resources. This may be a less reasonable proposition when the possibility of growth through merger is recognized. However before considering this it is worth looking at the changing structure of firms which has evolved with diversification.

The development of the M-form, divisionalized company was identified by Chandler (1963) to be a response to the growth and, more particularly, diversification of the modern

corporation. Subsequent research (e.g. Channon 1973; Rumelt 1974; Williamson 1970, 1975) has reinforced their inter-connection to such a degree that it is necessary to interpret the consequences of diversification within the context of the divisionalized company structure. In this structure responsibility for profitability is restored to divisional managers whose performance can be assessed. Top management is freed from day-to-day operational decisions and can concentrate on the allocation of funds between the divisions and other aspects of strategy. The divisional structure significantly reduces the organizational constraints of diversified growth, particularly growth by acquisition. The acquisition of new divisions, or sub-divisions, by takeover can be achieved quickly and with minimum disruption. Diversification through merger is often seen as less risky since it involves the acquisition of the physical assets, existing products and channels of distribution required and brings with it management and employees who are experienced in this area of activity. Furthermore, entry is achieved without initially having to compete for a market share. On the other hand if the motive for diversification is to utilize spare resources within the firm, or to exploit some technological development, then diversification by internal growth may be preferred. It appears that the importance of diversification mergers has increased in recent decades, partly as a response to the increase in strength of competition policy.

> The US merger laws have evolved into a potent deterrent against sizeable horizontal and vertical mergers. It is doubtful, however, whether they have had much impact on the overall level of merger activity which has continued at high levels (Scherer 1980, p. 588).

A substantial controversy surrounds the question of what impact diversification has on competition and efficiency. At first sight the creation of large, non-specialized firms would be expected to reduce both, but there are counter arguments. The evidence does not suggest that diversification raises market concentration. Indeed, broad spectrum diversification may be a force for reducing concentration in individual markets. Large firms diversifying are able to overcome many barriers to entry and may promote competition by their entry. Diversification may be the only means by which firms may grow large enough to reap pecuniary economies of scale, without becoming too dominant in a single market. It is also argued that the diversity of products, as well as large size, brings a greater potential benefit from research. Therefore large, diversified firms may be more likely to engage in intensive research and development, to the benefit of the whole economy. The associated introduction of the M-form organization is argued to lead to improved internal efficiency of the firm as divisions strive to meet profit targets and compete for funds. It is also argued that the internalizing of the capital market within the large, diversified firm can lead to improved allocative efficiency. This is created by top management, who hold better information than investors, allocating funds to their most profitable use. On the other hand there are several arguments which suggest that the growth of the diversified firm has the potential to create reduced competition and efficiency. It was noted earlier that there has been a high proportion of narrow spectrum diversification.

> At least one possible interpretation of this finding is that the diversification that has led to relatively rapid rates of corporate growth (or has accompanied it) has not in general been to markets where the entering firm is a new and potentially competitive force. Rather, that 'diversification' has been to markets that are related to – and potentially if not actively competitive with – those in which the entering firm will frequently share what ever market power already exists. This kind of diversification is only one small step removed from the consolidation of market power through horizontal acquisition (Berry 1975, pp. 74–5).

Furthermore, the internalizing of capital markets has led to the removal of information and decision-making from the investor and led to a concentration of economic power. 'This means that the diversified, divisionalized firm is increasingly becoming the arbiter of intersectional shifts in funds' (Rumelt 1974, p. 155). Another focus of concern has been the potential for predatory pricing behaviour in which the diversified firm uses cross-subsidization between divisions to eliminate, or discipline, more specialized rivals and

so achieve higher long-run profits. A further possibility is reciprocal purchasing agreements when a firm is significant both as a seller to and buyer from another firm. It is argued that such practices are more likely to be found amongst large, diversified firms, but there is little evidence for the widespread existence of either predatory pricing, or reciprocal purchasing behaviour. Finally there is the spheres of influence hypothesis which recognizes the pervasive influence of large, diversified firms in almost all markets. Conglomerates might recognize that aggressive behaviour against another conglomerate in one market would have adverse consequences in other markets. It is possible that a symmetry of market power might emerges which would blunt competition. The answer to many of the empirical issues concerning diversification are as yet unresolved. This is in part due to a lack of sufficient research, but also in part due to the fact that the process of diversification is continuing. When, and if, a more stable period emerges the uncompetitive consequences outlines above may become more apparent.

Bibliography

Berry, C.H. 1975. *Corporate growth and diversification*. Princeton: Princeton University Press.

Chandler, A.D. 1963. *Strategy and structure: Chapters in the history of the industrial enterprise*. Cambridge, MA: MIT Press.

Channon, D.F. 1973. *The strategy and structure of British enterprise*. London: Macmillan.

Gort, M. 1962. *Diversification and integration in American industry*. Princeton: Princeton University Press.

Jacquemin, A., and C.H. Berry. 1979. Entropy measure of diversification and corporate growth. *Journal of Industrial Economics* 27(4): 359–369.

Marris, R.L. 1964. *The economic theory of managerial capitalism*. London: Macmillan.

Penrose, E.T. 1959. *The theory of the growth of the firm*. Oxford: Basil Blackwell.

Robinson, E.A.G. 1958. *The structure of competitive industry*. Rev. ed. Cambridge: Cambridge University Press.

Rumelt, R.P. 1974. *Strategy, structure and economic performance*. Cambridge, MA: Harvard University Press.

Scherer, F.M. 1980. *Industrial market structure and economic performance*. Chicago: Rand McNally.

Utton, M.A. 1979. *Diversification and competition*, The National Institute of Economic and Social Research Occasional Paper No. 31. London: Cambridge University Press.

Williamson, O.E. 1970. *Corporate control and business behavior: An enquiry into the effects of organisation form on enterprise behavior*. Englewood Cliffs: Prentice-Hall.

Williamson, O.E. 1975. *Markets and hierarchies: Analysis and anti-trust implications, a study in the economics of internal organization*. New York: Free Press.

Divided Populations and Stochastic Models

D. G. Champernowne

Introduction

The title of this entry requires some explanation. I use the term ‘stochastic models’ to distinguish those theoretical models which include one or more stochastic variables from ‘determinist models’ which do not. I shall confine attention to some stochastic models which are obtained by introducing into a determinist model a single stochastic variable (which can be multivariate, but will in illustrative examples be univariate). I shall use the term ‘generating system’ to mean a determinist model in which from an initial state of the system an unending sequence of successive states of the system can be exactly predicted by means of a set of rules such as lagged equations. It is convenient to distinguish generating systems from stochastic models rather than extend the former class to include some or all of the latter. The important feature of stochastic models is that they can make allowance for wide margins of uncertainty and ignorance.

By a ‘divided population’ I shall generally mean a frequency distribution most of which is closely clustered around two or possibly more peaks, but fairly empty elsewhere: an extreme case would be that where the peaks were completely separated by an unoccupied stretch.

However, the term 'divided population' can occasionally be extended to refer to a society which is divided into groups with contrasted living conditions, prospects and aims.

The term 'crisis' refers to an unstable situation where a small disturbance could tip the scales between the prospects of two widely different eventual outcomes. In a determinist model the representation of such a crisis would be a point of unstable equilibrium, and in a stochastic model based on that determinist model, one would still regard the point of instability as indicating crises facing that part of the population found close to it, by 'crises' meaning here the crises that chance might play a predominant part in determining their future prospects.

As a preliminary to the main discussion, it will be helpful to consider some standard tools for use with determinist models involving divided populations and crises.

Some Standard Methods for the Study of Unstable Situations

A standard method of constructing a model of the response of an economic system to the passage of time, or to possible changes of policy or of outside influences, is to set up a generating system giving a set of initial conditions containing the present and recent values of a set of economic variables and policy parameters, together with a set of rules for calculating the set of the same variables one time-unit later and repeating this operation successively for any required number of time-units. Such rules would normally take the form of a number of equations giving the values of each variable as functions of the values of other variables, mainly at earlier dates, taking account of the present values assumed for any policy parameters. We may confine attention to very simple examples of such models.

It is quite usual to find in simple models that given the initial information, the application of the system of rules with fixed policy parameters will generate a sequence of sets of values of the variables which tend to a long-run equilibrium set, apart possibly from one or more constant common growth-rates. But it is also possible to frame fairly simple rules which lead to oscillations which persist at a constant amplitude. Often these will be smooth and sinusoidal, but there is another possibility which is the one relevant to crises, where there are periodic jumps from one smooth steady path (which we might call boom) to another smooth steady path (which we might call slump) alternately to and fro indefinitely.

A convenient tool for the representation of such systems when there are sufficiently few equations involved is the phase diagram. If we are dealing with difference equations of the kind just described, the axes of the diagram could measure the values of one important variable along the horizontal axis as independent variable and the change of that same variable over the next time-unit along the vertical axis as dependent variable. In such diagrams the curve relating the change of the variable as a function of the value itself will reveal points of equilibrium by its intersections with the horizontal axis: however, where the curve cuts the axis from below on the left, the equilibrium will evidently be unstable, and we shall call such equilibrium points crisis points. Figure 1 is a phase diagram applicable to the determinist model described below in section "Rules for a Model Generating Lines of Bequests".

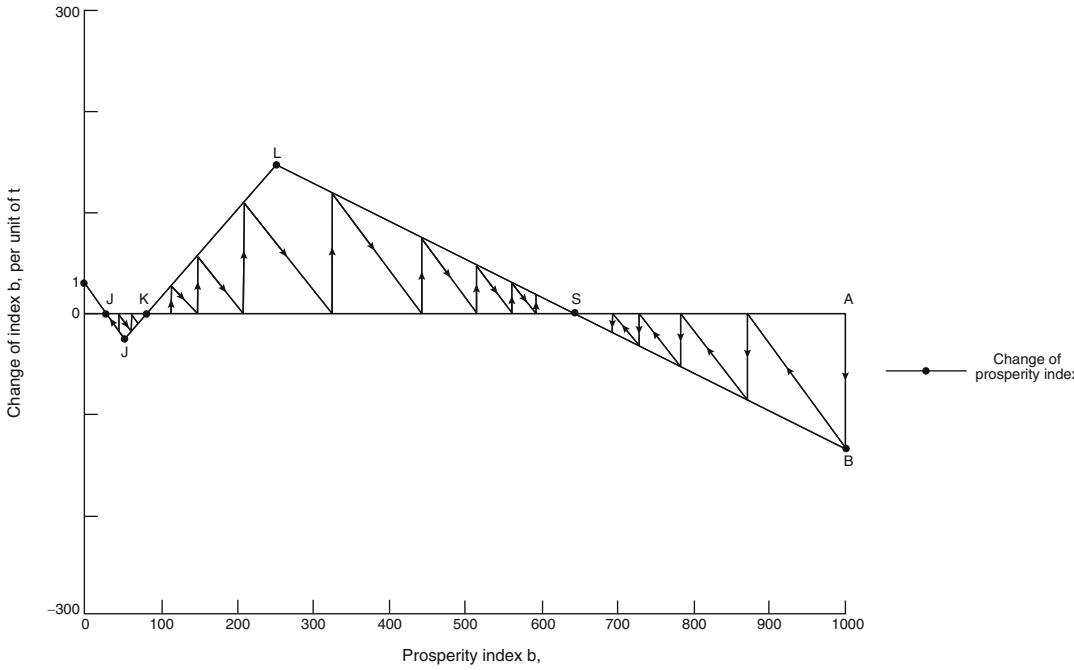
In Fig. 1 the horizontal axis is cut by the graph *ITJKLSB* in three points *J*, *K* and *L* denoting equilibrium levels of the index of prosperity, but the point *K* is a '*crisis* point' indicating an *unstable* equilibrium value. The arrows following the paths starting from *A* and from near *K* illustrate how, given the level of the index in an initial period, the chart may be used to predict its values in later periods assuming the rules of the model to be obeyed. For example, to follow the changes from the initial value of 1000 at *A* on the horizontal axis, measure horizontally the same (negative) distance AA_1 as the vertical distance of the graph from *A*. Having marked A_1, for the value after one unit of time, repeat the operation from A_1, to mark in A_2 and so continue as illustrated in Fig. 1. It is apparent that the series of such values must converge to the value marked by *S*, and similarly that starting from *C*, near *K* on the right, when the distances concerned will now be positive, (to the

right or upwards), it should again be clear that the series obtained must converge to the same stable equilibrium value *S*. Finally, starting from any point to the left of *K*, the usual procedure must result in a series arriving at the other stable equilibrium point *T*: (this is so because the gradient of the line *ITJ* is −1 in this example, which entails that as soon as the path hits *ITJ* it leads directly to *T*). This illustrates why the equilibrium at points, such as *K*, where the axis is cut by the curve from below on the left are unstable, while the points such as *T* and *S* mark stable equilibrium values.

Fig. 1 will be used again in section "Rules for a Model Generating Lines of Bequests" to illustrate the numerical example there, which involves equations (3) to (6): Table 1 in that section provides the first few values of the sequences that would be obtained by applying the rules, starting from values 1000, 100 and 70 respectively.

An early example of a determinist economic model involving oscillation between two points of stable equilibrium across a gap containing a crisis point of unstable equilibrium, due to the interventions of a disturbing force moving the phase-curve, was the model of the trade cycle published by Kaldor in the March 1940 issue of the *Economic Journal*. This contained a diagram closely related to a phase diagram of the elementary type shown here in Fig. 1, and which relied on the property that the curve itself moved upwards or downwards, depending on whether the currently relevant point representing equilibrium was on the right or left of the diagram.

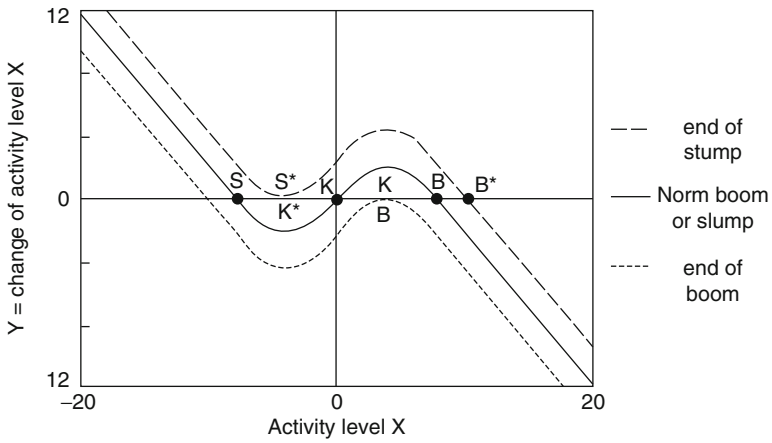
Figure 2 is a transposition of Kaldor's diagram into a phase diagram of the type outlined above. Three positions of the curve are shown marked 0, + and *. Initially the relevant point of intersection is *B* a stable prosperous stable equilibrium point: *K* and *S* mark the currently irrelevant *crisis* and slump equilibrium points on this curve. During the boom the curve moves down to the position + + + at which *B* and *K* meet at the point *K*+ of tangency and the curve loses contact with the horizontal equilibrium axis so that the relevant equlibrium shifts rapidly to *D**, the slump stable equilibrium, and now the curve moves upwards past position 0 to position +, at which *S* and *K* coalesce at the new point *K** of tangency, and again the curve loses contact with the equilibrium line, so that now the relevant



Divided Populations and Stochastic Models, Fig. 1 Phase diagram with crisis point K

Divided Populations and Stochastic Models, Table 1 Three lines of bequests over ten generations

Generation number	1	2	3	4	5	6	7	8	9	10
Level of bequest										
Line 1	70.0	61.3	45.1	24.0	24.0	24.0	24.0	24.0	24.0	24.0
Line 2	100	117	150	210	324	443	518	565	593	611
Line 3	1000	865	780	728	695	675	662	654	649	646



Divided Populations and Stochastic Models, Fig. 2 Phase diagram: three curve positions

stable equilibrium moves rapidly to the right to B^*, the boom stable equilibrium. Then the curve moves downwards again through the position 000 and the story is repeated again and again with alternative stable equilibria B and S in boom and in slump.

Since World War II a number of models have been based on such non-linear differential or difference equations to produce fairly regular switching between temporarily stable situations of slump and boom. An early and particularly neat example was provided by R. M. Goodwin in a paper delivered in June 1955 to a meeting of the International Economic Society in Oxford. An account of that and other such early models will be found in chapter 8 of *Mathematical Economics* by R. G. D. Allen (London: Macmillan; New York: St Martin's Press, 1956).

The model in section "Rules for a Model Generating Lines of Bequests" with its crisis point has much in common with those of Kaldor and Goodwin and later writers, but in section "Easy Rules for a Stochastic Model of a Divided Population" we shall develop it in a different direction by introducing stochastic disturbances so that it may be used for modelling the development of bimodal distributions. The point which that model is intended to illustrate is that a few simple ingredients which may underlie a number of complex situations in which bimodal frequency distributions may alone be sufficient to produce bimodality, without any of the many further influences which may also be possible explanations of it. It is quite plain that such a model is not in fact a complete explanation, but it may be helpful as illustrating a method of taking a first step in a variety of investigations of situations where divided populations are observed.

The simple ingredients alluded to above are as follows:

(1) A set of largely unidentifiable and unexpected disturbances to each member value of the population whose distribution is being generated. This may well increase the dispersion.
(2) A set of influences encouraging the growth of large member values and the declines of small member values.
(3) Opposing these influences, 1 and 2: specific measures taken to discourage further growth

of very large member values and reverse the fall of very small ones.

These three ingredients will often be sufficient to produce a bimodal distribution. We have not included in (3) the many influences that there may be operating to diminish or reverse the effects of ingredient (2) at intermediate levels: where this omitted set of influences is strong, a unimodal distribution is likely to be found.

Rules for a Model Generating Lines of Bequests

In this section we consider a population consisting of family lines within which bequests are passed down generation after generation according to a mechanical system of rules governing inheritance, earnings, consumption, taxation, subsidies and dividends, and which lead all family lines eventually to ruin or to considerable wealth.

The same rules apply to each and every line of bequests, which differ only in the level of the initial bequest. We may denote the level of the initial bequest in a representative line as B_0 and the level of the bequest in that line t generations later as B_t. We shall set out in the next paragraph a set of rules which entail that the following bequest, B_{t+1} in the line is always obtainable from one or two linear equations from the current bequest B_t. For reference these equations (1), (2), are set out below and followed by an explanation of the notation and by a description of the rules governing the accumulation of wealth for bequests and implying these equations.

When $B_t < WEX$,

$$B_{t+1} = e^{\mathrm{RT}}.[B_t - (C-E)/R] + (C-E)/R$$

if B_t exceeds P but otherwise

$$B_{t+1} = e^{\mathrm{RT}}.[P - (C-E)/(R)] + (C-E)/R \tag{1}$$

or if this is <0, $\mathrm{B}_{t+1} = 0$.

When $B_t > WEX$,

$$B_{t+1} = e^{RT}.[B_t - TAX.(B_t - WEX) - (C-E)/R] + (C-E)/R \tag{2}$$

Both (1) and (2) operate if $B_t = WEX$.

The meanings of the symbols *T, E, C, R, TAX* and *WEX* are as follows: T = length of generation in years: we take $T = 25$ for examples, E = level of earnings per annum: we take $E = 10$ for examples, C = consumption expenditure per annum: $C = 12$ for examples, R = interest rate for dividends per annum: $R = 2.5\%$ for examples, P = level up to which bequests less than it are subsidized, $P = 50$ for examples, TAX = rate of tax of bequests starting at exemption level WEX for tax on bequests: $TAX = 2/3$ or 3/5; $WEX = 250$ or 400, in examples.

The four rules which lead to the equations (1) and (2) are:

Rule 1. So long as any of a bequest remains it constitutes a fund attracting interest at the rate R per annum and provides a source from which the excess expenditure $(C - E)$ can be maintained.

Rule 2. If the whole of a bequest gets used up before the end of a generation, consumption is cut from C to E, (debt is ruled out) and in this case the bequest (before subsidy) must be zero.

Rule 3. Every bequest consists of the accumulated fund at retirement (before tax or subsidy), which fund may be zero.

Rule 4. The tax or subsidy on the bequest B_t is applied at the moment of payment to the heir, so that the heir receives W_t out of $B_t < P$ where $W_t = B_t - TAX.(B_t - WEX)$ if $B_t > WEX$, $W_t = P$ if $B_t < P$ and $W_t = B_t$ otherwise.

If we denote the value of the fund after u years by $F(u)$, the derivation of equations (1) and (2) follows directly from the rules by solving the differential equation $\mathrm{d}F/Fu = R \cdot F(u) - C + E$ by standard methods to obtain $F(T)$ given $F(0) = W_t$ which may be found by Rule 4.

The equations (1), (2) and a knowledge of the values of the parameters *T, E, C, R* etc. and of the initial bequest B_0 of any line now enable us to

derive the whole line of bequests $B_0, B_1, B_2, \ldots$ as far as we wish and to find the limiting value in long-run equilibrium, by repeated application of the relevant equations.

The operation of the model can be illustrated by a phase diagram if we select values for the parameters.

Putting

$$R = 2.5\%, T = 25, C = 12, E = 10, \quad TAX = 2/3, WEX = 250, P = 50 \tag{3}$$

we obtain, when $B_t < 250$,

$$B_{t+1} - B_t = 0.868246(B_t - 80), \quad \text{if } B_t > 40 \tag{4}$$

but if $B_t < 40$;

$$B_{t+1} - B_t = 23.95 - B_t. \tag{5}$$

But when $B_t > 250$,

$$B_{t+1} - B_t = -0.377251B_t + 241.91465 \tag{6}$$

and can derive the phase diagram, Fig. 1 in which the curve cuts the horizontal axis in the three equilibrium points at T, K and S at which $B_t = 23.95$, 80 and 641.26. The values of the turning points J and L are those of P and *WEX,* 50 and 250 (see section "Some Standard Methods for the Study of Unstable Situations").

Table 1 covers ten generations and gives the values of the bequests in three lines starting at 70, 100 and 1000.

Statistical Methods for Studying Distributions with Many Peaks

Twin-peaked distributions often arise in situations where there are three equilibrium points of which two are stable, but the third is unstable, and lies between them. In such unstable situations the fact that an initial distribution contains individuals on both sides of the unstable crisis point will ensure that the population will eventually be divided into two groups at or close to the two stable equilibrium points.

The same mathematical device that underlies the crisis models generating alternative progressions to the two stable equilibrium points, or in some models regular switching from one to the other across an unstable one, may be adapted to represent situations which produce a frequency distribution consisting of two peaked distributions each centred on stable equilibrium points on either side of an unstable one. The adaptation may consist of the introduction of rules for moving the curve that indicates the equilibrium points, as in Kaldor's models, or by the introduction of rules disturbing the point indicating the current state of affairs off that curve: that is the line we shall investigate.

The whole frequency distribution may either continue strictly positive across the neighbourhood of the unstable point between the two peaks, or be split into two entirely separate distributions, with the unstable point left in the gap between. In the class of models which will be discussed in section "Easy Rules for a Stochastic Model of a Divided Population" of the entry, the split version can only emerge if the rules governing the stochastic disturbances to the movements of the points (representing the individual values whose frequency distribution is being generated) do not enable individuals to arrive at or cross the unstable equilibrium point. This is a very stringent condition, but it represents an intermediate case between the stochastic model generating the unbroken two-peaked equilibrium distribution and the cruder determinist models generating a long-term equilibrium with all individual points concentrated at the two stable equilibrium points. More elaborate determinist models with lagged variables may lead to undamped regular oscillations about a single equilibrium point, but these will not be further discussed in this entry.

In the past, economists were largely concerned with the study of equilibrium positions towards which market competition and other social and economic forces would drive the economic individuals and conglomerations involved. The

particular concerns of the statisticians and econometricians were more often with the movements of those equilibrium points and with the dispersion of the individuals or groups around these points. In the simple cases where there was just a single equilibrium they might for example study the shapes of the frequency distributions of the one or more coordinates of the point, and suggest and test various theories to explain how such shapes could arise, as well as what caused the movement of the equilibrium point itself. Thus there have been theories to account for and predict the age-distributions of the populations of various territories, the size-distributions of their cities and the distribution of the shares of votes cast for a particular party in the various constituencies, and again the distribution of income, wealth and other measures of prosperity, both between individuals and between various groups of persons.

However, if stochastic disturbances can interfere with the equilibrating forces or even shift the three equilibrium points, there may be preserved a considerable spread of distribution around each stable equilibrium point, and if the stochastic disturbances are strong, there may still be movement between the two groups across the unstable equilibrium point. In the former case we should expect two separated equilibrium distributions, whose relative sizes would depend on the nature of the initial distribution, but in the latter case a single continuous but probably bimodal equilibrium distribution whose shape could well be independent of the initial distribution.

In those cases where a considerable valley between the two peaks of the long-term equilibrium distribution is preserved, the stochastic mechanism is quite different from the simple determinist explanation for such a bimodal equilibrium distribution: this determinist explanation is simply that two quite distinct populations have been juxtaposed and counted together as one population. For example, if a wealthy island were to annexe an impoverished island with roughly the same population and then compiled wealth-or income-distribution figures for the two combined, one might expect a fairly stable bimodal wealth-or income-distribution. However, with good transport between the two islands one might expect eventually that the later generations sprung from the impoverished island would acquire gradually some of the cultural and other advantages of the descendants of the wealthy islanders and, vice versa, some of the descendants of the wealthy islanders and, would be impoverished as a result of the competition of the more gifted immigrants from the other island. Thus in the long run the stochastic intermingling of the two races might make the stochastic model more relevant than the determinist analysis of the equilibrium distribution to be expected.

In section "Easy Rules for a Stochastic Model of a Divided Population" we shall explain how to introduce a stochastic variable into our determinist model of lines of bequests so as to change it into a stochastic model of the distribution of bequests in successive generations and in the following section will provide some numerical examples to suggest some questions which such models might be helpful in answering if they were suitably elaborated. These questions will be related to situations featuring an apparent contrast between two overlapping groups, 'poor' and 'rich', where the 'persons' to whom the distributions refer may be individuals or households or localities or larger groups such as whole economies.

Easy Rules for a Stochastic Model of a Divided Population

In our stochastic model we shall consider the distribution of bequests in each generation over a series of value-ranges of equal proportionate extent g. We shall suppose the top of range 0 to be P, and we shall number the ranges so that for each integer i, positive or negative, the top of range i is Pg^i. We assume that initially all bequests are at the centres of the ranges, and we impose rules to ensure that the same is true in every ensuing generation. This simplification makes rather narrow ranges desirable so as to avoid introducing considerable inaccuracy, but in illustrative examples we shall have to use wide intervals with $g = 10.^2 = 1.585$ or $10.^1 = 1.26$ so as to be able to set out the results in the space available.

The stochastic model differs from the deterministic one of section "Rules for a Model Generating Lines of Bequests", in modifying the value of B_{t+1}, there calculated from B_t. Denote that value by B_{t+1}(i) when B_t is at the centre of range i, then in the stochastic model we multiply it by a stochastic multiplier $\mathbf{m}_i$, which when $N = 2$ scatters the bequests of $B_{t+1}(i)$ to the centres of 4 (or more generally, of $N + 2$) consecutive ranges, in proportions that leave their arithmetic mean equal to $B_{t+1}(i)$.

In section "Rules Setting the Probabilities of the N + 2 Values of mi" the rules for choosing the ranges and the proportions of bequests moved to each of them will be set out but nonspecialists may prefer to skip that section and be content with Figs. 3 and 4 below, which illustrate typical effects of the multiplier with (i) $g = 1.585$, $N = 2$ and (ii) $g = 1.26$, $N = 7$.

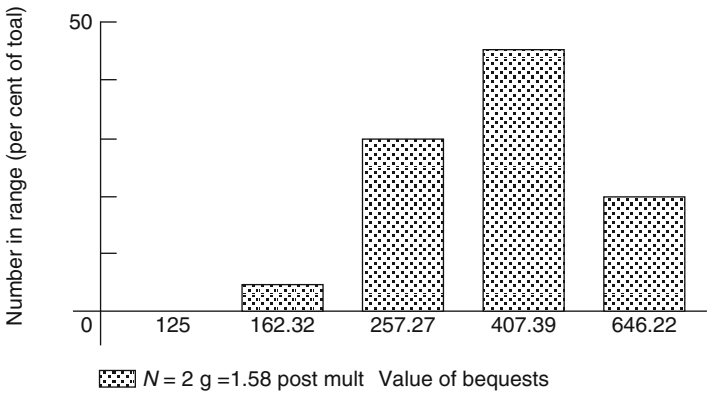
We may then work out for each range i from 0 upwards, the following bequest $B_{t+1}(i)$ by the formulae (1) and (2) of the determinist model and apply the stochastic multiplier to the bequests in lines from each range i, so as to split them into sets going, when $N = 2$ to the 4 (or more generally, $N + 2$) appropriate consecutive ranges. By using the information for every range containing at least one bequest where we round off to the nearest integer, assuming a total number of one million bequests in each generation, it is then a matter of arithmetic, to find the size-distribution of bequests in generation $t + 1$ from that of bequests in ranges with non-negative i in generation t.

We still have to explain how to handle bequests in the ranges with $i < 1$, namely the ranges below the level P. It is again assumed that all bequests in such ranges are subsidized up to the level P. Indeed, some such egalitarian measure as this is needed if we are to avoid all bequest lines eventually becoming permanently zero or in a range well below P, except possibly for a wealthy set all considerably above the crisis, level $(C - E)/R$. So in the calculations we merely have to lump all the bequests in ranges with i less than 1 into range 0. This need not prevent there being bequests *before subsidy* in each generation in other ranges below P, and it is the distribution of ranges before subsidy that we shall calculate in examples and which are relevant to the distributions of wealth and dividends which are all available towards the retiring age.

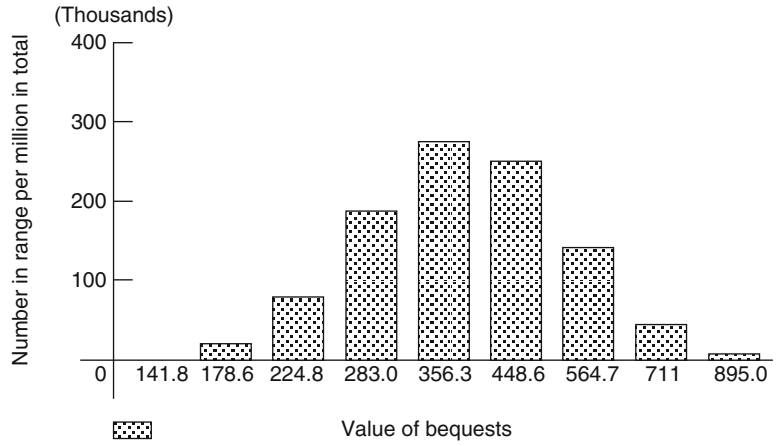
We shall give a very few numerical examples of such distributions in section "Model Generating 2-Peaked-Distribution: Illustrative Cases" to which those uninterested in the details of the rules for the stochastic multiplier are advised to skip. Those rules will now be outlined in section "Rules Setting the Probabilities of the N + 2 Values of mi".

Rules Setting the Probabilities of the N + 2 Values of mi

These rules will be illustrated by the case $N = 2$; where normally $\boldsymbol{m}_i$ may take 4 values g^j, g^{j+1}, g^{j+2}, g^{j+3}, where j is an integer. There are, however, two simple special cases where only three consecutive



Divided Populations and Stochastic Models, Fig. 3 Stochastic disturbance about 400



Divided Populations and Stochastic Models, Fig. 4 Disturbance about 400 (N = 7; g = 1.26)

integers are taken, the fourth value having no bequests dispersed to it. The rules ensure that these two cases give probabilities 0.25, 0.5, 0.25, 0 and 0, 0.25, 0.5, 0.25; the three non-zero terms are those of the binomial expansion $(0.5 + 0.5)^2$. Similarly, where N is any positive integer there are two special cases where $N + 1$ ranges only may be occupied with probabilities given by the $N + 1$ terms of the binomial expansion $(0.5 + 0.5)^N$. Returning to the special case $N = 2$, the rules further provide that the value of j and the proportions of the bequests should be so chosen that the four proportions are a weighted average with weights $1 - p$ and p $(0 < p < 1)$ of the two special cases with three terms each, and that the arithmetic mean of the bequests should equal the value $B_{t+1}(i)$ obtained in the determinist model. This entails that the four proportions should be the following: $(1 - p)/4$, $(2 - p)/4$, $(1 + p)/4$ and $p/4$. The arithmetic mean of the bequests must then be $(1 + pg)$. Pg^j $(1 + g)^2/4$ so that our rules require

$$(1 + p(g - 1)) \cdot Pg^j(1 + g)^2 = 4B_{t+1}(i) \quad (7)$$

where the right-hand side is known. This uniquely determines j and p and they may easily be derived.

In the general case where N is any positive integer the main modifications are that the binomial expansions in the special cases are now $(0.5 + 0.5)^N$ and that in equation (7) and the preceding line, 4 must be replaced by 2^N.

Model Generating 2-Peaked-Distribution: Illustrative Cases

In this section we shall illustrate the kinds of two-peaked distributions that are generated by such simplified stochastic models and the ways one might use them, by a few numerical exercises involving an imaginary set of a million lines of bequests. We shall mainly use arithmetic and diagrams for the exposition.

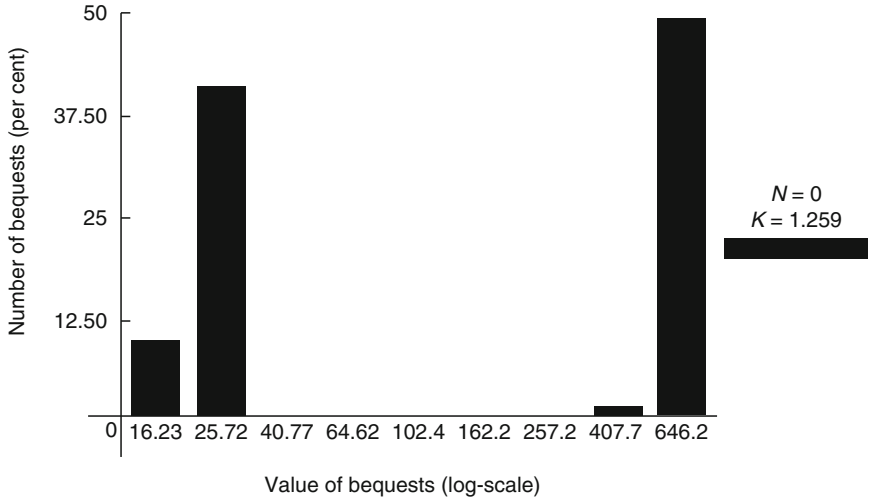
Let us start with the standard values for the parameters given in section "Rules for a Model Generating Lines of Bequests" equation (3) as

$$R = 2.5\%, T = 25, C = 12, E = 10,$$
$$TAX = 2/3 \text{ and } WEX = 250$$

and in section "Easy Rules for a Stochastic Model of a Divided Population" as

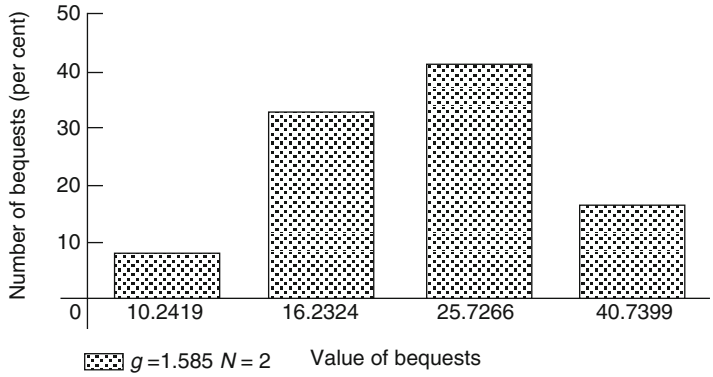
$$P = 50,\ g = 10.2 \quad (8)$$

The long-run equilibrium distribution obtained with this set-up will depend on how widely the stochastic multiplier disperses the bequests from each range in a single generation, and this is set by the parameter N. Figures 5, 6 and 7 show that widening the dispersion in this example through the first four values, $N = 0$, 1, 2 and 3 already exhibits a wide variety of types of solution. Case $N = 0$. Two separated distributions: in ranges -2

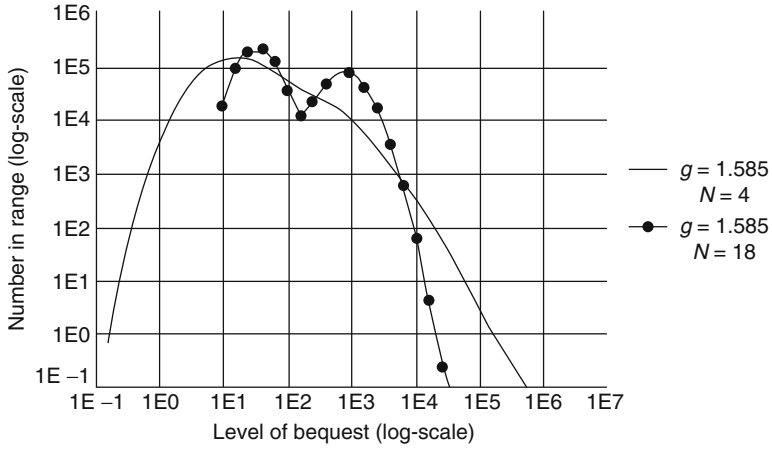


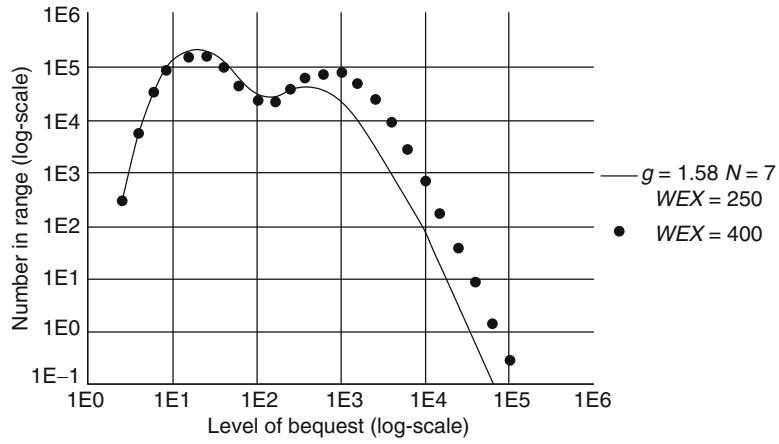
Divided Populations and Stochastic Models, Fig. 5 Equilibrium bequest–distributions

Divided Populations and Stochastic Models, Fig. 6 Equilibrium bequest distributions



Divided Populations and Stochastic Models, Fig. 7 Equilibrium Pareto Curve (non-cumulative)





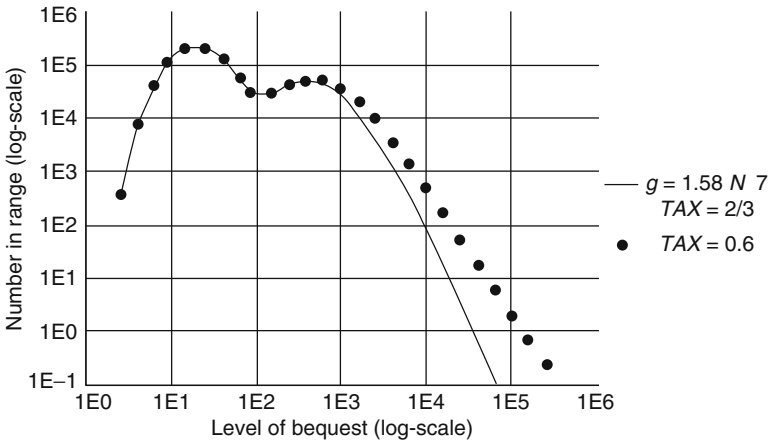
Divided Populations and Stochastic Models, Fig. 8 Effect of higher tax-exempt limit

and −1; and in ranges 10 and 11. The proportion of bequests in the two distributions will equal the initial proportions that were in ranges up to and including 1 and in ranges 2 and above. Cases $N = 2$ and $N = 3$. All bequests are in ranges immediately below P; 4 of them when $N = 2$ and 5 when $N = 3$. Cases $N = 4$ and over. All bequests are in a single distribution extending over many ranges from well below P up to well above *WEX* the tax exemption limit. These are the most interesting cases. Some have a pair of peaks separated by a valley, but those with N taking higher values have only a single peak.

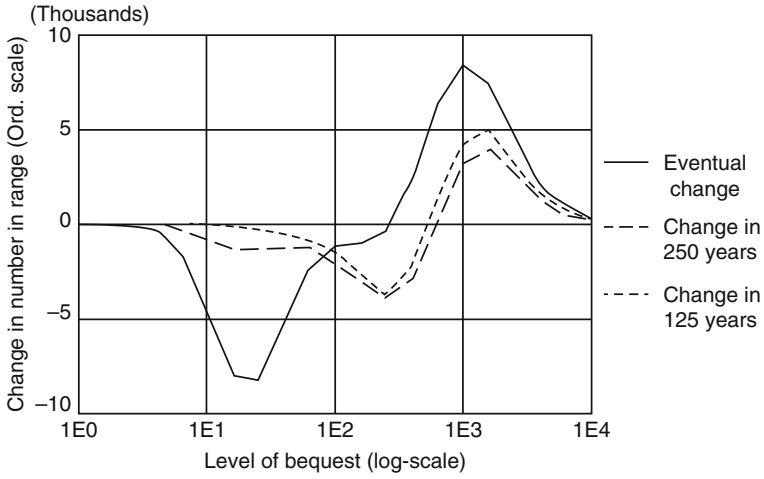
Figure 5 illustrates the case $N = 0$ where we have assumed that half the bequests in the initial generation were in ranges 3, 4, 5 etc. . . . and half in the ranges 0, −1, −2 etc. The result, due to the minimum value of N, is a very divided distribution little different from the complete division that would be found in the determinist model: the case with $N = 1$, not shown, is less stark, allowing a spread over five ranges in the upper peak and over three in the lower peak. Figure 6 shows the unusual cases where in the long term there are no bequests in any range above P. With the particular values we took for the other parameters this unusual feature only occurs when $N = 2$ and $N = 3$. Figure 7 compares the typical bimodal form when $N = 4$ with the typical unimodal form when $N = 18$. The logarithmic scale used may given the deceptive impression that when $N = 4$ the valley between the peaks is not deep and therefore easily crossed, but a more careful inspection will reveal that it is very deep, since the valley floor indicates a range with roughly 10,000 bequests, whereas even the lower peak indicates one containing roughly 100,000 bequests.

Figures 8 and 9 are mainly concerned with a potentially instructive use of stochastic models for investigating the effects of altering one or more of the policy parameters. They do this for two examples of reflationary fiscal policies: raising *WEX* the tax-emption level from 250 to 400 (Fig. 8) and altering *TAX* from 3/5 and 2/3 to 50 per cent (Fig. 9). The effects of the higher exemption level are to deepen and widen the valley and shift the tail to the right along with the level of the exemption limit, without affecting its Pareto slope. The effect of the tax reduction is mainly to lessen the Pareto slope of the tail. Each measure greatly increases the total number of bequests above 5000; the tax reduction by four or fivefold and the higher exemption level by more than tenfold. This is perhaps the right moment to reiterate the warning that such examples are not meant to be more than indications of elementary methods for testing hunches of what are the probable logical effects of such changes given any set of artificial rules being mechanically obeyed.

All our discussion has been concerned with equilibrium distributions. However, as in so many branches of economic theory, knowledge of the eventual equilibrium corresponding to the



Divided Populations and Stochastic Models, Fig. 9 Effect of lower tax on bequests



Divided Populations and Stochastic Models, Fig. 10 Short-term moves to equilibrium

present state of the economy and current policy decisions is virtually useless unless one knows how rapidly that equilibrium will be approached and particularly what will happen during the reasonably near future.

This also can be illustrated with our elementary example. It is to be expected that with low values of N, approach to equilibrium may be very slow and that is only too well confirmed by a wide variety of examples not further reported here. It is more interesting to give the stochastic disturbances considerable scattering influence by choosing quite a high value for N and then following the pace of approach towards equilibrium from an initial distribution of bequests chosen so as to differ considerably from that equilibrium distribution.

But in Fig. 10 we have taken our standard example, with $N = 7$, and shown for each range the difference between the equilibrium number of ranges that would result from altering *TAX* from 3/5 to 2/3: we also show by the two intermediate curves how much of the approach to the new equilibrium would in each range be achieved after five and after ten generations of 25 years each. It will be seen how far from completed the transition is even after the 250 years.

Space forbids showing further examples of the short-run effect of altering policy or other parameters in such models, by methods similar to those

used in studying long-run effects: naturally such short-term enquiry can be more important for obtaining conclusions remotely relevant to the real world, yet the theoretical approach to such investigations need differ little from that to the long-term ones discussed above.

Concluding Observations

In the later sections of this entry we have provided an illustration of how, without introducing any of the more detailed causes of divided, i.e. of bimodal, frequency distributions, one can obtain a skeletal model of the development of such distributions by merely including the dispersive elements: (a) an ability and willingness on the part of the richer, to save a higher proportion of their income than on the part of the poorer, and (b) stochastic disturbances; both (a) and (b) tending to increase inequality; and the egalitarian elements: (c) to curtail (a) and (b) on the part of the very rich; and (d) subsidies to set a limit on the poverty of the very poor. We have hinted how models including at least these four basic elements, and thereby generating biomodal distributions, can provide some insight into possible long-run and short-run effects of altering policy parameters: in particular we have argued that considerations of long-run equilibrium can be very poor guides to short-run effects.

The next step is, naturally, to introduce into the model the more obvious and significant other causes acting to modify distributions of wealth, income, health, nourishment and other measures of well-being. That step is far too long for inclusion in an entry of this nature. Moreover, since although situations of the critical and divisive kinds, on which such research aims to throw some light become more frequent every year, statistics of these phenomena are patchy and unreliable. The development of stochastic methods to make allowance for the unreliability and incompleteness of information is by no means a minor step in such enquiries. It was this belief that prompted the submission of this methodological entry.

Dividend Policy

David J. Denis and John J. McConnell

Abstract

Dividends represent the primary means by which invested capital is returned to common stockholders. In this article we summarize the development of academic thinking on dividend policy, focusing on three primary perspectives: (*a*) the effect of dividend policy on common stock value and firm performance, (*b*) the determinants of dividend policy, and (*c*) macroeconomic trends in the propensity of firms to pay dividends.

Keywords

Agency costs; Asymmetric information; Capital gains; Capital gains taxation; Dividend change; Dividend policy; Dividend taxation; Dividends; Manager–shareholder conflict; Modigliani–Miller theorem; Regular vs special dividends; Stock repurchases; Stockholders

JEL Classifications

G35

There are two major ways in which a firm can distribute cash to its common stockholders. The firm can either declare a cash dividend which it pays to all its common stockholders or it can repurchase shares. Stock repurchases may take the form of registered tender offers, open market purchases, or negotiated repurchases from a large shareholder. In a share repurchase, shareholders may choose not to participate. In contrast, dividends are direct cash payments to shareholders and are distributed on a pro rata basis to all shareholders.

Most firms pay cash dividends on a quarterly basis. The dividend is declared by the firm's board of directors on a date known as the 'announcement

date'. The board's announcement states that a cash payment will be made to stockholders who are registered owners on a given 'record date.' The dividend checks are mailed to stockholders on the 'payment date,' which is usually about two weeks after the record date. Stock exchange rules generally dictate that the stock is bought or sold with the dividend until the 'ex-dividend date', which is a few business days before the record date. After the ex-dividend date, the stock is bought and sold without the dividend.

Dividends may be either labelled or unlabelled. Most dividends are not given labels by management. Unlabelled dividends are commonly referred to as 'regular dividends'. When managers label a dividend, the most common label is 'extra'.

A Historical Perspective

Prior to 1961, academic treatments of dividends were primarily descriptive, as, for example, in Dewing (1953). To the extent that economists considered corporate dividend policy, the commonly held view was that investors preferred high dividend payouts to low payouts (see, for example, Graham and Dodd 1951). The only question was how much value was attached to dividends relative to capital gains in valuing a security (Gordon 1959). This view was concisely summarized with the saying that a dividend in the hand is worth two (or some multiple) of those in the bush. The only question was: what is the multiple?

In 1961, scientific inquiry into the motives and consequences of corporate dividend policy shifted dramatically with the publication of a classic paper by Miller and Modigliani. Perhaps the most significant contribution of the Miller and Modigliani paper was to spell out in careful detail the assumptions under which their analysis was to be conducted. The most important of these include the assumptions that the firm's investment policy is fixed and known by investors, that there are no taxes on dividends or capital gains, that individuals can costlessly buy and sell securities, that all investors have the same information, and that investors have the same information as the managers of the firm. With this set of assumptions, Miller and Modigliani demonstrate that a firm's stockholders are indifferent among the set of feasible dividend policies. That is, the value of the firm is independent of the dividend policy adopted by management.

Because investment policy is fixed in the Miller–Modigliani set-up, all feasible dividend policies involve the distribution of the full present value of the firm's free cash flow (that is, cash flow in excess of that required for investment) and are, therefore, equally valuable. If internally generated funds exceed required investment, the excess must be paid out as a dividend so as to hold investment constant. If internally generated funds are insufficient to fund the fixed level of investment, new shares must be sold. It is also possible for managers to finance a higher dividend with the sale of new shares.

The key insight from the Miller–Modigliani analysis is that investors will be indifferent among the feasible dividend choices because they can costlessly create their own dividend stream by buying and selling shares. If investors demand higher dividends than the amount paid by the firm, they can sell shares and consume the proceeds, leaving themselves in the same position as if the firm had paid a dividend. Alternatively, if shareholders prefer to reinvest rather than to consume, they can choose to purchase new shares with any dividends paid. In this instance, shareholders would be in the same position that that they would have been in had no dividends been paid. Thus, regardless of corporate dividend policy, investors can costlessly create their own dividend position. For this reason, stockholders are indifferent to corporate dividend policy, and, as a consequence, the value of the firm is independent of its dividend policy.

After a brief flurry of debate, the Miller–Modigliani irrelevance proposition was essentially universally accepted as correct under their set of assumptions. There nevertheless remained an underlying notion that dividend policy must 'matter' given that managers and security analysts spend time worrying about it. If so, and if the Miller–Modigliani proposition is accepted, it

must be due to violation of one or more of the Miller–Modigliani assumptions in the real world.

Since the early 1960s, the dividend debate has been lively and interesting. Economists have analysed theoretically whether the relaxation of the various Miller–Modigliani assumptions alters their irrelevance proposition. In addition, economists have analysed the data from several perspectives. First, they have undertaken an array of analyses to determine the effect, if any, of dividend policy on stock value and firm performance. Second, they have sought to identify the characteristics associated with dividend payments (or the lack thereof) by individual firms. Third, they have attempted to characterize macroeconomic trends in the level and propensity of firms to pay dividends, and in the form of the payout. Our discussion of these issues focuses primarily (though not exclusively) on studies of US firms since these are the studies most accessible to us.

Relaxing the Miller–Modigliani Assumptions

Taxes

Perhaps the obvious starting point for an investigation into the effect of relaxing the Miller–Modigliani assumptions is to introduce taxes. In the United States, dividend payments by a corporation do not affect that firm's taxes. However, at least historically, dividends have been taxed at a higher rate than capital gains at the personal level. Thus, superficially, the US tax code appears to favour a low dividend payout policy, with payouts occurring primarily through share repurchases.

Under the assumption that dividends and capital gains are taxed differentially, Brennan (1970) derives a model of stock valuation in which stocks with high payouts have higher required before-tax returns than stocks with low payouts. As a counterpoint to this proposition, Miller and Scholes (1978) argue that under the US tax code there exist sufficient loopholes so that investors can shelter dividend income from taxation, thereby driving the effective tax rate on dividends to zero. Early studies of the association between stock returns and dividend yield (for example, Black and Scholes 1974; Litzenberger and Ramaswamy 1979; Miller and Scholes 1982) yielded mixed results using different definitions of dividend yield. Subsequent studies indicated that the correlation between dividend yield and stock returns (if any) appeared to be due to omitted risk factors that were correlated with dividend yield. For example, Chen et al. (1990) report that dividend yield and risk measures are cross-sectionally correlated. Similarly, Fama and French (1993) show that, when a three-factor model for expected returns is used, there is no significant relation between dividend yields and stock returns.

Other studies have analysed the potential effects of the differential taxation of dividends and capital gains by studying the behaviour of stock prices and trading volume around ex-dividend days. The logic of these studies is that, in order for investors to be indifferent between selling a stock just before it goes ex dividend and just after, stocks should be priced so that the marginal tax liability would be the same for each strategy. Thus, if dividends are taxed more heavily than are capital gains, stock prices should fall by less than the size of the dividend on the ex-dividend day. Evidence consistent with a tax effect in stock price behaviour around ex-dividend days is provided in Elton and Gruber (1970), Eades et al. (1984), Green and Rydqvist (1999), Bell and Jenkinson (2002), and Elton et al. (2005). In addition, evidence of tax-motivated trading around ex-dividend days is provided in Lakonishok and Vermaelen (1986), Michaely and Vila (1995) and Green and Rydqvist (1999).

Collectively, the evidence in these studies indicates that the differential taxation of dividends and capital gains affects both ex-dividend day stock returns and trading activity. This conclusion has been reinforced in studies that examine changes in tax laws (for example, Poterba and Summers 1984; Barclay 1987; Michaely 1991). Nonetheless, the fact that individual investors in high tax brackets receive large amounts of taxable dividends each year (Allen and Michaely 2003) casts doubt on taxes being a first-order determinant of dividend policy.

Agency Costs

A second real-world violation of the Miller–Modigliani assumptions is the existence of agency costs associated with stock ownership. In particular, managers of firms maximize their own utility, which is not necessarily the same as maximizing the market value of common stock. The costs associated with this potential conflict of interest include expenditures for structuring monitoring and bonding contracts between shareholders and managers, and residual losses due to imperfectly constructed contracts (Jensen and Meckling 1976).

Several authors have argued that dividends may be important in helping to resolve manager–shareholder conflicts. If dividend payments reduce agency costs, firms may pay dividends even if these payments are taxed disadvantageously.

Easterbrook (1984) and Rozeff (1982) argue that establishing a policy of paying dividends enables managers to be evaluated periodically by the capital market. By paying dividends, managers are required to tap the capital market more frequently to obtain funds for investment projects. Periodic review by the market is one way in which agency costs are reduced, which in turn raises the value of the firm. Similarly, Jensen (1986) argues that establishing a policy of paying dividends reduces agency problems of overinvestment by reducing the amount of discretionary cash controlled by managers.

An implication of the agency models is that dividends will be more valuable in mature firms with substantial cash flow and poor investment opportunities. Early tests of this implication focused on the stock price reaction to dividend change announcements and produced mixed results. Lang and Litzenberger (1989) find that firms with less valuable growth opportunities exhibit a larger stock price reaction to dividend increase announcements than firms with more valuable growth opportunities. Although this finding is consistent with the agency cost hypothesis, Denis et al. (1994) find that when they control for other factors, particularly the change in dividend yield, they find no difference in the stock price reaction to dividend changes between firms with good growth opportunities and those with poor growth opportunities. Moreover, they find no evidence that increases in dividends reduce corporate investment.

More recent tests of the agency models have focused on the cross-sectional determinants of dividend policy. Fama and French (2001) find that the propensity to pay dividends is positively related to firm size and profitability, and negatively related to the value of future growth opportunities. DeAngelo et al. (2006) find that the propensity to pay dividends is strongly associated with the proportion of the firm's equity that comes from retained earnings. These findings support the primary prediction of the agency models that dividends are more valuable for mature firms with high cash flow and poor growth opportunities.

La Porta et al. (2000) and Faccio and Lang (2002) provide further support for the agency models of dividend policy by analysing international evidence. La Porta et al. hypothesize that agency conflicts will differ across countries because of differences in the extent of investor protection. In a sample of 33 different countries, they find that dividend payments are higher in countries with better investor protection. This indicates that when investors are better able to monitor managers, they are able to force higher dividend payouts. Faccio and Lang (2002) show that in western Europe and in Asia dividend payments are higher when controlling shareholders have a higher ratio of voting rights to cash flow rights – that is, those situations in which minority shareholders are otherwise at greatest risk of expropriation by the controlling shareholder.

Asymmetric Information

Contrary to the Miller–Modigliani assumption that investors have the same information as managers, a large number of studies assume that managers possess more information about the prospects of the firm than individuals outside the firm, and that dividend changes convey this information to outsiders. This idea was suggested by Miller and Modigliani and has roots in Lintner's (1956) classic study on dividend policy. Lintner interviewed a sample of corporate managers. One of the primary findings of the interviews is that a

high proportion of managers attempt to maintain a stable regular dividend. In Lintner's words, managers demonstrate a 'reluctance (common to all companies) to reduce regular rates once established and a consequent conservatism in raising regular rates' (1956, p. 84).

If managers change regular dividends only when the earnings potential of the firm has changed, changes in regular dividends are likely to provide some information to the market about the firm's prospects. More formal models in which dividends convey information to outsiders include Bhattacharya (1979 1980), John and Williams (1985), and Miller and Rock (1985). The common assumption in these models is that managers have information not available to outside investors. Typically, the information has to do with the current or future earnings of the firm.

Empirical evidence on the information content of dividends has taken three forms. First, a large set of studies has analysed whether dividend changes are associated with abnormal stock returns of the same sign. Second, studies have analysed whether dividend changes are associated with subsequent earnings changes. Third, studies have analysed the association between dividend changes and changes in investor expectations regarding future earnings.

Studies have consistently documented that stock returns around the announcement of a dividend change are positively correlated with the change in the dividend (Aharony and Swary 1980; Asquith and Mullins 1983; Brickley 1983; Healy and Palepu 1988; Grullon et al. 2002; Michaely et al. 1995; Pettit 1972). These studies are robust over time and are robust to controls for contemporaneous earnings announcements. Moreover, in general, the studies indicate that the market reacts more strongly to a dividend decrease than to a dividend increase.

The findings described above indicate that dividend announcements provide information to the market. Subsequent studies have investigated whether this information is correlated with current or future earnings. On this issue, the evidence is more mixed. In a study of dividend initiations and omissions, Healy and Palepu (1988) find that the initiation of dividends follows a period of abnormal earnings growth and that earnings continue to grow in subsequent years. For omissions, however, earnings decline in the year of omission, then rebound in the following years. Using a comprehensive sample of dividend changes, Benartzi et al. (1997) find no evidence that dividend changes are associated with subsequent earnings changes of the same sign. Miller's interpretation of the evidence (1987) is that dividends appear to be better described as lagging earnings than as leading earnings.

One difficulty in testing whether dividend changes 'signal' unexpected future earnings is that it is difficult to identify what level of earnings would be expected by the market if the dividend change did not take place. To address this issue, Ofer and Siegel (1987) study how analysts alter their estimates of current year earnings when firms announce dividend changes. They find that analysts revise their earnings estimates in the direction of the dividend change and that the size of the earnings revision is positively associated with the stock price reaction to the dividend change. Similarly, Fama and French (1998) report a positive association between dividends and firm value after controlling for past, current and future earnings, as well as investment and debt. They conclude that dividends contain information about value that is not contained in earnings, investment and debt.

The accumulated empirical evidence thus indicates that dividend announcements provide information to the market. Whether they convey information about future earnings is less clear. Moreover, other findings indicate that information signalling is unlikely to be a first-order determinant of dividend policy. For example, as noted earlier, dividends are paid primarily by larger, more mature firms with higher cash flow and poorer growth opportunities. These types of firm would seem to be least in need of signalling their true value to the market.

Firm Value and the Form of the Payout

As with increases in regular cash dividends, specially labelled cash dividends and share repurchases have been shown to be accompanied

by permanent increases in stock prices (Brickley 1983; Dann 1981; Vermaelen 1981). However, there is little agreement on the factors that lead managers to choose one method over another.

Given the Miller–Modigliani assumptions, the choice of the payout mechanism, like the choice of dividend policy itself, does not affect the value of the firm. Therefore, if the form of the payout is to matter, it must be due to violation of one or more of the Miller–Modigliani assumptions. To develop a theory to explain the choice of payout mechanism, it must be that there are differential costs or benefits associated with the alternative payout methods. Furthermore, the relative benefits or costs must be especially significant because, in general, dividends have been tax-disadvantaged (at the personal level) relative to share repurchases.

Economists have explored several possible explanations as to why a particular form of payout is chosen, including adverse selection effects (Barclay and Smith 1988; Miller and McConnell 1995), the impact on equity ownership structure (Stulz 1988; Denis 1990), the signalling power of alternative payout mechanisms (Ofer and Thakor 1987; Jagannathan et al. 2000), and the impact of executive stock options (Fenn and Liang 2001). The evidence indicates that share repurchases are more likely when recent earnings increases are temporary, when earnings are riskier, when firms make heavy use of stock options in executive compensation contracts and when firms seek to protect themselves from a hostile takeover.

As regards the choice between regular cash dividends and specially labelled cash dividends, reasonable explanations have been relatively scarce. Brickley (1983) does provide evidence that specially labelled dividends convey a less positive message about firm value than do increases in regular cash dividends. Nonetheless, it is unclear why this is so. Moreover, there has been little examination of the choice between special dividends and share repurchases.

What Managers Say

Lintner's (1956) classic empirical study began with a survey of corporate executives. The results of that survey and the accompanying evidence laid the foundation for much of the empirical and theoretical work that has followed over the succeeding half century. Brav et al. (2005) have conducted a new and more extensive survey of chief financial officers (CFOs) regarding their views of corporate payout policy. Their survey yields further insights into what managers think about dividend policy, and complements the existing empirical evidence.

Brav et al. report that CFOs view dividends as inflexible in that, once a dividend level has been established, any dividend cut is likely to have a significantly adverse impact on the company's stock price. Thus, consistent with Lintner's (1956) original observation, managers tend to be conservative when adjusting dividends upward in order to avoid having to cut the dividend at a later date. Rather than establishing a target payout ratio, managers set a per share payment that is downwardly inflexible. According to the survey, managers do not explicitly view dividends as a mechanism for signalling information that would distinguish their companies from competitors, and they consider tax effects only as an afterthought. These observations accord with the conclusions drawn from empirical studies in that both imply that taxes and signalling are not first-order determinants of dividend policy.

In contrast to dividends, repurchases are viewed by managers as a parallel but more flexible way to distribute cash to shareholders in that they can be initiated and discontinued as funds are available. This observation is consistent with the empirical evidence cited earlier that repurchases tend to be associated with temporary increases in earnings, while dividends are associated with earnings changes that are more permanent. Whether the modern survey of Brav et al. leads to the volume of additional empirical work that followed Lintner's study remains to be seen.

Summary and Recent Trends

Since the mid-1960s, rigorous consideration has added considerably to progress in what is known

about dividend policy. We know that firms pay out to stockholders substantial amounts of cash annually in the form of regular cash dividends, share repurchases and specially labelled dividends. We also know that stock prices increase permanently when regular dividends are increased, when special dividends are declared, and when shares are repurchased, and that stock prices decline when regular dividends are reduced. While these findings imply that dividend changes reflect information available to managers that is not otherwise available to outside investors, it is still not clear what information is being conveyed through the dividend payment. Moreover, although we now know a considerable amount about the empirical determinants of the size of payout and the form of payout, there is little agreement as to whether the level of cash payout affects the value of the firm or and whether the choice of the payout method matters.

We conclude by outlining several recent trends that pose additional challenges to our understanding of dividend policy. First, Fama and French (2001) document that the propensity to pay dividends has declined substantially since the late-1970s. Second, despite this decline in the propensity to pay dividends, aggregate dividends have not declined (DeAngelo et al. 2004). Rather, dividends and earnings have become increasingly concentrated among larger firms. Third, specially labelled dividends have nearly disappeared (DeAngelo et al. 2000). Fourth, share repurchases have increased substantially so that aggregate payouts through share repurchases now exceed those through regular dividends (Grullon and Michaely 2002). These trends are difficult to explain given our current understanding of dividend policy. Undoubtedly, therefore, economists will continue to devote substantial effort to understanding the puzzles of dividend policy.

See Also

- ▶ Finance (New Developments)
- ▶ Modigliani–Miller Theorem

Bibliography

Aharony, J., and I. Swary. 1980. Quarterly dividend and earnings announcements and stockholders' returns: An empirical analysis. *Journal of Finance* 35: 1–12.

Allen, F., and R. Michaely. 2003. Payout policy. In *Handbook of the economics of finance: volume 1a*, ed. G. Constantinides, M. Harris, and R. Stulz. Amsterdam: North-Holland.

Asquith, P., and D. Mullins. 1983. The impact of initiating dividend payments on shareholders' wealth. *Journal of Business* 56: 77–96.

Barclay, M. 1987. Dividends, taxes, and common stock prices: The ex-dividend day behavior of common stock prices before the income tax. *Journal of Financial Economics* 14: 31–44.

Barclay, M., and C. Smith. 1988. Corporate payout policy: Cash dividends versus open-market repurchases. *Journal of Financial Economics* 22: 61–82.

Bell, L., and T. Jenkinson. 2002. New evidence of the impact of dividend taxation and on the identity of the marginal investor. *Journal of Finance* 57: 1321–1346.

Benartzi, S., R. Michaely, and R. Thaler. 1997. Do changes in dividends signal the future or the past? *Journal of Finance* 52: 1007–1043.

Bhattacharya, S. 1979. Imperfect information, dividend policy, and 'the bird in the hand' fallacy. *Bell Journal of Economics* 10: 259–270.

Bhattacharya, S. 1980. Nondissipative signaling structures and dividend policy. *Quarterly Journal of Economics* 95: 1–24.

Black, F., and M. Scholes. 1974. The effects of dividend yield and dividend policy on common stock prices and returns. *Journal of Financial Economics* 1: 1–22.

Brav, A., J. Graham, R. Michaely, and C. Harvey. 2005. Payout policy in the 21st century. *Journal of Financial Economics* 77: 483–527.

Brennan, M. 1970. Taxes, market valuation and financial policy. *National Tax Journal* 23: 417–429.

Brickley, J. 1983. Shareholders wealth, information signaling, and the specially designated dividend: An empirical study. *Journal of Financial Economics* 12: 187–209.

Chen, N., B. Grundy, and R. Stambaugh. 1990. Changing risk, changing risk premiums, and dividend yield effects. *Journal of Business* 63: S51–S70.

Dann, L. 1981. Common stock repurchases: An analysis of returns to bondholders and stockholders. *Journal of Financial Economics* 9: 113–138.

DeAngelo, H., L. DeAngelo, and D. Skinner. 2000. Special dividends and the evolution of dividend signaling. *Journal of Financial Economics* 57: 309–354.

DeAngelo, H., L. DeAngelo, and D. Skinner. 2004. Are dividends disappearing? Dividend concentration and the consolidation of earnings. *Journal of Financial Economics* 72: 425–456.

DeAngelo, H., L. DeAngelo, and R. Stulz. 2006. Dividend policy and the earned/contributed capital mix: A test of

the life-cycle theory. *Journal of Financial Economics* 81: 227–254.
Denis, D. 1990. Defensive changes in corporate payout policy: Share repurchases and special dividends. *Journal of Finance* 45: 1433–1456.
Denis, D., D. Denis, and A. Sarin. 1994. The information content of dividend changes: Cash flow signaling, overinvestment, and dividend clienteles. *Journal of Financial and Quantitative Analysis* 29: 567–587.
Dewing, A. 1953. *The Financial policy of corporations*. Vol. 2. 5th ed. New York: Ronald Press Co.
Eades, K., P. Hess, and E. Kim. 1984. On interpreting security returns during the ex-dividend period. *Journal of Financial Economics* 13: 3–34.
Easterbrook, F. 1984. Two agency–cost explanations of dividends. *American Economic Review* 74: 650–659.
Elton, E., and M. Gruber. 1970. Marginal stockholders' tax rates and the clientele effect. *Review of Economics and Statistics* 52: 68–74.
Elton, E., M. Gruber, and C. Blake. 2005. Marginal stockholder tax effects and ex-dividend day behavior: Evidence from taxable versus nontaxable closed-end funds. *Review of Economics and Statistics* 87: 579–586.
Faccio, M., and L. Lang. 2002. The ultimate ownership of western European corporations. *Journal of Financial Economics* 65: 365–395.
Fama, E.F., and K. French. 1993. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* 33: 3–56.
Fama, E.F., and K. French. 1998. Taxes, financing decisions, and firm value. *Journal of Finance* 53: 819–843.
Fama, Eugene, and K. French. 2001. Disappearing dividends: Changing firm characteristics or lower propensity to pay? *Journal of Financial Economics* 60: 3–43.
Fenn, G., and N. Liang. 2001. Corporate payout policy and managerial stock incentives. *Journal of Financial Economics* 60: 45–72.
Gordon, M. 1959. Dividends, earnings and stock prices. *Review of Economics and Statistics* 41: 99–105.
Graham, B., and D. Dodd. 1951. *Security analysis: Principles and technique*. New York: McGraw-Hill.
Green, R., and K. Rydqvist. 1999. Ex-day behavior with dividend preference and limitation to short-term arbitrage: The case of Swedish lottery bonds. *Journal of Financial Economics* 53: 145–187.
Grullon, G., and R. Michaely. 2002. Dividends, share repurchases and the substitution hypothesis. *Journal of Finance* 57: 1649–1684.
Grullon, G., R. Michaely, and B. Swaminathan. 2002. Are dividend changes a sign of firm maturity? *Journal of Business* 75: 387–424.
Healy, P., and K. Palepu. 1988. Earnings information conveyed by dividend initiations and omissions. *Journal of Financial Economics* 21: 149–176.
Jagannathan, M., C. Stephens, and M. Weisbach. 2000. Financial flexibility and the choice between dividends and stock repurchases. *Journal of Financial Economics* 57: 355–384.
Jensen, M. 1986. Agency costs of free cash flow, corporate finance, and takeovers. *American Economic Review* 76: 323–329.
Jensen, M., and W. Meckling. 1976. Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics* 3: 305–360.
John, K., and J. Williams. 1985. Dividends, dilution, and taxes: A signaling equilibrium. *Journal of Finance* 40: 1053–1070.
La Porta, R., F. Lopez-De Silanes, A. Shleifer, and R. Vishny. 2000. Agency problems and dividend policy around the world. *Journal of Finance* 55: 1–33.
Lakonishok, J., and T. Vermaelen. 1986. Tax induced trading around ex-dividend dates. *Journal of Financial Economics* 16: 287–319.
Lang, L., and R. Litzenberger. 1989. Dividend announcements: Cash flow signaling vs. free cash flow hypothesis. *Journal of Financial Economics* 24: 181–192.
Lintner, J. 1956. Distribution of incomes of corporations among dividends, retained earnings, and taxes. *American Economic Review* 46: 97–113.
Litzenberger, R., and K. Ramaswamy. 1979. The effects of personal taxes and dividends on capital asset prices: Theory and empirical evidence. *Journal of Financial Economics* 7: 163–195.
Michaely, R. 1991. Ex-dividend day stock price behavior: The case of the 1986 tax reform act. *Journal of Finance* 46: 845–860.
Michaely, R., and J. Vila. 1995. Investors' heterogeneity, prices and volume around the ex-dividend day. *Journal of Financial and Quantitative Analysis* 30: 171–198.
Michaely, R., R. Thaler, and K. Womack. 1995. Price reactions to dividend initiations and omissions: Overreaction or drift? *Journal of Finance* 50: 573–608.
Miller, M. 1987. The information content of dividends. In *Macroeconomics: Essays in honor of Franco Modigliani*, ed. J. Bossons, R. Dornbush, and S. Fischer. Cambridge, MA: MIT Press.
Miller, J., and J. McConnell. 1995. Open-market share repurchase programs and bid-ask spreads on the NYSE: Implications for corporate payout policy. *Journal of Financial and Quantitative Analysis* 30: 365–382.
Miller, M., and F. Modigliani. 1961. Dividend policy, growth and the valuation of shares. *Journal of Business* 34: 411–433.
Miller, M., and K. Rock. 1985. Dividend policy under asymmetric information. *Journal of Finance* 40: 1031–1051.
Miller, M., and M. Scholes. 1978. Dividends and taxes. *Journal of Financial Economics* 6: 333–364.
Miller, M., and M. Scholes. 1982. Dividends and taxes: Empirical evidence. *Journal of Political Economy* 90: 1118–1141.
Ofer, A., and D. Siegel. 1987. Corporate financial policy, information, and market expectations: An empirical investigation of dividends. *Journal of Finance* 42: 889–911.
Ofer, A., and A. Thakor. 1987. A theory of stock price responses to alternative corporate cash disbursement

methods: Stock repurchases and dividends. *Journal of Finance* 42: 365–394.
Pettit, R. 1972. Dividend announcements, security performance, and capital market efficiency. *Journal of Finance* 27: 993–1007.
Poterba, J., and L. Summers. 1984. New evidence that taxes affect the valuation of dividends. *Journal of Finance* 39: 1397–1415.
Rozeff, M. 1982. Growth, beta and agency costs as determinants of dividend payout ratios. *Journal of Financial Research* 5: 249–259.
Stulz, R. 1988. Managerial control of voting rights: Financing policies and the market for corporate control. *Journal of Financial Economics* 20: 25–54.
Vermaelen, T. 1981. Common stock repurchases and market signaling: An empirical study. *Journal of Financial Economics* 9: 139–183.

Divisia Index

Charles R. Hulten

Abstract

The Divisia index, it its modern application, is a continuous-time index related to an underlying economic structure via a potential function. Under certain conditions, the index can retrieve important characteristics of the underlying structure using prices and quantities alone, without full knowledge about the structure itself. The Divisia index is widely used in theoretical discussions of productivity analysis, and has important applications elsewhere. In practice, it is approximated by discrete–time superlative indexes, like the Tornqvist, or by chain indexes. Older applications of the Divisia stressed its discrete-time axiomatic properties.

Keywords

Aggregation; Chain indexes; Continuous–time indexes; Discrete–time indexes; Divisia index; Divisia, F.; Duality; Path dependence; Production functions; Productivity (measurement problems); Solow, R.; Törnqvist index

JEL Classifications

C43; C80; E01

The Divisia index is a continuous-time index number formula due to François Divisia (1925–6) that has been widely used in theoretical discussions of data aggregation and the measurement of technical change. It is defined with respect to the time paths of a set of prices $[P_1(t), \ldots, P_N(t)]$ and commodities $[X_1(t), \ldots, X_N(t)]$. Total expenditure on this group of commodities is given by:

$$Y(t) = P_t(t)X_1(t) + \ldots + P_N(t)X_N(t). \quad (1)$$

With dots over variables indicating derivatives with respect to time, total differentiation of (1) yields:

$$\frac{\dot{Y}(t)}{Y(t)} = \sum_{i=1}^{i=N} \frac{P_i(t)X_i(t)}{Y(t)} \frac{\dot{P}_i(t)}{P_i(t)} + \sum_{i=1}^{i=N} \frac{P_i(t)X_i(t)}{Y(t)} \frac{\dot{X}_i(t)}{X_i(t)}. \quad (2)$$

The growth rates of the Divisia price and quantity indexes are the respective weighted averages of the growth rates of the individual $P_i(t)$ and $X_i(t)$, where the weights are the components' shares in total expenditure. The levels of these indexes are obtained by line integration over the trajectory followed by the individual prices and quantities over the time interval $[0, T]$. For the quantity index, the line integral has the following form:

$$I_q(0,T) = \exp\left\{ \int \left[\sum_{i=1}^{N} \frac{P_i(t)X_i(t)}{\sum P_j(t)X_j(t)} \frac{\dot{X}_i(t)}{X_i(t)} \right] \right\} = \exp\left\{ \int_r \varphi(X) \mathrm{d}X \right\}, \quad (3)$$

where φ is a vector-valued function whose arguments are $P_i(t)/Y(t)$, prices are assumed to be a function of the X_i, and Γ is the curve described by X_i. A similar expression characterizes the Divisia price index (for a more extensive discussion of Divisia line integrals, see Richter 1966; Hulten 1973; Samuelson and Swamy 1974).

The value of the index defined by (3) depends on the solution of the line integral. This can be obtained by identifying a 'potential function' Φ

whose partial derivatives are the vector-valued function φ, that is, $\varphi = \nabla\Phi$. Writing $\Phi = \log F$ function, the value of the index can be shown to equal F[X(T)]/F[X(0)], implying that the index is unique only up to a scalar multiple.

In economic terms, the solution to (3) is associated with some underlying economic relationship among the variables being indexed. Assume, for example, there is a constant returns to scale production function F(X) and $F_i = \lambda P_i$ (Fi denotes the partial derivative of F with respect to Xi and λ is a factor of proportionality). Then the function log F can serve as the requisite potential function for (3), and in this particular case, the Divisia index of inputs can be interpreted as the ratio of output at time T to output at time zero.

If the form of the potential function is known a priori, the value of the index could be computed directly from the function F. However, the rationale for the Divisia index is that it provides a way of obtaining the ratio F(X(T))/F(X(0)) by using data on prices and quantities alone, without direct knowledge of F. Intuitively, this is possible because, under sufficiently restrictive assumptions, information about the slope of the function F (as estimated by relative prices) over the path followed by the inputs is sufficient to characterize F up to a scalar multiple.

When the objective is to form an index of a subset of inputs – aggregate labour input, for example – the required potential function is a 'piece' of a production function. Specifically, if one wants to form a Divisia index of the first M inputs, the production function needs to be weakly separable into a function of these inputs, that is, $F\{G[X_1(t),\ldots,X_M(t)], X_{M+1}(t),\ldots,X_N(t)\}$. The function log G serves as the potential function for the line integration (see also Balk 2005).

These considerations apply to Divisia price indexes as well. The relevant potential function is now the factor price frontier $\Psi[P_1(t),\ldots,P_N(t)]$. A basic result of duality theory shows that the partial derivatives of Ψ are proportional to the corresponding $X_i(t)$.

The discussion suggests that the existence of the Divisia index is closely linked to the conditions for consistent aggregation. Furthermore, the required existence of a potential function implies that aggregation cannot proceed with just any set of prices or quantities. There must be an a priori reason for supposing that the variables to be indexed are theoretically related. This is an important characteristic of the Divisia index, one which it shares with the broader class of economic index numbers (in contrast to the non-structural axiomatic approach associated with Irving Fisher 1921; see also Balk 2005). The potential function theorem establishes the conditions under which the Divisia index is an 'exact' index number (to use the terminology of Diewert 1976) for some underlying economic structure.

Divisia indexes have the desirable property that they are invariant when the path of integration lies entirely in the same level set of the potential function. That is, ifone input is substituted for another along a given isoquant, the value of the index will not change. However, there is no guarantee of invariance when the path of integration lies across several level sets. This reflects the mathematical property that line integrals are, in general, path dependent.

Path dependence means that the index (3) will generally have a different value for a path $\beta(t) \in \Gamma_1$ than path $\alpha(t) \in \Gamma$, even though the beginning and end points of Γ_1 and Γ are identical. This can lead to the following situation: the economy moves along Γ_1 from X from X' (which is on a different isoquant); the economy then returns along Γ to the original point X; because of path dependence, the vector of quantities represented by the vector X will have a different Divisia index value after the trip around the composite path, and subsequent circuits will produce still different values. The value of the Divisia index at any point X is thus arbitrary under path dependence. The uniqueness of the Divisia index thus involves path independence.

The condition for path independence is the existence of a homothetic potential function, log F, such that $\varphi = \nabla logF$, where φ is defined in (3). Given the existence of the potential function, the value of (3) is $F(X(T))/F(X(0))$, implying path independence since (3) depends only on the end points of the path, $X(0)$ and $X(T)$. Conversely, if (3) is path independent, there exists a potential function log F such that $\nabla \log F = \varphi$. In some

D

applications in productivity analysis, the homothecticity condition must be strengthened to linear homogeneity, but this can be weakened depending on data availability (Hulten 2001, pp. 11–12).

We note, finally, that the Divisia index is defined using time as a continuous variable. Data on prices and quantities typically refer to discrete points in time, and the indexes constructed from them must therefore have a discrete–time form. The continuous-time Divisia index is nevertheless useful, both for informing the structure of these discrete-time indexes (for example, for the determining which variables are conceptually related), and for interpreting the results. The Divisia framework is also appropriate for the theoretical analysis of many economic problems, such as the use of Divisia indexes by Solow (1957) in growth accounting.

One approach to linking discrete and continuous index numbers is to approximate the continuous variables of (2) with their discrete time counterparts. Under the Törnqvist (1936) approach, the growth rates of prices and quantities are approximated by logarithmic differences, and the continuous weights by two period arithmetic averages. The Tornqvist approximation to the growth rate of the Divisia quantity index can then be written:

$$\sum_{i=1}^{i=T} \frac{1}{2}\left[\frac{P_{i,t}X_{i,t}}{Y_i} + \frac{P_{i,t-1}X_{i,t-1}}{Y_{t-1}}\right]\left[\log X_{i,t} - \log X_{i,t-1}\right] \quad (4)$$

A similar approximation applies to the growth rate of the Divisia index of prices.

While the Törnqvist index may be regarded as approximate, Diewert (1976) has shown that it is exact when the underlying potential function has the (continuous) translog form. This result is very important in its own right, but can also be regarded as an important conceptual link between the discrete and continuous–time families of index numbers, given the exact properties of the Divisia index in continuous time.

The continuous Divisia index can also be approximated by using chain indexing procedures (the Divisia index is sometimes regarded as a chain whose links are defined over infinitesimal time periods). Other numerical approximation techniques can also be employed.

See Also

- ▶ Divisia, François Jean Marie (1889–1964)
- ▶ Index Numbers

Bibliography

Balk, B. 2005. Divisia price and quantity indices: 80 years after. *Statistica Neerlandica* 59: 119–158.

Diewert, W. 1976. Exact and superlative index numbers. *Journal of Econometrics* 4 (2): 115–145.

Divisia, F. 1925–6. L'indice monétaire et la théorie de la monnaie. *Revue d'Economie Politique* 39(4): 842–864; (5): 980–1008; (6): 1121–1151; 40(1): 49–81. Also separately: Paris: Société Anonyme du Recueil Sirey, 1926.

Fisher, I. 1921. *The making of index numbers*. Boston: Houghton Mifflin Co.

Hulten, C. 1973. Divisia index numbers. *Econometrica* 41: 1017–1025.

Hulten, C. 2001. Total factor productivity: A short biography. In *New developments in productivity analysis*, Studies in income and wealth, ed. C. Hulten, E. Dean, and M. Harper, vol. 63. Chicago: University of Chicago Press for the NBER.

Richter, M. 1966. Invariance axioms and economic indexes. *Econometrica* 34: 739–755.

Samuelson, P., and S. Swamy. 1974. Invariant economic index numbers and canonical duality: Survey and synthesis. *American Economic Review* 64: 566–593.

Solow, R. 1957. Technical change and the aggregate production function. *Review of Economics and Statistics* 39: 312–320.

Törnqvist, L. 1936. The bank of Finland's consumption price index. *Bank of Finland Monthly Bulletin* 10: 1–8.

Divisia, François Jean Marie (1889–1964)

David E. R. Gay

Keywords

American Statistical Association; Divisia Index; Econometric Society; International Econometric Society.; Mathematical economics; Statistics and economics

JEL Classifications

B31

Divisia was born in Tizi-Ouzou, Algeria. He received baccalaureate degrees in mathematics and philosophy at Algiers. After two years in the Ecole Polytechnique he worked for the government as a civil engineer (Ponts et Chaussées). His graduate engineering work at the Ecole Nationale des Ponts et Chaussées was completed in 1919 after the interruption of the First World War. After nearly ten years as a government engineer he joined the ministry of national education to continue research and teaching economics. He became a professor of applied economics at the Ecole Nationale des Ponts et Chaussées (1932–50), the Conservatoire National des Arts et Métiers (1929–59), and the Ecole Polytechnic (1929–59). He was a founding member of the Econometric Society and its president in 1935. Subsequently he was also president of the Paris Statistics Society (1939) and of the International Econometric Society. He was a Fellow of the American Statistical Association and of the American Association for the Advancement of Science.

His major contributions to economics can be found centred in several books on economics and applied statistics. The Divisia Index, a variable-weight price index, was developed in *L'indice monétaire et la théorie de la monnaie* (1926). His *Economique rationnelle* (1928) was widely acclaimed in mathematical economics and was awarded prizes by the Academy of Sciences and by the Academy of Moral Sciences and Politics. Using a microeconomic perspective he cautioned against uncritical acceptance of macroeconomic research in *Traitement économétrique de la monnaie, l'intérêt, l'emploi* (1962).

Selected Works

1926. *L'indice monétaire et la théorie de la monnaie*. Paris.
1928. *Economique rationnelle*. Paris.
1931. *L'epargne et la richesse collective*. Paris.
1951–65. *Exposés d'économique*. Paris.
1962. *Traitement économétrique de la monnaie, l'intérêt, l'emploi*. Paris.

Division of Labour

Peter Groenewegen

Abstract

Division of labour has been a very important topic for economic writings from the earliest times, and was treated in great detail by major economists, including especially Adam Smith and Alfred Marshall. This article surveys the development of 'division of labour' from its beginnings in the writings of Greek philosophers through the centuries and up to the 21st century. It therefore also reflects on its offshoots: international division of labour, sexual division of labour and its contemporary revival as an essential adjunct to the theory of economic growth, labour productivity, inter-firm cooperation, and its modern limits in coordination and communication costs.

Keywords

Agricultural productivity; Aristotle; Austrian capital theory; Babbage, C.; Beccaria, C. B.; Bagehot, W.; British Association; Capital accumulation; Capitalism; Carl, E. L.; Carlyle, T.; Clustering; Communication and coordination costs; Comte, A.; Concentration; d'Alembert, C.; Dexterity; Diderot, D.; Division of labour; Durkheim, E.; Economies of scale; Engels, F.; Engineering; Equilibrium analysis; Ferguson, A.; Firm, theory of; Foreign direct investment; Hegel, G. W. F.; Hodgskin, T.; Housework; Human capital; Hutcheson, F.; IbnKhaldun; Increasing returns; Industrial organization; Industrial psychology; Industrial revolution; Inequality; Inter-firm cooperation; International division of labour; Kaldor, N.; Labour productivity; Mandeville, B.; Manufacturing; Manufacturing division of labour; Marshall, A.; Marshall, M. P.; McCulloch, J. R.; New classical microeconomics; New growth economics; Nicholson, J. S.; Petty, W.; Plato; Productivity growth; Putting-out system; Quesnay, F.; Rae,

J. (1796–1872); Robbins, L. C.; Ruskin, J.; Segmented labour markets; Senior, N.; Sexual division of labour; Sidgwick, H.; Social division of labour; Specialization; Spencer, H.; Taussig, F. W.; Technical change; Torrens, R.; Tucker, J.; Turgot, A. R. J.; Ure, A.; Verdoorn's Law; Women's work; Xenophon; Young, A. A.

JEL Classifications
B1

Division of labour, or specialization, may be defined as the division of a process or employment into parts, each of which is carried out by a separate person, or any system of production in which tasks are separated to enable specialization to occur. This includes the separation of employments and professions within society at large or *social division of labour* as well as the division of labour which takes place within the walls of a factory building or within the limits of a of a single industry, the *manufacturing division of labour.* Division of labour as a form of specialization can also be practiced by small firms which all contribute to the production of parts (inputs) for the manufacturing of a complex output, as in the case of aircraft production or sophisticated electronic equipment. This form of business organization requires excellent coordination and communication between its various parts to ensure continuous supply of the necessary parts for the manufacturer of the final output. It is a geographical form of division of labour, developed from the notion of clustering related firms in a particular area or industrial district (for a survey, see Dosi 1988).

Division of labour and its consequences for productivity were analysed as early as the time of the Greek philosophers, including Plato, Aristotle and Xenophon. Early analysis of the manufacturing division of labour had to await industrial developments of the 17th and 18th centuries and underwent further qualitative change in the 19th, 20th and 21st centuries. Hence manufacturing and more detailed division of labour should not be seen as a simple continuum of the social division of labour. By the end of the Middle Ages, social division of labour was extensively practiced; manufacturing division of labour, generally speaking, came with the Industrial Revolution. Under modern capitalism, social division of labour remains largely a market influenced phenomenon but manufacturing division of labour is enforced by those who plan and control the manufacturing process. Furthermore, the one divides society: the other human activity within the workshop, or within an industry: labour generally enhances 'the individual and the species, [a manufacturing division] of labour, when carried on without regard to human capabilities and needs, is a crime against the person and against humanity' (Braverman 1974, p. 73). Division of labour was first practiced within the household, a *sexual division of labour* between women's activities in or near the house, and those of men further afield. When applied to local specialization of industries both nationally and internationally, it has produced a variety of conceptions of the *territorial* or *international (global) division of labour.*

Adam Smith (1776) placed the division of labour at the forefront of his discussion of economic growth and progress. Neither in its social nor in its manufacturing forms did the idea originate with him. It retained a varying, but often very prominent, place in 19th-century writings (particularly those of Senior, Babbage, John Stuart Mill, Marx and Marshall).'About 1890, Schmoller, Semmel, Bücher, Durkheim and Maunier all wrote on religious and sociological aspects of specialization' (Salz 1934, p. 284). For much of the 20th century, division of labour and specialization virtually disappeared as a major topic from economic texts. Reasons for this varied. Some economists believed such discussions were more appropriate to technical handbooks of production engineering and factory management. Other writers wished to confine analysis of its effects to sociological studies assessing the general impact of division of labour on society. The return of economic growth as an important part of the economist's research programme from the 1950s onwards, and earlier the work of Young (1928), brought renewed interest in the division of labour in its wake, as did growing

dissatisfaction with the narrow view confining economics to studying 'the disposal of scarce commodities'(Robbins 1932, p. 38). Global organization of manufacturing made possible by improvements in transport and communication implies modern adaptations of the division of labour which economists cannot ignore. An example is the formation of industrialized districts, first observed and analysed by Alfred Marshall (1890), to be rediscovered and adapted to the post-Second World War Italian situation by Becattini (for example 1990, 2001) and his colleagues (for a survey, see Goodman and Pamford 1989). The various dimensions of division of labour raised in these introductory paragraphs suggest that a broad-based treatment of the subject is warranted by featuring highlights within its continuous development.

The Greeks

Many of the major Greek philosophers discussed aspects of the division of labour in their writings. In Book 2 of the *Republic,* Plato stated the necessity for a division of labour or specialization in occupations for social well-being and the adequate satisfaction of primary wants linking the phenomenon with exchange, the requirements of 'a market, and a currency as a medium of exchange' (Plato 380 BC, pp. 102–6). Aristotle, though very conscious of the social need for a division of labour, did not depart much from Plato's earlier discussion (see Bonar 1893, p. 34). More importantly, Xenophon linked division of labour and specialization to great cities, because they provided a substantial demand for individual products while the subdivision of work raised the skill of individual workers. Extracts from the work of these Greek pioneers on the division of labour have been often reprinted (see, for example, Sun 2005, chs. 2–4). Knowledge of these Greek texts among Arabian Islamic scholars during the middle ages enabled them to produce sophisticated treatments of the division of labour. Examples are the writings of Islamic theologian, al-Ghazali (1058–1111) and, more importantly, the writings of fourteenth century Islamic philosopher and historian, IbnKhaldun, whose *Muqaddima* contains a detailed account of the division of labour (Sun 2005, pp. 7–8, ch. 5).

Subsequent Pre-Smithian Developments

Towards the end of the seventeenth century, English economic literature rediscovered the concept of the division of labour and began to analyse the more modern manufacturing forms, linking them to productivity growth, cost reduction, increased international competitiveness and associating its scope with the more extensive markets made possible through urbanization. For example, Petty's *Political Arithmetick* written in 1671 compared the benefits of division of labour in textile production with specialization in ship building:

> For as Cloth must be cheaper made, when one Cards, another Spins, another Weaves, another Draws, another Presses and Packs; than when all the Operations above-mentioned, were clumsily performed by the same hand; so those who command the Trade of Shipping [need] to build. . .a particular sort of Vessels for each particular Trade. (Petty 1671, pp. 260–1)

Ten years later, in *Another Essay on Political Arithmetick Concerning the Growth of the City of London* (1683, p. 473), Petty showed that a major gain from a vast city like London came from the improvement and growth of manufactures it encouraged:

> For in so vast a City *Manufacturers* will beget one another, and each *Manufacture* will be divided into as many parts as possible, whereby the Work of each *Artisan* will be simple and easy; As for example in the making of a *Watch,* if one Man shall make the *Wheels,* another the *Spring,* another shall Engrave the *Dial-plate,* and another shall make the *Cases,* then the *Watch* will be better and cheaper, than if the whole Work be put upon any one man.

In continuing this argument Petty also suggested that specialization benefits could be achieved from concentrating certain manufactures on a particular location, partly because of the savings in transport and communication costs such concentration entailed (Petty 1683, pp. 471–2). The anonymous author of *Considerations on the East India Trade* (1701, pp. 590–2) illustrated

productivity gains from the division of labour by examples drawn from cloth making, watch making and shipbuilding. He clearly indicated that sufficient demand and regular trade were a precondition for such improvements, which lowered manufacturing labour costs without the need to lower wages. During the 18th century, examples of authors aware of the benefits and preconditions for a division of labour become more common. Practical writers like Patrick Lindsay (1733), Richard Campbell (1747) and Joseph Harris (1757) tended to concentrate on manufacturing division of labour using examples from linen and pin production as well as from the familiar watch making. Those writing from the position of moral or political philosophy, like Mandeville (1729), Hutcheson (1755), Ferguson (1767) and Josiah Tucker (1755, 1774) concentrated more on aspects of the social division of labour.

Discussion of the division of labour was of course not confined to English economic literature. A treatise on wealth published in the 1720s by Ernst Ludwig Carl discussed the benefits of the division of labour, applying them also to demonstrate the gains from free trade through an international division of labour based on different climates, resource availability and locational advantages (cited in Hutchison 1988, pp. 161–2). Among the Physiocrats, Quesnay dealt briefly with the social aspects of the division of labour in his article 'Natural Right' (1765, p. 51). Turgot developed the subject more thoroughly, making it the starting point of his *Reflections,* subsequently associating it with the introduction of money, the extension of commerce and the accumulation of capital (1766, pp. 44–6, 64, 70). Earlier, Turgot (1751, pp. 242–3) had linked the spread of social division of labour to inequality, arguing that this particular consequence of inequality improved living standards for even the humblest members of society and made possible cultivation of the arts and sciences. Among the general principles with which Beccaria (1771, pp. 387–8) commenced the argument of his *Elementi*, the division of labour and its benefits in terms of increased skills and dexterity are clearly set out. Finally, it may be noted that the *Encyclopédie* of Diderot and d'Alembert in its article 'Art' discussed the essentials of the manufacturing division of labour, listing its consequences as improvements in skill, better quality products, saving of time and of materials, and 'of making the time or the labour go further, whether by the invention of a new machine or the discovery of a more suitable method'. In its article on pins ('Epingle') their manufacture is described as being generally subdivided into eighteen separate operations and thereby a prime example of the manufacturing division of labour (see Cannan 1929, pp. 94–5).

Adam Smith's Treatment of the Division of Labour

Adam Smith's discussion of the division of labour deserves separate treatment not because of its 'originality' or 'completeness of exposition' (Cannan 1929, p. 96) but because 'nobody either before or after [him], ever thought of putting such a burden upon division of labour. With A. Smith, it is practically the only factor in economic progress' (Schumpeter 1954, p. 187). The first three chapters of the *Wealth of Nations* were devoted to its analysis because it provided one of the two causes explaining increases in per capita output by which Smith defined the wealth of the nation. Although therefore only one of two causes, the other being 'the proportion between the number of those who are employed in useful labour, and that of those who are not so employed' (Smith 1776, p. 10), it is the dominant one. Smith seems to have believed that scope for substantial increases in the proportion of the labour force to productive activities was limited. Using the equation, $g = (k \,.\, p/w) - 1$, developed by Hicks (1965, p. 38) to summarize the Smithian growth progress, if a change in k, the proportion of productive labour in the labour force, is more or less ruled out, a substantial growth rate (g), given the real wage (w), depends exclusively on rising productivity (p) through extensions of the division of labour. Smith's emphasis on the division of labour as a factor in growth via its enormous influence on productivity makes his treatment of the subject so novel. Surprisingly, this aspect of his contribution

was taken up by few 19th-century writers and had to be largely rediscovered in the work of Young (1928) and Kaldor (1972) who reiterated dynamic aspects of the phenomenon Smith was analysing.

Even though it was the most frequently revised part of his economics (see Meek and Skinner 1973), Smith's basic account of the division of labour contains a number of weaknesses. First, Smith failed to develop aspects of the manufacturing division of labour with which he ought to have been familiar. Marglin (1974) points out that Smith ignored organizational features from a division of labour taking place within the one building of relevance to some well-established industries like textiles and the manufacture of metal implements. These organizational features which Smith omitted were associated with growing labour discipline problems, wasting time and materials, inherent in the putting-out system, then the dominant form of manufacturing organization. In fact it can be suggested that if this aspect of the division of labour is more fully taken into account, its important role in explaining economic growth so much emphasized by Smith is more easily integrated as a major factor explaining the industrial revolution (see Groenewegen 1977). Marglin (1974) also questioned the force of 'the three different circumstances' by which Smith (1776, p. 17) explained the productivity gains from the division of labour: increased dexterity, saving of time, and invention of machinery. Although increased dexterity is clearly a product of a division of labour in a manufacturing process, its scope there is rather limited when compared to that of the continual practice of surgeons, concert pianists and opera singers, to give some examples. Time saved in eliminating time lost in passing from job to job is trivial and not the 'very considerable' benefit Smith (1776, pp. 18–19) had suggested. Savings in materials and time through transforming a putting-out to a factory system, an organizational feature of the division of labour Smith had ignored, was more important, particularly through eliminating losses from pilfering. Rae (1834, pp. 164–5) saw savings in the use of tools as far more significant than time saved, and for him (pp. 352–7) this provided the basic reason for extending the division of labour. Other 19th century writers, particularly Babbage (1832), expanded further on this aspect of the matter. Smith's association of division of labour with inventions (1776, pp. 19–22) covered both 'on the job improvements' and scientific inventions by specialists originating from within a more sophisticated division of professions. It ignores, as Hegel (1821, p. 129) was one of the first to point out (cf. Stewart 1858–75, vol. 8, pp. 318–19), that as division of labour makes 'work more and more mechanical,. . . man is able to step aside and install machines in his place'. This feature of the process was subsequently noted by Babbage (1832, pp. 173–4), Ure (1835, p. 21) and developed by Marx (1867). In short, the three circumstances Smith saw as explaining the productivity consequences from the division of labour derive their basic validity from reasons different to those Smith advanced. Further, Smith's remarks (1776, pp. 16–17) on the smaller benefits from applying the division of labour to agriculture than to manufacturing can be contrasted with his quite different and controversial analysis of the primacy of agricultural investment in terms of its employment of productive labour. Agriculture's more substantial contribution to gross revenue as Smith (1776, Book II, ch. 5) subsequently argued, was used by him to define the 'natural' course of economic development (Book II, ch. 1) and recommended as superior practice for newly settled regions like the American colonies. Perelman (1984, p. 185) explained this seeming contradiction in Smith by suggesting Smith was the 'first theorist of neo--imperialism' because his strategy of development forces developing regions to specialize in raw material production whose terms of trade with manufactures are invariably poor. More likely, Smith's views on the productivity of agriculture relative to manufacturing are posed in terms of different yardsticks: agricultural activity by the very nature of its processes is less amenable to division of labour, even though its ability to employ productive labour is greater than produced by equal investments in manufacture and trade. However, growing mechanization of agriculture, especially in the 20th century, together with the greater scope for exporting agricultural

surplus with modern transport, encouraged specialization in agriculture and very large scale farming (Salz 1934, p. 283).

A final controversial issue from Smith's treatment of the division of labour concerns its social consequences, an argument he placed in the context of public education. The 'few simple operations' which under a division of labour most ordinary labouring people are asked to perform, renders them 'as stupid and ignorant as it is possible for a human creature to become' and increased 'dexterity at his own particular trade' is purchased with a reduction in 'intellectual, social and martial virtues.. .unless government take some pains to prevent it' through providing general education (Smith 1776, pp. 781–5). Smith was not alone in presenting this disadvantage in an extensive division of labour: similar views were put by Ferguson (1767, p. 280) and Kames (1774). Ferguson described 'ignorance as the mother of industry' and argued that prosperous manufactures arise 'where the mind is least consulted, and where the workshop may.. .be considered as an engine, the parts of which are men. 'At the turn of the century, and after, German philosophers (for example, Schiller 1793; Hegel 1821 and the young Marx 1844) developed this into a humanist critique of industrial society, suggesting like Smith that these detrimental consequences were removable by education, especially aesthetic education. Such sentiments were resurrected in mid-19th century England by Carlyle (1843) and Ruskin (1851–3, pp. 197–8). For others, Smith's remarks were an aberration, 'as unfounded [a statement] as can well be imagined' (McCulloch 1850, p. 350) or even a contradiction with the division of labour's ability to inspire inventive faculties in labourers (West 1964).

Despite its deficiencies, Smith's account of the division of labour proved particularly hardy and was invariably praised in most general terms by major textbook writers of the 19th century and after, though few followed the emphasis Smith gave it as the key factor explaining growth. Cannan (1929, p. 97) ascribed this success to 'the popularity of its form'. It can also be attributed to the striking productivity increase inherent in the pin example (cf. Mill 1821, p. 215) and the unambiguous connection Smith drew between increased division of labour, extending the market and human proclivities 'to truck and barter' (McCulloch 1825, pp. 54–5). The account of the division of labour is undoubtedly one of Smith's best remembered performances in economics.

19th-Century Developments

With the growth of the factory system and more extensive use of increasingly sophisticated machinery, the manufacturing form of division of labour was considerably expanded. Consequently, some economic writers focused on a number of new aspects of the phenomenon, linking the division of labour with developments in the machine tool industry, large scale production and its advantages, and hence, on a more theoretical level, with increasing returns to scale and explicit recognition of a different pattern of productivity growth in manufacturing from that in agriculture.

Charles Babbage was in many respects the pioneer in presenting the division of labour as 'the most important principle on which the economy of a manufacture depends' (1832, p. 169). He therefore carefully revised the advantages of a division of labour as first expounded by Adam Smith. In this discussion, time (and cost) savings were also related to time saved in learning a skill and reduced waste of materials during the learning process (pp. 170–1), as well as economy in tool using (p. 172), while the association between division of labour, dexterity and the introduction of new machines was developed more precisely and rigorously. More significantly, Babbage pointed to a hitherto ignored additional advantage of the division of labour he had derived from observation. This had earlier been discussed by Gioja (1815–17) whose interesting contribution on this subject was analysed by Scazzieri (1981, ch. 3).

> *By dividing the work to be executed into different processes of skill or of force, . . .the master manufacturer. . . can purchase exactly that precise quantity of both which is necessary for each process; whereas, if the whole work were executed by one workman, that person must possess sufficient skill to perform the most difficult, and sufficient*

> *strength to execute the most laborious, of the operations into which the art is divided.* (Babbage 1832, pp. 175–6; emphasis in original)

This economy of skill, Babbage demonstrated from a pin example, not only reinforced the cost advantages traditionally associated with division of labour, but was also a major cause of establishing large factories: 'When the number of processes into which it is most advantageous to divide it, and the number of individuals to be employed in it, are ascertained then all factories which do not employ a direct multiple of this number, will produce the article at a greater cost' (Babbage 1832, p. 213). Detailed division of labour, Babbage also argued, as in its manufacturing form, can also be applied to mental labour (p. 191). An illustration of its application to mining highlights these control and information gathering features, two aspects of the division of labour to which Babbage paid particular attention. His analysis of the division of labour is even more important because the process as he described it is made interdependent with machine production, increased factory size, lower costs and prices from such concentration of industry and hence induces growth in demand and an extended market (see Corsi 1984).

Ure's (1835) contribution must also be noted. It likewise linked development of the factory system to division of labour, summarizing 'the principle of the factory system. . . as substituting mechanical science for hand skills, and the partition of a process into its essential constituents' (1835, p. 20). Ure commented on two other consequences of the division of labour in modern factories: deskilling of the workforce when workers become 'mere overlookers of machines' and the development of mechanical engineering since the 'machine factory displayed the division of labour in manifold gradations' and facilitated the substitution of skilled hands by 'the planning, the key-groove cutting, and the drilling machines' (pp. 20–1).

Accounts of the division of labour by economists of the middle of the century were generally less innovative than those of Babbage and Ure, though they did occasionally provide some new points of departure. Senior (1836, pp. 74–5, 77), after classifying division of labour as one major advantage from the use of capital, concentrated on listing its benefits additional to those given by Smith. Illustrating from the post office, he argued that the fact that 'the same exertions which are necessary to produce a single given result are often sufficient to produce many hundreds or many thousands similar results' was one aspect of the division of labour omitted by Smith. The development of retailing as a separate profession was likewise something Smith had failed to consider adequately. More importantly, for a number of reasons, but particularly the division of labour, Senior suggested 'additional Labour when employed in Manufactures is MORE, when employed in Agriculture is LESS efficient in proportion', linking manufacturing activity implicitly to increasing returns to scale (1836, pp. 81–2). Mill (1848) treated division of labour as an important aspect of cooperation, arguing that irrespective of its well-known productivity advantages, without this complex cooperation in the modern division of labour 'few things would be produced at all' (Mill 1848, p. 118). In discussing the productivity advantages, Mill cited the modification and additional advantages provided by Babbage (1832) and Rae (1834), adding little to their discussion. However, in Chapter 9 dealing with large scale and small scale production, he highlighted the point, so 'ably illustrated by Mr Babbage. . . [that] the larger the enterprise, the farther the division of labour may be carried.. .as one of the principal causes of large manufactories' (Mill 1848, p. 131), thereby bringing the argument firmly into the corpus of economics. Mill's account was largely followed by Fawcett (1863) and in most of its essentials by Nicholson (1893).

Marx's account (1867, chs. 13–15) combines much of this discussion, endowing it in the process with sharper analytical insights derived from his study of both the technical literature and his appreciation of the significance of the qualitative changes underlying the evolution of the division of labour. To Marx is owed the important distinction between manufacturing and social division of labour, as well as the precise assessment of the organizational features of its application to modern manufacture, derived from his careful study of

Babbage, Ure and many other sources. No wonder that Nicholson (1893, p. 105) described Marx's treatment as 'both learned and exhaustive and... well worth reading'. More recently, Rosenberg (1976) expressed regret that Marx's close study of 'both the history of technology, and its newly emerging forms' has had so few imitators among contemporary economists.

Marshall is another economist from the second half of the 19th century who fully appreciated the importance of the division of labour and revealed it in its more modern forms. In 1879, the *Economics of Industry,* written with his wife (Marshall and Marshall 1879), devoted Chapter 8 of Book I to the division of labour, immediately after its Chapter 7 on organization of industry. It distinguished the opportunity to apply a division of labour as inherent in the nature of the work, as dependent on direction and control by an entrepreneur as earlier indicated by Bagehot, and as applied to firms: 'If there are any producers, large and small, all engaged in the same process, *Subsidiary Industries* will grow up to meet their special wants.' These include special machine tool makers for the industry, improved transport to enhance communication between related firms, as well as auxiliary enterprises in banking and credit provision (Marshall and Marshall 1879, p. 52). Localization of industry also fosters 'education of skills and taste' and 'diffusion of linked knowledge', and encourages large firms. Hence division of labour is closely related to economies of scale, where size has enabled specialization to grow more and more. Marshall also devoted no less than three chapters to division of labour in his *Principles* (1890, Book IV, chs. 9–11), not only covering points traditionally dealt with under this heading, but often introducing subtle modifications. For example, Marshall (1890, p. 263) discounted detrimental social consequences from monotonous work by pointing to the mental stimulus from the 'social surroundings of the factory' and the view that factory work was not inconsistent with 'considerable intelligence and mental resources'. Likewise, he extended Babbage's principle of 'economy of skill' to economy of machinery and materials (1890, p. 265), used it as a major explanatory factor for the localization of specialized industry (p. 271) and made it the chief advantage of large scale production in his famous discussion of economies of scale (p. 278). Later, Marshall applied these aspects of his work to his detailed study of industry and trade to explain such things as America's leadership in standardized production (seen by Marshall 1919, p. 149, as an 'unprecedented' application of Babbage's 'great principle of economical production'), the successful specialization of plant during the First World War, and new issues concerning the growth of the firm. It is therefore paradoxical that Marshall's work in other respects induced the demise of the division of labour in theoretical literature. This arose from the incompatibility of increasing returns to scale with stable demand and supply equilibrium (Marshall 1890, Appendix H). Apart from this, modern equilibrium analysis found it difficult to come to grips with the dynamic features of the division of labour process, and it is presumably at least partly for this reason that division of labour was dropped as an important subject from the economic textbooks (see Kaldor 1972). However, the locational aspects of the division of labour were further addressed by Becattini (for example 1990, 2001) in his development of the notion of industrial districts as a concentration of related firms. Marshall had discovered this aspect of industrial organization through the factory tours in the British midlands and Scotland he engaged in from the late 1860s, on which he first reported in 1879. When division of labour for technical reasons could not take place within the same building, small firms spring up specializing in part of the manufacturing process, thereby generating a division of labour among firms concentrated in a particular geographical area (for a survey, see Goodman and Pamford 1988).

International Division of Labour

Torrens (1808) appears to have been the first economist to distinguish the territorial division of labour from the mechanical division, suggesting that the former is inspired by 'different soils and climates [being] adapted to the growth of

different production' thereby inducing regional specialization in those products which best suit 'the varieties of their soil' and climate. Taking advantage of territorial division of labour through regional and international trade enhances productivity and increases the wealth of nations as much as a manufacturing division of labour. Senior (1836, p. 76) also drew attention to this aspect of the division of labour, attributing its discovery to Torrens. Marshall (1890, pp. 267–77) covered territorial division of labour under localization of industry while Taussig (1911, pp. 41–7) called it 'the geographical division of labour' with gains arising from 'the adaptation of different regions to specific articles' for climatic and resource endowment reasons as well as from the general increase in proficiency which all specialization brings. During the 1970s a new dimension of the international division of labour was analysed, concentrating on its direct foreign investment aspects. Its novel features were a tendency to 'undermine the traditional bisection of the world into a few industrialized countries on the one hand, and a great majority of developing countries integrated into the world economy solely as raw material producers on the other, and [secondly, to compel] the increasing subdivision of manufacturing processes into a number of partial operations at different industrial sites throughout the world' to take advantage of favourable labour market circumstances, relatively cheap transport opportunities, tax breaks and other government inducements for foreign investors (Fröbel et al. 1980, p. 45). This multinational dimension to application of the division of labour is a direct descendent from the concept as understood by Smith, Babbage, Ure and Marx.

The characteristics of the contemporary global division of labour have been well captured by Hobsbawm (2000, pp. 65–6):

> Thus, while the global division of labour was once confined to the exchange of products between particular regions, today it is possible to produce across the frontiers of states and continents. This is what the process is founded on. The abolition of trade barriers and liberalization of markets is, in my opinion, a secondary phenomenon. This is the real difference between the global economy before 1914 and today. Before the Great War, there was pan global movement of capital, goods and labor. But the emancipation of manufacturing and occasionally agricultural products from the territory in which they were produced was not yet possible. When people talked about Italian, British and American industry, they meant not only industries owned by citizens of these countries, but also something that took place almost entirely in Italy, Britain, or America, and was then traded with other countries. This is no longer the case. How can you say that a Ford is an American car, given that it is made of Japanese and European components, as well as parts manufactured in Detroit?

Sexual Division of Labour

The first explicit reference to a sexual division of labour in economic literature I could find is Hodgskin (1827, pp. 111–12). He argued that

> There is no state of society, probably, in which division of labour between the sexes does not take place. It is and *must* be practiced the instant a family exists. Among even the most barbarous tribes, *war* is the exclusive business of the males; they are in general the principal hunters and fishers … the woman labours in and about the hut… In modem as well as in ancient times,… we find the men as the rule taking the out-door work to themselves, leaving the women most of the domestic occupations The aptitude of the sexes for different employments, is only an example of the more general principle, that every human being … is better adapted than another to some particular occupation.

Marx and Engels (1845–6, pp. 42–3) ascribed beginnings of the division of labour 'originally [to] nothing but the division of labour in the sexual act' and only later to that 'spontaneously' or 'naturally' derived from predisposition, needs, accidents, and so on. Engels (1884, esp. p. 311) elaborated further on the matter presenting the sexual division of labour in the family as a barrier to the 'emancipation of women'. Such an emancipation, he argued, was 'possibly only as a result of modern large- scale industry [which] actually called for the participation of women in production and moreover, strives to convert private domestic work also into a public industry'. Both aspects of the sexual division of labour to which Engels referred in the context of women's emancipation have been taken up in more recent research. The role of domestic labour has been

analysed by contemporary writers (see, for example, Himmelweit and Mohun 1977; Gershuny 1983) while attention has also been drawn to the shift in the provision of services from domestic production to production for the market (laundromats, take-away-food) as a result of the gradual break-down of the traditional sexual division of labour within the family (Gouverneur 1978). Sexual division of labour issues have also been applied in segmented labour market analysis, thereby enriching this particular aspect of labour economics.

Becker (1985) has analysed the sexual division of labour in the context of human capital investment and allocating the work load of parties within the household. Thus both the allocation of effort within a household, and the advantages of investing in specific human capital are designed to enhance the social division of labour and its benefits without necessarily diminishing the exploitative aspects of such arrangements (Becker 1985, p. S41). Social factors are, however, equally important. Increasing returns by itself cannot explain the traditional division of labour within the household; a division of labour itself subject to change. The increased contribution to housework by men during the 1970s is one observed aspect of this social change (Becker 1985, p. S56). Furthermore, as Posner (1992, pp. 54, 129) has noted in particular, women were not fully brought into the work place on a large scale until the two world wars, and this only became a dominant pattern in employment from the 1950s onwards. Cigno (1991) discusses many of these issues as part of his economics of the family.

Decline and Rehabilitation of Division of Labour in the 20th and 21st Centuries

The association between division of labour and increasing returns, the consequent possibility of falling supply and cost curves, created problems for equilibrium analysis already noticed as a factor explaining decline in emphasis on the division of labour and induced its virtual elimination from much of the theoretical literature. Attempts to remove division of labour from economics were also based on other grounds. Robbins (1932, pp. 32–8) argued that study of the 'technical arts of production' belonged to engineering and not to economics or, in the case of 'motion study', to industrial psychology even if this meant removal of traditional topics like division of labour from economics. Robbins's approach followed Sidgwick's (1883, pp. 104–7) treatment, removing all technical aspects from the topic, leaving only what he called the pure economics side. Others suggested it was better to leave discussion of division of labour to sociologists because Durkheim, and before him Comte and Herbert Spencer, had absorbed it within this emerging discipline. However, some economists in the 20th century objected to removal of the division of labour from economics. In particular, this would reduce understanding of the dynamics of economic progress.

Allyn Young (1928) was one of these economists. He made Adam Smith's theorem that the division of labour is limited by the extent of the market the central theme of his address to section F of the British Association, arguing this was 'one of the most illuminating and fruitful generalizations which can be found in the whole literature of economics' (Young 1928, p. 529). Rather than covering all aspects of the division of labour, Young concentrated on two interdependent matters: 'growth of indirect and roundabout methods of production and the division of labour [or increased specialization] among industries' (Young 1928, p. 529) but the former, as Kaldor (1975, pp. 355–6) pointed out, was not to be confounded with the Austrian capital theoretic notion. From this he deduced division of labour as a cumulative, self-reinforcing process, because every re-organization of production, sometimes described as a new invention, involves fresh application of scientific progress to industry,

> alters the conditions of industrial activity and initiates responses elsewhere in the industrial structure which in turn have further unsettling effects . . . The apparatus of supply and demand in their relation to prices does not seem to be particularly helpful for the purpose of an inquiry into these broader aspects of increasing returns. (Young 1928, p. 533)

However, apart from this damaging conclusion for competitive price theory, the 'possibility of

economic progress' could not really be grasped by ignoring these factors of greater specialization, better combinations of advantages of location, and a consequent increased number of specialized producers between basic raw materials and final producers (Young 1928, pp. 538–40).

Kaldor was a major economist who took up Young's challenge in both its critical (Kaldor 1972, 1975) and more constructive aspects (Kaldor 1966, 1967).The major thrust of Kaldor's positive argument proclaimed that faster growth is derived from faster growth in the manufacturing sector, partly from the cumulative features linking the growth of manufacturing to growth of labour productivity via static and dynamic economies of scale, or the notion of increasing returns as developed by Young from the division of labour. This strong and powerful interaction of productivity growth and manufacturing growth is also posited in Verdoorn's Law (1949) but its association with aspects of the division of labour is what is relevant here. Faster manufacturing growth draws labour from other sectors of the economy, inducing faster productivity growth, but as the scope of transferring such labour from lower productivity sectors like agriculture dries up, the growth process slows down (see Thirlwall 1983). A key feature of the process, as Rowthorn (1975, p. 899, n. 1) noticed in one his skirmishes with Kaldor on the subject, is that it is an interdependent, cumulative historical process where 'higher productivity means more exports which means greater industrial output which via its effects on investment, innovation and *scale of production* reacts back on productivity growth'. The importance of such a process was given detailed empirical examination in a discussion of the Taiwan machine tools industry in the 1970s as an application of the division of labour, envisaged as increases in output increasing productivity, with 'technological change, broadly defined, sandwiched in between' (Amsden 1985, p. 271). Writers in the new growth economics, who emphasized the impact of increasing returns from specialization on growth performance (Romer 1987) drew in part for their inspiration on the literature of the division of labour, in Romer's case as represented by Marshall (1890) and possibly Young (1928).

Research from the 1990s has particularly stressed the importance of communication and co-ordination costs of the division of labour. Becker and Murphy (1992) portray these costs as setting limits on the division of labour more important than that exerted by the extent of the market so heavily emphasized by Adam Smith. Subsequent, Camacho (1996) has studied this aspect in more detail, drawing a clear and direct relationship between increases in the division of labour and rises in both communication and co-ordination costs, as an essential extension to the modern theory of the firm and the market. Pernin (1993) has treated inter-firm cooperation and its benefits from a similar angle, assessing the benefits for production from such cooperation as an economy of conventions and inter-firm agreements. This analysis thereby treats division of labour once again as part of the organizational theory of the firm or a production unit in which much emphasis is placed on the potential trade-offs between the economies reaped from specialization and the transaction costs it generates (Yang and Ng 1993). In this way, division of labour has also become an important part of the foundations for a new classical micro-economic analytic framework.

Conclusion

Viewed dynamically within the context of economic growth, as Smith (1776) and others had intended the division of labour to be contemplated, it continues to be a powerful tool for understanding the process of growth and development. On this ground alone it can therefore not be jettisoned from economics as unwanted baggage, as Robbins (1932) mistakenly suggested. When its importance for understanding aspects of the labour process, the labour market, the theory of production and the theory of the firm contemplated at the plant and the industry level are included, this argument is even stronger. As mentioned in the previous paragraph, on these grounds division of labour is making a definite come-back as part of the theory of a new classical micro-economics. Last, but not least, the importance of

the division of labour for economics is underlined by the fact that some of the major economic minds from both past and present have invariably included it as an important part of their economic analysis.

Bibliography

Amsden, A.H. 1985. The division of labour is limited by the rate of growth of the market; The Taiwan machine tool industry in the 1970s. *Cambridge Journal of Economics* 9: 271–284.

Anonymous. 1701. Considerations on the East-India trade. In *A select collection of early English tracts on commerce,* ed. J. R. McCulloch. London, 1856; Cambridge: Cambridge University Press, 1954.

Babbage, C. 1832. *On the economy of machinery and manufactures*. London. Fourth enlarged edition of 1835; New York: Augustus M. Kelley, 1963.

Becattini, G. 1990. The Marshallian industrial district as a socio-economic notion. In *Industrial districts and inter-firm co-operation in Italy,* eds. F. Pyke, G. Becattini, and W. Sengenberger. Geneva: International Institute of Labour Studies.

Becattini, G. 2001. *The caterpillar and the butterfly*. Florence: Felice de Monier.

Beccaria, C. 1771. Elementi di economiapublica. In *Opere*, ed. S. Romagnoli. Florence: Sansoni, 1958.

Becker, G.S. 1985. Human capital, effort, and the sexual division of labour. *Journal of Labour Economics* 3 (1): S33–S58.

Becker, G.S., and K.M. Murphy. 1992. The division of labour, coordination costs, and knowledge. *Quarterly Journal of Economics* 107: 1137–1160.

Bonar, J. 1893. *Philosophy and political economy,* 3rd ed. London: Allen and Unwin, 1967.

Braverman, H. 1974. *Labour and monopoly capital: The degradation of work in the twentieth century.* New York: Monthly Review Press.

Camacho, A. 1996. *Division of labour, variability, coordination and the theory of the firm and markets*. Dordrecht: Kluwer.

Campbell, R. 1747. *The London tradesman*. London.

Cannan, E. 1929. *Review of economic theory.* London: P.S. King & Son.

Carlyle, T. 1843. *Past and present.* London: Chapman & Hall. Another ed. London: G. Routledge & Sons, 1893.

Cigno, A. 1991. *Economics of the family.* Oxford: Oxford University Press.

Corsi, M. 1984. Ilsistema di fabbrica e la divisione del lavoro, ilpensiero di Charles Babbage. *Quaderni di storiadell'economiapolitica* 3: 111–123.

Dosi, G. 1988. Sources, procedures, and micro-economic effects of innovation. *Journal of Economic Literature* 26: 1120–1171.

Engels, F. 1884. Origin of the family, private property and the state. In *MarxEngels selected works*, vol. 2. Moscow: Progress Publishers.

Fawcett, H. 1863. *Manual of political economy.* London and Cambridge: Macmillan & Co.

Ferguson, A. 1767. *An essay on the history of civil society.* Edinburgh.

Fröbel, F., J. Heinrichs, and O. Kreye. 1980. *The new international division of labour.* Trans. P. Burgess. Cambridge: Cambridge University Press.

Gershuny, J. 1983. *Social innovation and division of labour*. Oxford: Oxford University Press.

Gioja, M. 1815. *Nuoveprospettodellescienzeeconomiche*. Milan: G. Pirotta.

Goodman, E., and J. Pamford, eds. 1989. *Small firms and industrial districts in Italy.* London: Routledge.

Gouverneur, J. 1978. *Contemporary capitalism and marxist economics.* Trans. R. le Farnu. Oxford: Robertson, 1983.

Groenewegen, P. 1977. Adam Smith and the division of labour: A bi-centenary estimate. *Australian Economic Papers* 16: 161–174.

Harris, J. 1757. *An essay upon money and coins.* Part I. London: G. Hawkins.

Hegel, G.W.F. 1821. *Philosophy of right.* Trans. T.M. Knox. Oxford: Clarendon Press, 1982.

Hicks, J.R. 1965. *Capital and growth*. Oxford: Clarendon Press.

Himmelweit, S., and S. Mohun. 1977. Domestic labour and capital. *Cambridge Journal of Economics* 1: 15–31.

Hobsbawm, E. 2000. *The new century.* London: Abacus.

Hodgskin, T. 1827. *Popular political economy.* London/New York: C. Tait/A.M. Kelley, 1966.

Hutcheson, F. 1755. *A system of moral philosophy.* Glasgow.

Hutchison, T.W. 1988. *Before Adam Smith: The emergence of political economy 1662–1776*. Oxford: Basil Blackwell.

Kaldor, N. 1966. *Causes of the slow rate of economic growth in the United Kingdom*. Cambridge: Cambridge University Press.

Kaldor, N. 1967. *Strategic factors in economic development*. Ithaca: State School of Industrial and Labour Relations, Cornell University.

Kaldor, N. 1972. The irrelevance of equilibrium economics. *Economic Journal* 82: 1237–1255.

Kaldor, N. 1975. What is wrong with economic theory? *Quarterly Journal of Economics* 89: 347–357.

Kames, H.H. 1774. *Sketches of the history of man*. Edinburgh/London: W. Creech/W. Strahan and T. Cadell.

Lindsay, P. 1733. *The interest of Scotland considered*. Edinburgh: R. Fleming & Co.

McCulloch, J.R. 1825. *Principles of political economy.* London: Murray, 1870.

McCulloch, J.R. 1850. Introduction and notes to Adam Smith. In *An inquiry into the nature and causes of the wealth of nations*, 4th ed. Edinburgh/London: A. & C. Black/Longman.

Mandeville, B. 1729. *The fable of the bees*. London: A. Roberts.
Marglin, S. 1974. What do bosses do? The origins and functions of hierarchy in capitalist production. *Review of Radical Political Economics* 6: 60–112.
Marshall, A. 1890. *Principles of economics,* 8th ed. London: Macmillan & Co., 1920.
Marshall, A. 1919. *Industry and trade*, 4th ed. London: Macmillan & Co., 1923.
Marshall, A., and M.P. Marshall. 1879. *The economics of industry.* London: Macmillan and Company.
Marx, K. 1844. *Economic and philosophic manuscripts*. Moscow: Foreign Languages Publishing House, 1959.
Marx, K. 1867. *Capital*, vol. 1. Moscow: Foreign Languages Publishing House, 1959.
Marx, K., and F. Engels. 1845–6. *The German ideology.* Moscow: Progress Publishers, 1964.
Meek, R.L., and A.S. Skinner. 1973. The development of Adam Smith's ideas on the division of labour. *Economic Journal* 83: 1094–1116.
Mill, J.S. 1821. Elements of political economy. In *The selected economic writings of James Mill,* 3rd ed, ed. D.N. Winch. Edinburgh: Oliver & Boyd for the Scottish Economic Society, 1966.
Mill, J.S. 1848. Principles of political economy. In *The collected works of John Stuart Mill,* vols. 2 and 3, ed. J.M. Robson. Toronto: Toronto University Press, 1965.
Nicholson, J.S. 1893. *Principles of political economy,* 2nd ed. London: A. & C. Black.
Perelman, M. 1984. *Classical political economy: primitive accumulation and the social division of labour*. London: Rowman & Allenheld.
Pernin, J.L. 1993. La cooperation entre firmes: uneapproche par l'économie des conventions. *Economie appliqué* 46: 105–126.
Petty, W. 1671. Political arithmetick. In *Economic writings of Sir William Petty*, ed. C.H. Hull. New York: A.M. Kelley, 1963.
Petty, W. 1683. Another essay on political arithmetick concerning the growth of the city of London. In *Economic writings of Sir William Petty*, ed. C.H. Hull. New York: A.M. Kelley, 1963.
Plato. 380 BC. *The republic.* Harmondsworth: Penguin Classics, 1955.
Posner, R.A. 1992. *Sex and reason*. Cambridge, MA: Harvard University Press.
Quesnay, F. 1765. Natural right. Extracts. Trans. R.L. Meek. *The economics of physiocracy.* London: Allen &Unwin, 1962.
Rae, J. 1834. *Statement of some new principles on the subject of political economy.*New York*:* A.M. Kelley, 1964.
Robbins, L. 1932. *An essay on the nature and significance of economic science.* London*:* Macmillan, 1935.
Romer, P.M. 1987. New theories of economic growth: Growth based on increasing returns due to specialization. *American Economic Review* 77: 56–62.
Rosenberg, N. 1976. Marx as a student of technology. *Monthly Review* 28: 56–77. In *Inside the black box: Technology and economics,* ed. N. Rosenberg. Cambridge: Cambridge University Press, 1982.
Rowthorn, R.E. 1975. A reply to Lord Kaldor's comment. *Economic Journal* 85: 897–901.
Ruskin, J. 1851–1853. The stones of Venice.In *The complete works of John Ruskin*, eds. E.T. Cook and A. Wedderburn. London: George Allen, 1904.
Salz, A. 1934. Specialisation. In *International encyclopaedia of the social sciences,* vol. 13, ed. E.R.A. Seligman. New York: Macmillan Company, 1948.
Scazzieri, R. 1981. *Efficienza, produttività e livelli di attitività*. Bologna: Il Mulino.
Schiller, F. 1793. *On the aesthetic education of man*. Trans. R. Snell. New York: Ungar, 1980.
Schumpeter, J.A. 1954. *History of economic analysis*. London: Oxford University Press.
Senior, N. 1836. *An outline of the science of political economy.* London: George Allen & Unwin, 1938, 1951.
Sidgwick, H. 1883. *Principles of political economy*, 2nd ed. London: Macmillan & Co.
Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*, ed. R.H. Campbell and A.S. Skinner. Oxford: Clarendon Press, 1976.
Stewart, D. 1858. In *Collected works of Sir Dugald Stewart*, ed. Sir William Hamilton. Edinburgh.
Sun, G.Z. 2005. *Readings in the economics of the division of labour*. Singapore: World Scientific Publishing Company.
Taussig, F.W. 1911. *Principles of economics,* 3rd ed. New York: Macmillan, 1936.
Thirlwall, A.P. 1983. A plain man's guide to Kaldor's growth law. *Journal of Post Keynesian Economics* 5: 345–358.
Torrens, R. 1808. *The economists refuted.* Sydney: Department of Economics, University of Sydney, Reprints of Economic Classics, 1984.
Tucker, J. 1755. *The elements of commerce and the theory of taxes*. London.
Tucker, J. 1774. *Four tracts on political and economic subjects,* 2nd ed. Gloucester: R. Raikes.
Turgot, A.R.J. 1751. Lettre à Madame de Graffignysur les letters d'un Péruvienne. In *Oeuvres de Turgot et documents le concernant*, vol. 1, ed. G. Schcllc. Paris: F. Alcan, 1913.
Turgot, A.R.J. 1766. Reflections on the production and distribution of wealth. In *The economics of A.R.J. Turgot*, ed. P.-D. Groenewegen. The Hague: Nijhoff, 1977.
Ure, A. 1835.*The philosophy of manufactures*. London: C. Knight/Frank Cass, 1967.
Verdoorn, P.J. 1949. Fattori che regolano lo sviluppo della produttività del lavoro. *L'industria* 1: 45–53.
West, E.G. 1964. Adam Smith's two views on the division of labour. *Economica* 31: 23–32.
Yang, X.-K., and Y.-K. Ng. 1993. *Specialisation and economic integration: A new classical micro-economic framework*. New York: Elsevier Science.
Young, A. 1928. Increasing returns and economic progress. *Economic Journal* 38: 327–342.

Dmitriev, Vladimir Karpovich (1868–1913)

D. M. Nuti

Vladimir Karpovich Dmitriev was the first Russian mathematical economist. His *Economic Essays on Value, Competition and Utility* (1898, 1902; English edition 1974) are a classic text in economic literature.

Vladimir Karpovich Dmitriev was born on 24 November 1868 on the Rai Estate in Smolensk Gubernia, Smolensk Uezd. On completing his classical education at the Tula Classical Gymnasium he went to Moscow University to study medicine but subsequently transferred to the Law Faculty where he began his studies in Political Economy. After graduation in 1896 he married T.A. Vatatsi and left to take the post of excise controller in the small town of Von'kovitsy in Podol'sk Gubernia. He served there for three years but contracted lung tuberculosis and had to leave the service. He was in great need all his life and his chronic illness eventually aggravated and caused his death on 30 November 1913.

Dmitriev's First Essay on *The theory of value of D. Ricardo* was published in 1898, followed in 1902 by a Second Essay on *The competition theory of A. Cournot* and a Third Essay on *The theory of marginal utility* (published together and with a Conclusion); the three essays were reprinted and issued together in 1904. He also published a large volume on the consumption of alcohol in Russia (with an introduction by P.V. Struve) in 1911 and half a dozen articles on the same topics as his books. He was planning at least three further Essays on rent, on industrial crises and on monetary circulation, which apparently were never written or at any rate published.

Dmitriev's contributions to economic theory include: (i) the development of an input-output method for the determination of the quantity of labour directly and indirectly embodied in commodities; (ii) a theory of production prices based on dated labour, similar to that of Piero Sraffa; (iii) a statement of 'wage-profit frontier' derived from technology and alternative level of real wage; (iv) a theory of non-productive costs in competition between firms. While these contributions do not amount quite to the 'organic synthesis of the labour theory of value and the theory of marginal utility' promised in the title page of the 1902 and 1904 editions of the *Essays*, they are highly original and remarkable in their anticipation of subsequent work. Dmitriev's propositions on labour values and production prices gained early recognition (Chuprov 1905; Bortkiewicz 1906, who praised and used extensively 'this remarkable work'; Struve 1908, who hailed Dmitriev as a 'logically and mathematically thought-out Ricardo'; and Shaposhnikov's memorial lecture a year after his death, 1914). Until shortly after the October revolution Dmitriev was widely mentioned in Russian economic literature, then he was entirely forgotten until the Soviet school of mathematical economists brought him out of his official oblivion circa 1960 (Nemchinov 1959; Belkin, Grobman, Lunts, in Aganbegyan and Belkin 1961) attracting the attention of Western scholars (Nove and Zauberman 1961; Zauberman 1962).

In his first Essay Dmitriev considers the question 'how is it possible to calculate the amount of labour expended for the production of a given economic good from the very beginning of history, when man managed without capital, down to the present time' (p. 43 of the English edition, to which all page references are made here). He answers that there is no need for 'historical digressions' of this kind; the quantity of labour N_A which goes directly and indirectly into the production of commodity A is expressed by the equation

$$N_A = n_A + \frac{1}{m_1}N_1 + \frac{1}{m_2}N_2 + \cdots + \frac{1}{m_M}N_M \quad (1)$$

where n_A is the *direct* labour input of a unit of commodity A: $1/m_i$ is the amount of the i th commodity *used up* in the production of commodity A, where $i = 1, 2, \ldots, M$; and N_i is the labour directly and indirectly embodied in the ith commodity (this is Eq. 6 in the First Essay, p. 44). The coefficient $1/m_i$ here is to be interpreted either as

the intermediate inputs requirement for the production of the A commodity, or as the straight-line amortization of the i th fixed capital good (assuming uniform productiveness over its lifetime); some of these coefficients may be equal to zero, as in Dmitriev's system of Eq. 7 in the First Essay. For each of the M other commodities there is an equation of the same form, relating labour (directly and indirectly) embodied to input coefficients and the labour embodied in the inputs (p. 44). We obtain a system of $(M + 1)$ equations in $(M + 1)$ unknowns,

> which is always adequate for the determination of N, giving the required sum of the labour expended on the production of product A. Therefore, without any digressions into the prehistoric times of the first inception of technical capital, we can always find the total sum of the labour directly and indirectly expended on the production of any product *under present day production conditions*, both of this product itself and of those capital goods involved in its production (p. 44, emphasis in the text).

This is clearly a full-fledged input-output system, where N_i are the full coefficients of labour, the N_i are the direct labour inputs, and the $1/m$ are identical with Leontief's input-output coefficients. The analytical apparatus provided by Leontief four decades later adds two things: (i) a method for the actual computation of the solution, namely the inversion of the matrix $(I\text{-}A')$, where I is the identity matrix and A' is the transpose of the matrix of technical coefficients; and (ii) the generalization of notion of full input (i.e. direct and indirect input requirements) from labour to other production inputs. In Leontief's type of notation, if we call a_{ij} the amount of i th product required per unit of the j th product, A the $[a_{ij}]$ matrix: a_{oj} the direct labour input of product j, and a the column vector $[a_{oj}]$; and f_{ij} the full-input coefficient, i.e. the element of the $(I\text{-}A')^{-1}$ matrix, we obtain

$$f_{ik} = \sum_{j=1}^{n} a_{ij} f_{ik} + \delta_{ik} \tag{2}$$

where $i, k, j = 1, 2, \ldots, n$; and f_{ik} is Kronecker's delta, i.e. is equal to zero except for $i = k$ when it is equal to unity. If we indicate full labour inputs (i.e. Dmitriev's Ns) by f_{ok}, Leontief's approach gives

$$f_{ok} = \sum_{j=1}^{n} a_{oi} f_{jk} \tag{3}$$

or

$$\mathbf{f}_o = (\mathbf{I} - \mathbf{A}\prime)^{-1}\mathbf{a} \tag{4}$$

where $F_o = [f_{ok}]$. Dmitriev's formulation of full labour inputs is

$$f_{ok} = a_{ok} + \sum_{j=1}^{n} f_{oj} a_{jk} \tag{3'}$$

or

$$\mathbf{f}_o = \mathbf{a} + \mathbf{A}'\mathbf{f}_o \tag{4'}$$

which is just another way of rewriting Leontief's Eq. 4.

The importance of Dmitriev's approach for socialist planning was already understood in the 1920s; A.V. Chayanov (1926) developed Dmitriev's scheme into an input-output table for agriculture. In the 1960s the ability to claim Russian priority in the discovery of input-output equations in the work of Dmitriev was an important step in the struggle for the use of mathematical methods in socialist planning. In 1962 the Central Statistical Administration produced an 83 × 83 intersectoral balance of labour outlays in the Soviet economy for 1959–60, using the first ex-post input-output tables for the Soviet economy, compiled for 1959. This balance shows, in terms of labour, the inter-industrial flows, the formation of the final bill of goods, the formation of national product and cost incurred in the non-productive sphere (see Eidel'man 1962; Zauberman 1963). This calculation corresponds exactly to the Dmitriev-Leontief full labour coefficients.

Dmitriev also had a theory of *prices of production* which is a reformulation and development of Ricardian price theory and corresponds to Marxian production prices. Dmitriev starts from the

D

refutation of the criticism levied in his time (for instance by Walras) against the 'classical' theory of price determination based on production costs, 'that it defines price from prices, that it defines one unknown from other unknowns' (p. 41). This allegation, Dmitriev argues, can be levied against Adam Smith, who did not deal with the problem of the determination of the profit rate, except for a vague reference to the demand for and supply of capital, i.e. going outside the sphere of production. But Ricardo is not subject to this criticism; indeed 'The most important point in Ricardo's theory is undoubtedly his theory of the conditions defining the "average" profit rate …' and 'Ricardo's immortal contribution was his brilliant solution of this seemingly insoluble problem' (pp. 50 and 58, First Essay).

For the study of *prices* (or *values*, in his terminology) Dmitriev uses a framework slightly different from that employed for the study of *labour values* (or labour embodied in commodities). Instead of extending his *point input-point output* framework, whereby commodities are produced by means of labour and other commodities (Eq. 1), he uses an Austrian-type model where commodities are produced by dated labour, i.e. a *flow input-point output* framework, whereby commodities are produced by dated labour. For each commodity Dmitriev formulates a price equation of the type:

$$X_A = n_A a X_a (1+r)^{tA} + n_1 a X_a (1+r)^{tA1} + \mathrm{L} \\ \mathrm{L} + n_m a X_a (1+r)^{tAm} \tag{5}$$

where X_A is the price of commodity A, a is the amount of wage good (say, corn) consumed by workers, X_a is the unit price of the wage good; $n_A, n_1, \ldots, n_m$ are the labour inputs required respectively $t_A, t_{A1}, \ldots, t_{\mathrm{Am}}$ time units before the output of commodity A becomes available (this is Eq. 25, p. 54). If there are M commodities in addition to the wage good, we have $(M + 1)$ equations; there are M relative prices to be determined, in terms of an arbitrary commodity whose price is taken as unit of account, plus the profit rate; the system is complete and can simultaneously determine relative price and the profit rate.

> It is to Ricardo's credit that he was the first to note that there is one production equation by means of which we may determine the magnitude of *r directly* (i.e. without having recourse for assistance to the other equations). This equation gives us the production conditions of the product *a* to which in the final analysis the expenditure on all the products, *A*, *B*, *C*, . . ., is reduced (p. 59).

For the wage good, with labour inputs N_i,

$$X_a = a X_a \left[N_a (1+r)^{t_a} + N_t (1+r)^{t_{a1}} + \cdots + N_1 (1+r)^{t_{aq}} \right]. \tag{6}$$

From this (Eq. 44, First Essay) we can obtain

$$a_i = \frac{1}{\sum_i N_i (1+r)^i} \tag{7}$$

which today is familiar as the 'wage-profit frontier': Dmitriev writes it instead in the implicit form

$$r = F\left(N_a, N_1, \ldots, N_q; t_a, t_{a1}, \ldots, t_{aq}; a\right). \tag{8}$$

Dmitriev then extends this analysis to the case where workers consume not a single commodity but a number of commodities in fixed proportions. The condition for a positive profit rate to arise is that 'we can obtain a *larger* quantity of the same product within some finite period of time as a result of the production process' (p. 62).

Dmitriev, in sum, considers 'production of commodities by means of dated labour', not 'production of commodities by means of commodities' (at least when discussing the determination of the profit rate), with wages being advanced, not 'posticipated' as in Sraffa (1960). Their similarity descends from the common Ricardian root. Although Dmitriev's approach is close to Marx, he goes out of his way to *deny* the Marxian theory of exploitation and to show, 'proceeding from Ricardo's analysis, that the origin of industrial profit does not stand in any "special" relationship to the human labour used in production' (p. 64), In order to do this, Dmitriev investigates the properties of an imaginary system where work is performed exclusively by animals and machines. The conditions

for a positive profit rate are shown to be quite general; however, the fact that we do not usually talk of 'exploitation' of animals and machines does not in any conceivable sense rule out the proposition of *human* exploitation when *human* labour *is* actually used in production.

Having formulated and developed Ricardian propositions on prices of production Dmitriev proceeds to show that these propositions hold only under the most restrictive assumptions. Among these are constant returns to scale, i.e. zero rents, *and* perfect competition of a kind that brings prices down to the (constant) necessary costs of commodities (including profit at a rate determined by technology and the real wage). He decidedly parts company from Ricardo and shows that whenever at least one of these conditions is not satisfied prices depend on *demand conditions* as well, and not even 'long-run' equilibrium prices can be obtained purely from the knowledge of technology and the real wage.

Already at the end of the First Essay, Dmitriev shows that a demand price based exclusively on production conditions cannot handle the cases of monopoly prices and of positive rent. But the greatest blow to the Ricardian theory of price determination is given in the Second Essay, where Dmitriev most emphatically argues that demand conditions contribute to price determination also for 'goods which are infinitely reproducible by labour under conditions excluding the possibility of the occurrence of rent' (p. 92) even under competitive conditions. In order to do this, Dmitriev challenges the proposition that 'competition lowers prices' (p. 93) and starting from Cournot's analysis of competition he constructs a theory of unrestricted but not-so-perfect competition.

Dmitriev argues that the assumption that supply = production contradicts not only economic reality, but also the other basic hypothesis of competitive analysis, 'that every individual tends to pursue the greatest advantage' (p. 118). He relaxes the assumption to allow for stocks and unused capacity, representing *potential* supply. Dmitriev postulates that *for a given volume of production* rational behaviour of producers leads them to a tacit collusion on price, i.e. joint profit maximisation as in the monopoly case, but (i) such collusion is enforceable only because of the existence of a potential threat in the form of a potential supply greater than the collusion sales level, and (ii) competition between producers takes the form of expanding the level of potential supply, with sales lagging behind. For a given number n of producers there is an equilibrium potential supply such that the price corresponds to what would be charged by a monopolist. For n tending to infinity, the cost of the potential supply tends to equal the revenue from actual sales; profit (over and above the interest component of production costs) is zero, as in the customary competitive equilibrium, not because price is equal to the necessary production cost of the output sold, but because the additional cost of holding stocks or installing unused capacity brings the total cost of potential output up to the level of actual sales revenue and wipes out profits completely (p. 134).

A further instance of unproductive expenditure is mentioned by Dmitriev in his Conclusion, namely 'advertising' to expand sales of an individual entrepreneur *when the total sales level remains the same*. In a notable passage Dmitriev compares the role of commodity stocks with the strategy of 'intensified armament of the Powers in peace time' (pp. 148–9). It follows from this analysis that unrestricted competition has a cost for the economy, i.e. a *social* cost of wasted output, excess inventories, unused capacity or redundant advertising. This is only partly compensated by consumers' gain from prices lower than monopoly prices.

A most important implication of Dmitriev's analysis is his account of the economic consequences of technical progress (Section 7), which raises the level of potential supply at which the *temporary profit*, obtained by individual producers breaking their tacit price-collusion, disappears. 'Therefore an *expansion of output following a reduction of production costs will*, *in general*, *extend not only to an expansion of supply but also to an increase in excess commodity inventories* (p. 171). The building up of excess commodity inventories following technical progress

gives rise to fluctuations in the levels of output capacity, capacity utilization, and inventory levels (pp. 173–8). When technical progress takes place, 'over-production' periodically occurs, and this '*is in no sense a result of errors of economic judgement*, i.e. it is not a consequence of the inability of production to adapt to excessively variable demand . . . but is a direct result of the struggle of competing entrepreneurs, *each of whom is motivated in his own actions by quite correct economic judgement*' (p. 117).

The only way of eliminating wasted output, excess inventories and unused capacity, and the non-productive costs which these involve, is the establishment of forward markets (*Terminhandel*): 'forward contracts make non-productive "reserve stocks" unnecessary since they make it possible to sell goods which have still not been produced but merely can be produced . . .' (p. 178, footnote 1). Dmitriev relegates this qualification to a footnote, but this is really a central point in his argument, except that – we now know – forward markets would have to be not only complete but also exclusive (i.e. no future spot markets could reopen), which is neither practical nor advisable.

For Dmitriev the short-run equilibrium of an economic system is determined by the given levels of supply and the demand functions. He concludes that if prices of commodities happen to coincide with their necessary reproduction costs, actual prices will correspond to the solution of the Walrasian system. But if the supply level of a commodity is such that its price exceeds its necessary reproduction costs, the question of the distribution of the extra-normal profit lies, for Dmitriev, 'outside the sphere of economic research', because it is the result of a 'struggle' and is taken as a question of fact by economic theory. There may be 'a general sociological solution' (p. 207); 'Otherwise we should have to admit that the question cannot have any general solution at all' (*ibid*). Ultimately, price theory becomes the theory of the self-defeating attempts, by economic agents, to gain from a social struggle which is rational by the standards of individuals though not of society, and the theory of the ensuing waste and fluctuations.

Bibliography

Belkin, V.D. 1961. Natsionalnyi dokhod i mezhotraslevoy balans [National income and intersectoral balance]. In *Primenenie matematiki i elektronnoy tekhniki v planirovanii* [The use of mathematics and electronic techniques in planning], ed. A.G. Aganbegyan and V.D. Belkin. Moscow.

Bortkiewicz von L. 1906. Wertrechnung und Preisrechnung im Marxschen Sistem, (in three parts). *Archiv für Sozialwissenschaft und Sozialpolitik* 23(1) 1907, 25(1); 1907, 25(2). The second and the third parts are translated into English, as 'Value and Price in the Marxian System', *International Economic Papers,* 1952, no. 2.

Eidel'man, M.R. 1962. Pervyi mezhotraslevoi balans zatrat truda v narodnom khoziaistve SSSR [The first intersectoral balance of labour expenditures in the national economy of the USSR]. *Vestnik Statistiki,* no. 10, 1962.

Leontief, W.W. 1941. *The structure of the American economy 1919–1939*. New York: Oxford University Press.

Leontief, W.W., et al. 1953. *Studies in the structure of the American economy: Theoretical and empirical explorations in input-output analysis*. White Plains: International Arts and Science Press.

Nemchinov, V.S. 1959. The use of mathematical methods in economics. In *The use of mathematics in economics* (English trans: Nove, A.)., ed. V.S. Nemchinov Moscow. London, 1964.

Nemchinov, V.S. 1961. *A model of an economic region*. Moscow. Trans. *Mathematical studies in economics and statistics in the USSR and Eastern Europe*. vol. 1, 1964, p. 14.

Nemchinov, V.S. 1963. Basic elements of a model of planned price formation. *Voprosy Ekonomiki,* no. 12. *Socialist economics* (trans: Nove, A., Nuti, D.M.). Harmondsworth: Penguin, 1972.

Nove, A., and A. Zauberman. 1961. A resurrected Russian economist of 1900. *Soviet Studies* 13: 96–101.

Shaposhnikov, N.N. 1914. *Pervyi Russkii ekonomist-matematik Vladimir Karpovich Dmitriev, Doklad v posvyashchennom pamyati Dmitrieva zasêdanii O-va im. A.I. Chuprova* [The first Russian mathematical economist V.K. Dmitriev, a lecture as a meeting of the A.I. Chuprov Society, held in memory of Dmitriev]. Moscow.

Sraffa, P. 1960. *Production of commodities by means of commodities*. Cambridge: Cambridge University Press.

Zauberman, A. 1962. A few remarks on a discovery in Soviet economics. *Bulletin of the Oxford Institute of Economics and Statistics* 24: 437–445.

Zauberman, A. 1963. A note on the Soviet inter-industry labour input balance. *Soviet Studies* 15: 53–57.

Selected Works

Dmitriev, Vladimir Karpovich. 1898. *Ekonomicheskie Ocherki,* Vyp, I, 'Teoriya tsênnosti D. Ricardo (opyt' tochnago analyza)' [*Economic Essays,* Issue I, 'The

theory of value of D. Ricardo, an attempt at a rigorous analysis']. Moscow.
Dmitriev, Vladimir Karpovich. 1902. *Ekonomicheskie Ocherki,* 'Chast' l-aya (opyt' organicheskago sinteza trudovoi teorii tsênnosti i teorii predêl'noi poleznosti)', Vypuski 2-i i 3-i. Ocherk 2-i: 'Teoriya konkurrentsii Og. Kurno (Velikago "zabytago" ekonomista)'. Ocherk 3-i: 'Teoriya predêl'noi poleznosti' [*Economic Essays,* Part I, Attempt at an organic synthesis of the labour theory of value and the theory of marginal utility, Issues 2 and 3. Second Essay: The theory of competition of A. Cournot (the great 'forgotten' economist). Third Essay: The theory of marginal utility]. Moscow.
Dmitriev, Vladimir Karpovich. 1904. *Ekonomicheskie Ocherki* (Seriya I-aya: 'opyt' organicheskago sinteza trudovoi toerii tsênnosti i teorii predêl'noi poleznosti') [*Economic Essays,* First Series: Attempt at an organic synthesis of the labour theory of value and the theory of marginal utility]. Moscow.
Dmitriev, Vladimir Karpovich. 1911. *Kriticheskie izslêdovaniya o potreblenii alkogolya v Rossii,* s predisl. P.B. Struve, Issledovaniya i raboty po polit. ekonomii i obshchestv. znaniyam, izd. pod red. P.B. Struve, Vyp. I [*Critical Studies on the consumption of alcohol in Russia,* with an introduction by P.V. Struve: Studies and works in political economy and social sciences, edited by P.V. Struve, Issue I]. Moscow. (With an English translation of the table of contents).
Dmitriev, Vladimir Karpovich. 1974. *Economic essays on value, competition and utility.* Edited with an introduction by D.M. Nuti. Cambridge: Cambridge University Press.

Dobb, Maurice Herbert (1900–1976)

Amartya Sen

Keywords

Dobb, M.; Economic development; Equality; Exploitation; Feudalism; Labour theory of value; Langer–Lerner price mechanism; Market socialism; Marxism; Planning; Revealed preference theory; Saving and investment; Socialist pricing theory; Sraffa, P.; Sweezy, P. M.; Utility theory of value

JEL Classifications

B31

Maurice Dobb was undoubtedly one of the outstanding political economists of this century. He was a Marxist, and was one of the most creative contributors to Marxian economics. As Ronald Meek put it, in his obituary of Dobb for the British Academy, 'over a period of fifty years [Dobb] established and maintained his position as one of the most eminent Marxist economists in the world'. Dobb's *Political Economy and Capitalism* (1937) and *Studies in the Development of Capitalism* (1946) are his two most outstanding contributions to Marxian economics. The former is primarily concerned with economic theory (including such subjects as value theory, economic crises, imperialism, socialist economies), and the latter with economic history (particularly the emergence of capitalism from feudalism). These two fields – economic theory and economic history – were intimately connected in Dobb's approach to economics. He also wrote an influential book on Soviet economic development. This was first published under the title *Russian Economic Development since the Revolution* (1928), and later in a revised edition as *Soviet Economic Development since 1917* (1948).

Maurice Dobb was born on 24 July 1900 in London. His father Walter Herbert Dobb had a draper's retail business and his mother Elsie Annie Moir came from a Scottish merchant's family. He was educated at Charterhouse, and then at Pembroke College, Cambridge, where he studied economics. This was followed by two postgraduate years at the London School of Economics, where he did his Ph.D. on 'The Entrepreneur'. The thesis formed the basis of his book *Capitalist Enterprise and Social Progress* (1925). Dobb returned to Cambridge at the end of 1924 on being appointed as a lecturer in economics. He taught in Cambridge until his retirement in 1967. He was a Fellow of Trinity College, and was elected to a University Readership in 1959. He received honorary degrees from the Charles University of Prague, the University of Budapest, and Leicester University, and was elected a Fellow of the British Academy. After retirement he and his wife, Barbara, stayed on in the neighbouring village of Fulbourn. He died on 17 August 1976.

Dobb was a theorist of great originality and reach. He was also, throughout his life, deeply concerned with economic policy and planning. His foundational critique of 'market socialism' as developed by Oscar Lange and Abba Lerner, appeared in the *Economic Journal* of 1933, later reproduced along with a number of related contributions in his *On Economic Theory and Socialism* (1955). His relatively elementary book *Wages* (1928) presented not merely a simple introduction to labour economics, but also an alternative outlook on these questions, including their policy implications, leading to interesting disputations with John Hicks, among others. In later years Dobb was much concerned with planning for economic development. In three lectures delivered at the Delhi School of Economics, later published as *Some Aspects of Economic Development* (1951), Dobb discussed some of the central issues of development planning for an economy with unemployed or underutilized labour, and his ideas were more extensively developed in his later book, *An Essay on Economic Growth and Planning* (1960).

Maurice Dobb also published a number of papers on more traditional fields in economic theory, including welfare economics, and some of these papers were collected together in his *Welfare Economics and the Economics of Socialism* (1969). In his *Theories of Value and Distribution since Adam Smith*: *Ideology and Economic Theory* (1973), he responded *inter alia* to the new developments in Cambridge political economy, including the influential 'Prelude to a Critique of Economic Theory' by Piero Sraffa (1960). Maurice Dobb's association with Piero Sraffa extended over a long period, both as a colleague at Trinity College, and also as a collaborator in editing *Works and Correspondence of David Ricardo*, published in 11 volumes between 1951 and 1973 (on the latter, see Pollitt 1990).

In addition to academic writings, Maurice Dobb also did a good deal of popular writing, both for workers' education and for general public discussion. He wrote a number of pamphlets, including *The Development of Modern Capitalism* (1922), *Money and Prices* (1924), *An Outline of European History* (1926), *Modern Capitalism* (1927), *On Marxism Today* (1932), *Planning and Capitalism* (1937), *Soviet Planning and Labour in Peace and War* (1942), *Marx as an Economist*: *An Essay* (1943), *Capitalism Yesterday and Today* (1958), and *Economic Growth and Underdeveloped Countries* (1963), and many others. Dobb was a superb communicator, and the nature of his own research was much influenced by policy debates and public discussions. Dobb the economist was not only close to Dobb the historian, but also in constant company of Dobb the member of the public. It would be difficult to find another economist who could match Dobb in his extraordinary combination of genuinely 'high-brow' theory, on the one hand, and popular writing on the other. The author of *Political Economy and Capitalism* (from the appearance of which – as Ronald Meek (1978) rightly notes – 'that future historians of economic thought will probably date the emergence of Marxist economics as a really serious economic discipline': was also spending a good deal of effort writing pamphlets and material for labour education, and doing straightforward journalism. It is not possible to appreciate fully Maurice Dobb's contributions to economics without taking note of his views of the role of economics in public discussions and debates.

Another interesting issue in understanding Dobb's approach to economics concerns his adherence to the labour theory of value. The labour theory has been under attack not only from neoclassical economists, but also from such anti-neoclassical political economists as Joan Robinson and, indirectly, even Piero Sraffa. In his last major work, *Theories of Value and Distribution since Adam Smith* (1973), Maurice Dobb speaks much in support of the relevance of Sraffa's (1960) major contribution, which eschews the use of labour values (on this see Steedman 1977), but without abandoning his insistence on the importance of the labour theory of value. It is easy to think that there is some inconsistency here, and it is tempting to trace the origin of this alleged inconsistency to Dobb's earlier writings, which made Abram Bergson remark that 'in Dobb's analysis the labour theory is not so much an analytic tool as excess baggage' (Bergson 1949, p. 445).

The key to understanding Dobb's attitude to the labour theory of value is to recognize that he did not see it just as an intermediate product in explaining relative prices and distributions. He took 'the labour-principle' as 'making an important qualitative statement about the nature of the economic problem' (Dobb 1937, p. 21). He rejected seeing the labour theory of value as simply a 'first approximation' containing 'nothing essential that cannot be expressed equally well and easily in other terms' (Dobb 1973, pp. 148–9). The description of the production process in terms of labour involvement has an interest that extends far beyond the role of the labour value magnitudes in providing a 'first approximation' for relative prices. As Dobb (1973, pp. 148–9) put it,

> there is something in the first approximation that is lacking in later approximations or cannot be expressed so easily in those terms (e.g., the first approximation may be a device for emphasising and throwing into relief something of greater generality and less particularity).

Any description of reality involves some selection of facts to emphasize certain features and to underplay others, and the labour theory of value was seen by Dobb as emphasizing the role of those who are involved in 'personal participation in the process of production *per se*' in contrast with those who do not have such personal involvement.

> As such 'exploitation' is neither something 'metaphysical' nor simply an ethical judgement (still less 'just a noise') as has sometimes been depicted: it is a factual description of a socio-economic relationship, as much as is Marc Bloch's apt characterisation of Feudalism as a system where feudal Lords 'lived on the labour of other men'. (Dobb 1973, p. 145.

The possibility of calculating prices without going through value magnitudes, and the greater efficiency of doing that (on this see Steedman 1977), does not affect this descriptive relevance of the labour theory of value in any way. Maurice Dobb also outlined the relationship of this primarily descriptive interpretation of labour theory of value with evaluative questions, for example, assessing the 'right of ownership' (see especially Dobb 1937).

The importance for Dobb of descriptive relevance is brought out also by his complex attitude to the utility theory of value. While he rejected the view that the utility picture is the best way of seeing relative values ('by taking as its foundation a fact of individual consciousness'), he lamented the descriptive impoverishment that is brought about by replacing the subjective utility theory by the 'revealed preference' approach.

> If all that is postulated is simply that men *choose,* without anything being stated even as to how they choose or what governs their choice, it would seem impossible for economics to provide us with any more than a sort of algebra of human choice. (Dobb 1937, p. 171.

Indeed, as early as 1929, a long time before the 'revealed preference theory' was formally inaugurated by Paul Samuelson, Dobb (1929, p. 32) had warned:

> Actually the whole tendency of modern theory is to abandon such psychological conceptions: to make utility and disutility coincident with observed offers in the market; to abandon a 'theory of value' in pursuit of a 'theory of price'. But this is to surrender, not to solve the problem.

Maurice Dobb's open-minded attitude to non-Marxian traditions in economics added strength and reach to his own Marxist theorizing. He could combine Marxist reasoning and methodology with other traditions, and he was eager to be able to communicate with economists belonging to other schools. Dobb's honesty and lack of dogmatism were important for the development of the Marxist economic tradition in the English-speaking world, because he occupied a unique position in Marxist thinking in Britain. As Eric Hobsbawm (1967, p. 1) has noted,

> for several generations (as these are measured in the brief lives of students) he was not just the only Marxist economist in a British university of whom most people had heard, but virtually the only don known as a communist to the wider world.

The Marxist economic tradition was well served by Maurice Dobb's willingness to engage in spirited but courteous debates with economists of other schools. Dobb achieved this without compromising the integrity of his position. The distinctly Marxist quality of his economic writings was as important

as his willingness to listen and dispassionately analyse the claims of other schools of thought with which he engaged in systematic disputation. The gentleness of Dobb's style of disputation arose from strength rather than from weakness.

Dobb's willingness to appreciate positive elements in other economic traditions while retaining the distinctive qualities of his own approach is brought out very clearly also in his truly far-reaching critique of the theory of socialist pricing as presented by Lange, Lerner, Dickinson and others in the 1930s. Dobb noted the efficiency advantages of a price mechanism, especially in a static context. He was, however, one of the first economists to analyse clearly the conflict between the demands of efficiency expressed in the equilibrium conditions of the Langer–Lerner price mechanism (and also of course in a perfectly competitive market equilibrium), and the demands that would be imposed by the requirements of equality, given the initial conditions. In his paper called 'Economic Theory and the Problems of a Socialist Economy' published in 1933, Maurice Dobb argued thus:

> If carpenters are scarcer or more costly to train than scavengers, the market will place a higher value upon their services, and carpenters will derive a higher income and have greater 'voting power' as consumers. On the side of supply the extra 'costliness' of carpenters will receive expression, but only at the expense of giving carpenters a differential 'pull' as consumers, and hence vitiating the index of demand. On the other hand, if carpenters and scavengers are to be given equal weight as consumers by assuring them equal incomes, then the extra costliness of carpenters will find no expression in costs of production. Here is the central dilemma. Precisely because consumers are also producers, both costs and needs are precluded from receiving simultaneous expression in the same system of market valuations. Precisely to the extent that market valuations are rendered adequate in one direction they lose significance in the other. (1933, p. 37)

The fact that given an initial distribution of resources the demands of efficiency and those of equity may – and typically will – conflict is, of course, one of the major issues in the theory of resource allocation, with implications for market socialism as well as for competitive markets in a private ownership economy. As a matter of fact, Marx had *inter alia* noted this conflict in his *Critique of Gotha Programme,* but in the discussion centring around Langer–Lerner systems, this deep conflict had attracted relatively little attention, except in the arguments presented by Maurice Dobb. The fact that even a socialist economy has to cope with inequalities of initial resource distribution (arising from, among other things, differences in inherited talents and acquired skills) makes it a relevant question for a socialist economy as well as for competitive market economies, and Dobb's was one of the first clear analyses of this central question of resource allocation.

The second respect in which Maurice Dobb found the literature on market socialism inadequate concerns allocation over time. In discussing the achievements and failures of the market mechanism, Maurice Dobb argued that the planning of investment decisions

> may contribute much more to human welfare than could the most perfect micro-economic adjustment, of which the market (if it worked like the textbooks, at least, and there were no income-inequalities) is admittedly more fitted in most cases to take care. (Dobb 1960, p. 76)

In his book *An Essay in Economic Growth and Planning* (1960), Dobb provided a major investigation of the basis of planned investment decisions, covering overall investment rates, sectoral divisions, choice of techniques, and pricing policies related to allocation (including that over time).

This contribution of Dobb relates closely to his analysis of the problems of economic development. In his earlier book *Some Aspects of Economic Development* (1951), Dobb had already presented a pioneering analysis of the problem of economic development in a surplus-labour economy, with shortage of capital and of many skills. While, on the one hand, he anticipated W.A. Lewis's (1954) more well-known investigation of economic growth with 'unlimited supplies of labour', he also went on to demonstrate the far-reaching implications of the over-all savings rates being socially sub-optimal and inadequate. Briefly, he showed that this requires not only policies directly aimed at raising the rates of saving and investment, but it also has implications for the choice of techniques, sectoral balances, and price fixation.

In such a brief note, it is not possible to do justice to the enormous range of Maurice Dobb's contributions to economic theory, applied economics and economic history. Different authors influenced by Maurice Dobb have emphasized different aspects of his many-sided works (see, for example, Feinstein, 1967, and the *Cambridge Journal of Economics'* Maurice Dobb Memorial Issue (1978)). He has also had influence even outside professional economics, particularly in history, especially through his analysis of the development of capitalism.

Dobb argued that the decline of feudalism was caused primarily by 'the inefficiency of Feudalism as a system of production, coupled with the growing needs of the ruling class for revenue' (1946, p. 42). This view of feudal decline, with its emphasis on *internal* pressures, became the subject of a lively debate in the early 1950s. An alternative position, forcefully presented by Paul Sweezy in particular, emphasized some *external* developments, especially the growth of trade, operating through the relations between the feudal countryside and the towns that developed on its periphery. No matter what view is taken as to 'who won' the debates on the transition from feudalism to capitalism, Dobb's creative role in opening up a central question in economic history as well as a major issue in Marxist political economy can scarcely be disputed. Indeed, *Studies in the Development of Capitalism* (1946) has been a prime mover in the emergence of the powerful Marxian tradition of economic history in the English-speaking world, which has produced scholars of the eminence of Christopher Hill, Rodney Hilton, Eric Hobsbawm, Edward Thompson and others.

It is worth emphasizing that aside from the explicit contributions made by Maurice Dobb to economic history, he also did use a historical approach to economic analysis in general. Maurice Dobb's deep involvement in descriptive richness (as exemplified by his analysis of 'the requirements of a theory of value'), his insistence on not neglecting the long-run features of resource allocation (influencing his work on planning as well as development), his concern with observed phenomena in slumps and depressions in examining theories of 'crises', and so on, all relate to the historian's perspective. Dobb's works in the apparently divergent areas of economic theory, applied economics and economic history are, in fact, quite closely related to each other.

Maurice Dobb was not only a major bridge-builder between Marxist and non-Marxist economic traditions (aside from pioneering the development of Marxist economics in Britain and to some extent in the entire English-speaking world): he also built many bridges between the different pursuits of economic theorists, applied economists and economic historians. Dobb's political economy involved the rejection of the narrowly economic as well as the narrowly doctrinaire. He was a great economist in the best of the broad tradition of classical political economy.

Selected Works

1925. *Capitalist enterprise and social progress*. London: Routledge.

1928. *Russian economic development since the revolution*. London: Routledge.

1928. *Wages*. London: Nisbet; Cambridge: Cambridge University Press.

1929. A sceptical view of the theory of wages. *Economic Journal* 39: 506–519.

1933. Economic theory and the problems of a socialist economy. *Economic Journal* 43: 588–598.

1937. *Political economy and capitalism: Some essays in economic tradition.* London: Routledge.

1946. *Studies in the development of capitalism*. London: Routledge.

1948. *Soviet economic development since 1917.* London: Routledge.

1950. Reply (to Paul Sweezy's article on the transition from feudalism to capitalism). *Science and society* 14(2): 157–167.

1950. *Some aspects of economic development*: *Three lectures.* Delhi: Ranjit Publishers, for the Delhi School of Economics.

1955. *On economic theory and socialism*. London: Routledge.

1960. *An essay on economic growth and planning*. London: Routledge.

1969. *Welfare economics and the economics of socialism.* Cambridge: Cambridge University Press.
1973. *Theories of value and distribution since Adam Smith: Ideology and economic theory.* Cambridge: Cambridge University Press.

Bibliography

Bergson, A. 1949. Socialist economics. In *A survey of contemporary economics*, ed. H.S. Ellis. Philadelphia: Blakiston.
Cambridge Journal of Economics. 1978. Maurice Dobb memorial issue. vol. 2(2), June.
Hobsbawm, E.J. 1967. Maurice Dobb. In *Socialism, capitalism and economic growth: Essays presented to Maurice Dobb*, ed. C. Feinstein. Cambridge: Cambridge University Press.
Lewis, W.A. 1954. Economic development with unlimited supplies of labour. *Manchester School* 20 (2): 139–191.
Meek, R. 1978. Obituary of Maurice Herbert Dobb. *Proceedings of the British Academy 1977* (53): 333–344.
Pollitt, B.H. 1990. Clearing the path for 'Production of Commodities by Means of Commodities': Notes on the collaboration of Maurice Dobb in Piero Sraffa's edition of 'The Works and Correspondence of David Ricardo'. In *Essays on Piero Sraffa: Critical perspectives on the revival of classical theory*, ed. K. Bharadwaj and B. Schefold. London: Unwin Hyman.
Sraffa, P. 1960. *Production of commodities by means of commodities: Prelude to a critique of economic theory.* Cambridge: Cambridge University Press.
Sraffa, P. with the collaboration of M.H. Dobb. 1951–73. *Works and correspondence of David Ricardo*, 11 vols. Cambridge: Cambridge University Press.
Steedman, I. 1977. *Marx after Sraffa*. London: New Left Books.

Dollarization

Roberto Chang

Abstract

This article focuses on dollarization, a situation in which a foreign currency (often the US dollar) replaces a country's currency in performing one or more of the basic functions of money. The distinction between official dollarization and endogenous dollarization is discussed, as are the concepts of currency substitution and liability dollarization. Implications for monetary and exchange rate policy are emphasized.

Keywords

Aggregate demand; Capital asset pricing model; Currency substitution; Dollarization; Euro; Financial dollarization; Inflation; Lender of last resort; Liability dollarization; Monetary policy; Net worth effects; Portfolio balance; Search theory; Seigniorage; Stabilization; Transaction costs

JEL Classifications

F3

Dollarization is a situation in which a foreign currency (often the US dollar) replaces a country's currency in performing one or more of the basic functions of money.

Thus in Ortiz (1983) the term 'dollarization' refers to the widespread usage of US dollars for transaction purposes in Mexico. More recently, Ize and Levy-Yeyati (2003) use 'financial dollarization' for episodes in which domestic financial contracts are denominated in dollars or another foreign currency.

In some countries, dollarization has been the outcome of official government policy. Examples include Ecuador in 2000 and El Salvador in 2001, where the domestic currency was retired from circulation and the US dollar became the official currency. An immediate implication of such 'official dollarization' is that domestic prices of tradable goods are tied to world prices, so domestic inflation is closely related to US inflation. Hence official dollarization has been advocated for countries suffering from chronic, high, and volatile inflation.

On the other side of the ledger, official dollarization implies the surrender of independent monetary policy, leaving only fiscal policy available as a stabilization tool. In addition, the domestic government gives up seigniorage, or the revenue from money creation, which accrues to the US Federal Reserve. While both effects are widely regarded as costly for the domestic economy, their welfare

implications depend on details about the policymaking process and, in particular, on whether the monetary authorities can credibly commit to implement optimal policy (see Chang and Velasco 2002, for a discussion).

Finally, official dollarization implies that the domestic central bank is no longer available as a lender of last resort, which may be conducive to financial fragility and crises. Calvo (2005) argues, however, that last resort lending can be provided by alternative arrangements.

Impetus for official dollarization as a policy alternative was greatest at the turn of the millennium, as emerging economies had to cope with a sequence of financial and exchange rate crises while several European countries were abandoning their national currencies in favour of the newly created euro. Support for official dollarization appears to have subsided since, however.

More frequently, dollarization has emerged as a spontaneous response of domestic agents to inflation. The special case in which such a process has resulted in the dollar becoming a widespread medium of exchange is known as 'currency substitution'. Currency substitution has been the subject of a large literature, much of it focused on the determinants of the relative demand for domestic vis-à-vis foreign currencies and on implications for monetary management. Early research followed Girton and Roper (1981) in postulating ad hoc aggregate demand functions for domestic and foreign currency, in the portfolio balance tradition. Somewhat later, Calvo (1985) derived similar demand functions from an optimizing model in which domestic and foreign currencies entered the representative household's utility function. Those approaches emphasized the possibility that increasing substitutability between the domestic and the foreign currencies would lead to monetary and exchange rate instability. However, they did not identify the basic determinants of substitutability, which was buried in the specification of the postulated demand function for foreign currency or the properties of the representative agent's utility function. Hence the early studies were of little use in understanding how to cure the ills associated with dollarization, and, in particular, they failed to trace the consequences of common policies designed to deal directly with currency substitution, such as outright prohibitions on the holdings of foreign currency.

Subsequent studies have attempted to address these shortcomings by modelling more explicitly the fundamental frictions underlying currency substitution. Thus Guidotti and Rodriguez (1992) developed a cash-in-advance model of currency substitution on the assumption that using foreign currency entailed fixed transaction costs, while Chang (1994) studied the implications of a similar assumption in an overlapping generations setting. These models still left unexplained where the assumed transaction costs were coming from. Therefore, recent work on this area models currency substitution entirely from first principles, in the search theoretic tradition (see, for instance, Craig and Waller 2004).

Another focus of recent literature has been the increased use of the dollar as the currency of denomination of the debts of domestic residents in emerging economies, a problem that Calvo (2005) terms 'liability dollarization'. A substantial degree of liability dollarization places an economy in a vulnerable situation, since presumably many of the agents with dollar debts have assets denominated in domestic currency. Such a currency mismatch situation means that a depreciation of the domestic currency reduces the net worth of domestic agents. If, in turn, aggregate demand depends on net worth (as would be the case in the presence of financial imperfections), a currency depreciation may lead to a reduction in income and employment. In other words, liability dollarization may render depreciations contractionary, not expansionary as assumed by conventional analysis (Aghion et al. 2001; Cespedes et al. 2004). The combination of liability dollarization and net worth effects has been blamed for the severity of the income and output contractions in recent emerging markets crises.

At this point, no consensus exists as to the causes of liability and financial dollarization, although research on this question is rather active. Ize and Levy-Yeyati (2003), in particular, have examined the choice of currency denomination of assets and liabilities from a capital asset pricing model (CAPM) perspective, while Jeanne (2005)

models liability dollarization as the private sector response to the lack of credibility in monetary policy. Finally, several studies estimate how measures of financial dollarization depend empirically on other characteristics of an economy. For example, Arteta (2005) has found that the dollarization of bank deposits is empirically more frequent in countries with a higher degree of exchange rate flexibility.

See Also

- ▶ Currency Unions
- ▶ Money

Bibliography

Aghion, P., P. Bachetta, and A. Banerjee. 2001. Currency crises and monetary policy in an economy with credit constraints. *European Economic Review* 45: 1121–1150.

Arteta, C. 2005. Exchange rate regimes and financial dollarization: Does flexibility reduce currency mismatches in bank intermediation? *Topics in Macroeconomics* 5(1): 1226–1246.

Calvo, G.A. 1985. Currency substitution and the real exchange rate: The utility maximization approach. *Journal of International Money and Finance* 4: 175–188.

Calvo, G.A. 2005. Capital markets and the exchange rate with special reference to the dollarization debate in Latin America. In *Emerging capital markets in turmoil*, ed. G. Calvo. Cambridge, MA: MIT Press.

Cespedes, L., R. Chang, and A. Velasco. 2004. Balance sheets and exchange rate policy. *American Economic Review* 94: 1183–1193.

Chang, R. 1994. Endogenous currency substitution, inflationary finance, and welfare. *Journal of Money, Credit, and Banking* 26: 903–916.

Chang, R., and A. Velasco. 2002. Dollarization: Analytical issues. In *Dollarization*, ed. E. Levy-Yeyati and F. Sturzenegger. Cambridge, MA: MIT Press.

Craig, B., and C. Waller. 2004. Dollarization and currency exchange. *Journal of Monetary Economics* 51: 671–689.

Girton, L., and D. Roper. 1981. Theory and implications of currency substitution. *Journal of Money, Credit, and Banking* 13: 12–30.

Guidotti, P.E., and C.A. Rodriguez. 1992. Dollarization in Latin America: Gresham's law in reverse? *IMF Staff Papers* 39: 518–544.

Ize, A., and E. Levy-Yeyati. 2003. Financial dollarization. *Journal of International Economics* 59: 323–347.

Jeanne, O. 2005. Why do emerging economies borrow in foreign currency? In *Other people's money*, ed. B. Eichengreen and R. Hausmann. Chicago: University of Chicago Press.

Ortiz, G. 1983. Currency substitution in Mexico: The dollarization problem. *Journal of Money, Credit, and Banking* 15: 174–185.

Domar, Evsey David (1914–1997)

E. Cary Brown

Keywords

Domar, E. D.; Economic growth; Harrod–Domar growth model; Portfolio theory; Proportional income tax; Slavery; Technical change

JEL Classifications

B31

Domar (Domashevitsky) was born in 1914 in Lodz, Russia (now Poland), spent most of his early life in Harbin, Manchuria, and moved permanently to the United States in 1936. His undergraduate degree in economics (1939) was from the University of California (Los Angeles); his graduate work was at the Universities of Michigan (MA, Mathematical Statistics) and Harvard (Ph.D., 1947), where he studied with Alvin Hansen, the leading American Keynesian and most important single intellectual influence on Domar. Domar is best known for his leadership role, along with Roy Harrod, in the initiation of modern growth theory.

His first position was with the research staff of the Board of Governors of the Federal Reserve System, where he worked on fiscal problems from 1943 to 1946. His subsequent academic career took him briefly to the Carnegie Institute of Technology, the Cowles Foundation and the University of Chicago, the Johns Hopkins University in 1948 for ten years, and the Massachusetts Institute of Technology in 1958, from which he retired in 1984. An avid traveller, he held more than a dozen visiting professorships in universities at home and abroad.

While the claim to the earliest statement of the famous Harrod–Domar growth model was clearly Harrod's (1939), Domar arrived independently at a structurally similar model but from a different point of view (1946, 1947). By incorporating into static Keynesian analysis the capacity changes associated with investment, he found that steady-state capacity growth required investment to grow at a rate equal to the savings rate multiplied by the capital–output ratio. From this simple beginning, growth theory took off to become a major focus, one might almost say obsession, of the profession in the 1950s and 1960s. Domar also made important contributions to some of its conceptual and measurement problems, such as the proper treatment of depreciation (1953) and the measurement of technological change (1961), and he coined the term 'residual' for the fraction of expanding output unexplained by the contribution of factors of production.

In fiscal theory, his early investigation, with Richard Musgrave (1944a), of the effect of a proportional income tax, with and without loss offsets, on portfolio choice was very similar in style and approach to portfolio theory of a decade later. Given individual preferences, the portfolio decision was modelled as a choice between alternative portfolios weighing their expected net returns against their risks (expected losses). The unconventional conclusion was reached that, given risk aversion, the imposition of a proportional income tax with symmetrical treatment of gains and losses would induce individuals to adjust their portfolios towards riskier assets. The reminder that expected risks and yields are both reduced by an income tax was an important correction to a simplistic focus on yields alone.

As an applied theorist, Domar had the knack of getting important results with simple theory. At a time when deficit finance was harshly criticized for increasing the debt burden and tax rate, Domar showed (1944b) that in a growing economy even continuous deficit finance resulted in only limited debt–income ratios and tax rates. Second, he made a fertile historical hypothesis (1970) – that the economic basis for the introduction of serfdom (or slavery) was a low land-to-labour cost. Third, he ingeniously modified the administrative rules that guided the behaviour of collective farms (1966) or that determined the compensation of socialist managers (1974) to induce them towards more efficient price–output decisions.

Domar's work was informed by a rare combination of historical, empirical and theoretical breadth. His profound scholarship, in several languages, periods, and areas, often resurrected important findings of earlier writers previously overlooked.

D

Selected Works

1944a. (With R.A. Musgrave.) Proportional income taxation and risk-taking. *Quarterly Journal of Economics* 58: 388–422.

1944b. The burden of the debt and the national income. *American Economic Review* 34: 798–827. Reprinted in Domar (1957).

1946. Capital expansion, rate of growth, and employment. *Econometrica* 14: 137–47. Reprinted in Domar (1957).

1947. Expansion and employment. *American Economic Review* 37: 34–55. Reprinted in Domar (1957).

1948. The problem of capital accumulation. *American Economic Review* 38: 777–94. Reprinted in Domar (1957).

1952. Economic growth: An econometric approach. *American Economic Review, Papers and Proceedings* 42: 479–95. Reprinted in Domar (1957).

1953. Depreciation, replacement and growth. *Economic Journal* 63: 1–32. Reprinted in Domar (1957).

1957. *Essays in the theory of economic growth.* New York: Oxford University Press.

1961. On the measurement of technological change. *Economic Journal* 71: 709–29.

1966. The Soviet collective farm as a producer cooperative. *American Economic Review* 56: 734–57.

1970. The causes of slavery or serfdom: A hypothesis. *Journal of Economic History* 30: 18–32.

1974. On the optimal compensation of a socialist manager. *Quarterly Journal of Economics* 88: 1–18.

Bibliography

Harrod, R.F. 1939. An essay in dynamic theory. *Economic Journal* 49: 14–33, Errata, 377.

Harrod, R.F. 1948. *Towards a dynamic economics. Some recent developments of economic theory and their applications to policy.* London: Macmillan.

Domesday Book

F. W. Maitland

Domesday Book is the name which, at least since the 12th century, has been borne by the record of the great survey of England made by order of William the Conqueror. Apparently the decree for the survey was issued at a moot held at Gloucester at the midwinter of 1085–86, and the work was completed in the course of the following year. Royal commissioners (*legati*) were sent into each shire with a list of interrogatories, to which they were to obtain sworn answers from local juries. Their procedure seems to have been this – they held a great shire moot, at which every hundred or wapentake of the shire was represented by a jury, while every vill was represented by a deputation of villagers. From each hundred-jury they obtained a verdict about all the land in the hundred, the villagers being at hand to correct or supplement verdicts, while 'the whole shire' was also present, and from time to time appeal could be made to its testimony. The statement thus supplied was

> reduced into writing and duly transmitted to the king. It was afterwards methodised and abstracted, and fairly transcribed in the great volume of Domesday and deposited in the royal treasury at Winchester, amongst the other muniments of the realm. It still exists, fresh and perfect as when the scribe put pen to parchment, the oldest cadastre, or survey of a kingdom, now existing in the world (Palgrave, *History of Normandy and England*, vol. iii, p. 575).

Our best information about the form of the original verdicts is contained, not in Domesday Book itself, but in a document known as the 'Inquisitio Comitatus Cantabrigiensis'. This seems to be a copy made in the 12th century of the verdicts delivered by the juries which represented some of the hundreds of Cambridgeshire. The verdicts having been obtained, they were sent to the king's treasury, and a digest was made of them by the royal officers. This digest is Domesday Book. If we may draw a general inference from Cambridgeshire, the materials supplied by the commissioners were subjected to a process of rearrangement. A scheme that was wholly geographical gave way to one which was partly geographical, partly proprietary. Domesday book deals with each shire separately, but within the shire it collects, under the name of each 'tenant in chief', all the estates that he holds, no matter in what hundred they may be. For example, the Cambridgeshire verdicts showed that Count Alan had lands in many hundreds. In the original verdicts the entries relating to his estates were therefore scattered about; in Domesday Book they are all collected together. Domesday Book consists of two volumes, sometimes called 'Great Domesday' and 'Little Domesday'. The latter deals with Essex, Norfolk, and Suffolk; the former with so much of the rest of England as was surveyed. A document in the keeping of the cathedral chapter of Exeter, and known as 'the Exon Domesday', contains an account of a large part of the south-western shires, which is very closely connected with that given by what, for distinction's sake, is sometimes called 'the Exchequer Domesday'. Seemingly this Exon Domesday is independent of the Exchequer Record, and goes back by a different route to the original verdicts. The same may perhaps be said of the 'Inquisitio Eliensis', an account of the estates held by the church of Ely. This Ely inquest must not be confused with the Cambridgeshire inquest.

Domesday Book was printed and published in 1783 in two folio volumes. A third volume containing indexes was published in 1811, and this was followed in 1816 by a fourth volume containing the Exon Domesday, the Ely Inquest, and some other matters. Of late years useful facsimiles have been published by the Ordnance Survey Office of various parts of the great Exchequer Record, and can be obtained at moderate

prices. The important Cambridgeshire Inquisition was first published by N.E. Hamilton in 1876.

A large literature has gradually been collecting round Domesday Book. Among the older books Robert Kelham's *Domesday Book Illustrated* (1788) and the essays of Philip Carteret Webb deserve to be mentioned. Sir Henry Ellis, in his *General Introduction to Domesday Book* (1833), supplied valuable indexes, and summed up the older learning. In the fifth volume of E.A. Freeman's *Norman Conquest* good use has been made of all that bears on political history, on the history of great men, great churches, great events. James F. Morgan's *England under the Norman Occupation* (1858) is a good introduction to the study of Domesday, and the like may be said of W. de Gray Birch's *Domesday Book* (1887). A new epoch in the scientific exploration of the record is marked by the various works of R.W. Eyton dealing with Dorset, Somerset, Lincoln, and Stafford, especially by the key to Domesday Book. Two volumes of essays by various writers, called *Domesday Studies* (1888–91), contain two valuable papers by J.H. Round, besides other matters. In some county histories Domesday has been well used, but here it is possible to name only the books of general importance. F. Seebohm's *English Village Community* has done much to awaken a new and an economic interest in our oldest statistics.

Much remains to be done. The student who approaches Domesday from the economic side will at once see that he has before him a vast mass of detailed statistics which ought to tell him much about agriculture, prices, rents, and the like. At the same time he will feel that he is debarred from making use of these precious materials by the difficulty of discovering the meaning of the crabbed formulas which are repeated on page after page. The difficulty is a very real one. Domesday Book stands alone. It is so far removed in time from the documents which most nearly resemble it, the extents of manors which are found in monastic cartularies, that we have to explain it out of itself or not at all, for we shall look in vain for help elsewhere. Then again the terms that it employs as technical terms are, we may say, derived from two different languages which have only of late come into contact with each other. About half of them have been introduced by the Norman conquerors, while the other half are words which were in use in England under Edward the Confessor. Hence many puzzles; for example, what word did English juries say when French clerks wrote down *villanus*? Then again, the more our record is studied, the more plainly do we see that one main purpose governs both its form and its matter. King William is not collecting miscellaneous information in the spirit of a scientific inquirer. He is in quest of geld. Domesday Book is a geld book, a tax book. Geldability, actual or potential, this is its main theme. If then we are to understand its statistics, the first thing necessary is a theory of geld, of the manner in which the great tax has been and is assessed and collected. Towards the construction of such a theory not a little has been done by modern writers, especially by Eyton and Round, but until the work has been completed, speculations about rents and values seem doomed to failure. Everywhere, for example, the question meets us whether we are reading of real areal units of land or of units which are the results of a rude system of taxation, and a great deal of labour must yet be spent on the book before this question will have been adequately answered.

References

Brich, W. 1887. *Domesday book*. New York/London: E. & J.B. Young & Co/Society for Promoting Christian Knowledge.

Dove, P.E. (ed.) 1888–91. *Domesday studies.* London/New York: Longmans, Green & Co.

Ellis, H. 1833. *General introduction to Domesday book*. London.

Eyton, R.W. 1878. *A key to Domesday.* London.

Freeman, E.A. 1869–79. *The history of the Norman Conquest,* vol. 5. Oxford.

Kelham, R. 1788. *Domesday book illustrated*. London.

Maitland, F.W. 1897. *Domesday book and beyond*. Cambridge: Cambridge University Press.

Morgan, J.F. 1858. *England under the Norman occupation*. London.

Palgrave, F. 1851–64. *The history of Normandy and of England,* vols. 1 and 2. London: J.W. Parker; vols. 3 and 4, London: Macmillan.

Seebohm, F. 1883. *The English village community.* London: Longmans & Co.

Webb, P.C. 1756. *A short account of some particulars concerning Domes-day.* London.

Domestic Labour

S. Himmelweit

The term domestic labour entered economic vocabulary in the early 1970s as a result of feminist interest in criticizing and expanding economic categories to incorporate women's activities. Both mainstream and critical traditions in economics tried to grapple with the problem of how to account for the difference between men's and women's position on the labour market. One approach was to relate women's lesser training and skills in paid employment to competing demands made on a (married) woman's time by domestic commitments, with a tacit, though unexplained, acceptance that for women paid employment has to fit into time left over after the allocation of that needed for domestic labour, while for men it is the other way round. It is only by the addition of such an assumption that the analysis of domestic labour can be said to have had anything to say about *women.*

Neoclassical economists have seen domestic labour as one of three competing claims on people's time, the others being paid work and leisure. A household maximizes 'its' utility, which is a function of the consumption goods bought with income received from paid work by members of the household, the direct consumption of the products of time spent in domestic labour and a variety of ways of spending remaining leisure time (Becker 1965). Women have a comparative advantage to men in domestic labour over paid work and so one or other partner should specialize; either a woman should not take paid employment or her husband should do no housework. Even if there is no intrinsic difference between men and women initially, specialized human capital can be acquired in each type of labour, so it makes sense for a division of labour to take place and for at most only one member of a household to work both in the home and outside. This is taken to explain both why the majority of domestic labour is performed by women and also why women work shorter hours in paid employment than men, accumulate less market-oriented training and skills, and have broken employment histories.

Two criticisms can be mounted of this approach. The first is that the comparative advantage itself needs explanation. At an individual level it can be accounted for by the lower relative earnings of women. But the outcome of individual household choices cannot, without circularity, then in turn be used to explain women's inferior earnings by the lesser time spent in the labour market acquiring appropriate human capital. At best such an approach can account for the division between houseworkers and paid workers, and if combined with an assumption that sex is used as a screening mechanism by employers, a form of rational statistical discrimination, why one sex as a whole will be more likely to constitute the houseworkers and the other the employees. But to explain why sex presents itself as a variable by which to screen, and why it is the *female* sex that constitutes the homeworkers, recourse must be made to biological differences in aptitude, an acceptable fall-back to some, but not to those who wish to show the power of the neoclassical economic approach to explain everything, nor to the feminist movement whose claims that a woman's place was socially rather than *naturally* in the home had led to the initial interest in the question.

The second criticism poses more fundamental problems for this type of analysis. The concept of a household's 'utility function' is a very shaky one. Individualism, upon which neoclassical economics is based, takes individuals as the only actors and decision makers, rejecting thereby, for example, the marxist notion of class interests and forces. The idea of a household utility function cannot therefore be entertained unless either all members of the household have identical preferences concerning the allocation of resources and leisure time among themselves or some rule for aggregating diverse preferences is adopted. Quite apart from the difficulty of devising such a rule which satisfies fairly minimal criteria to ensure that household preferences represent some meaningful aggregate of those of its members

(Samuelson 1956; Arrow 1951), there is little evidence that households, rather than individuals, make decisions at all. Indeed, feminists would argue that such an approach obscures one of the key questions it was supposed to illuminate, differential power and thus an unequal division of labour within the household (Pahl 1980).

This problem can be overcome by assuming one member of the household is sufficiently powerful, well-endowed and altruistic that all other members of the household are or aspire to be 'his' beneficiaries (Becker 1974). Then the interest of all family members are serviced by the maximization of family income and the whole family can behave as one single decision-making unit. The assumptions of this model, which might seem appropriate only to an idealized picture of a Victorian patriarchal family, are necessary in order to avoid oligopolistic decision making, and even then care has to be taken to ensure that the paterfamilias is not driven into a corner solution, whereupon the unity of the family breaks down.

Marxist approaches criticize neoclassical analyses for failing to take account of the different social relations involved in wage and housework. The categories of marxist analysis are particularly appropriate to the analysis of unequal power relations, making marxism seem to some feminists more likely to offer a useful approach. The marxist notion of exploitation is based upon the characterization of *specific* forms of surplus extraction. The attempt to analyse domestic labour in these terms would therefore illuminate power relations within the household, without falling into the trap of conflating housework with paid labour, by recognizing its relations of production, not just its product, to be specific.

Accounts which characterized domestic labour as a separate mode of production came both from writers claiming to be orthodox marxists, extending rather than revising Marx's work, and from others who saw themselves more as using parts of Marx's mode of analysis to criticize and reformulate orthodox Marxism. Of the latter group Christine Delphy, for example, argued that there is a transhistorical family mode of production in which wives' labour power is exploited by their husbands which has coexisted with and outlasted the modes of production Marx described (Delphy 1970). Harrison, on the other hand, sees the domestic mode of production as a specific subordinate counterpart to a capitalist mode of production unchanged from that of traditional marxist analysis.

Other accounts rejected the characterization of domestic labour as a separate mode of production on the grounds that a mode of production must be capable of independent self-perpetuation, since the term was used for the characterization of whole societies. The notion of a social formation encompassing two or more modes of production articulated with each other, while appropriate to the analysis of transition between modes, was not appropriate to the continued mutually dependent symbiotic relationship which exists between housework and wage work for capital. The alternative was to extend the notion of the capitalist mode of production to include housework (Gardiner et al 1975). That extension was needed because the transformation of the wage into reproduced labour power is a process requiring labour and taking place under specific relations of production, and not the unproblematic natural process that Marx took it to be (O'Brien 1981).

The effect within marxist theory of characterizing housework as a separate mode of production is to make housewives a class, exploited through performing surplus labour above the amount needed to reproduce their own labour power. This surplus was appropriated, according to different versions, either by their husband directly or transferred through lowering the value of his labour power to the capitalist who employed him. But if housework was seen as part of the capitalist mode of production, a housewife's class position, like that of anyone else, would be determined by her access to the means of production, and for most women that would put them in the working class along with their husbands.

Another area of dispute was whether domestic labour should be seen as value and/or surplus value producing. Some argued that it did produce value, because it produced the commodity labour power (Dalla Costa 1973). In so far as the housewife worked longer hours than that needed to reproduce her own labour she also produced

surplus value. Against this it could be argued that the housewife by producing use-values needed to reproduce labour power did not thereby make labour power her product, any more than the baker, butcher or obstetrician did (Seccombe 1975). Labour power is an attribute of a living human being and is not, *pace* Marx, a commodity like any other in that it is not directly produced by labour at all. In that case the labour that a housewife expands is use-value but not value creating, and therefore *a fortiori* not surplus value creating.

The dispute as to whether domestic labour counted as productive labour turned upon the same issue, since within the capitalist mode of production labour is productive, according to the Marxist definition, if and only if it produces surplus value. Those who argued that domestic labour produced surplus value could therefore also claim that it was productive labour. But against this could be put Marx's own demonstration that productive labour must, to produce surplus value, take place between two exchanges: in the first labour power is bought for a wage, in the second the product is sold. Domestic labour requires neither exchange and therefore is technically outside the classification into productive and unproductive labour, which applies only to wage labour (Fee 1976).

The 'domestic labour debate' as it became known failed to answer the question to which it was addressed: what is the material basis of women's oppression? To do so, it would have had to do more than classify domestic labour using the existing categories of Marxist analysis. By using those developed for the study of wage labour for capital it fell into a similar trap to the neoclassical approach.

The neoclassical approach failed to recognize that the different social relations under which domestic labour went on rendered the use of the theory of utility maximization developed to model market decision-making inappropriate. The assumptions needed in order that the division of labour within the home could be set up as a soluble decision-making problem had to turn the gender-divided household into a homogeneous single decision-making unit. Divisions within the household disappeared and its individual members became indistinguishable by anything that could be remotely related to gender except by recourse to some form of biological reductionism. Circularity is a common problem with utility analysis and in this case the only way to avoid it was by appeal to supposed biological differences, the very suppositions which feminists had rejected as insufficient to explain the social construction of gender-divided work patterns. Marxism can escape the charge of circularity because its method is a historical one. Circularity thus becomes recast as the reproduction through time of the conditions which give rise to a gender-divided society. But ultimately marxism fell into the same trap. Although it did recognize that domestic labour and wage labour go on under different relations of production, it failed to give those different relations any constructive effect, seeing domestic labour as simply labour that did not have all the attributes of waged labour for capital. To have got further it would have been necessary to relate the analysis of domestic labour to the sex of those who performed it and to its fundamental characteristic of being labour involved in *reproduction* rather than just another form of production (Himmelweit 1984).

See Also

- ▶ Family
- ▶ Housework
- ▶ Labour Supply of Women

Bibliography

Arrow, K.J. 1951. *Social choice and individual values*, Cowles Commission Monograph No. 12. New York: John Wiley & Sons.

Becker, G. 1965. A theory of the allocation of time. *Economic Journal* 75: 493–517.

Becker, G. 1974. A theory of social interactions. *Journal of Political Economy* 82(6): 1063–1093.

Dalla Costa, M. 1973. Women and the subversion of the community. In *The power of women and the subversion of the community*, 2nd ed. Bristol: Falling Wall Press.

Delphy, C. 1970. The main enemy. *Partisans* (Paris), Nos. 54–55.

Fee, T. 1976. Domestic labour: an analysis of housework and its relation to the production process. *Review of Radical Political Economy* 8(1): 1–8.
Gardiner, J, Himmelweit, S., and M. Macintosh. 1975. Women's domestic labour. *Bulletin of the Conference of Socialist Economists* 4(2). Reprinted in *On the political economy of women,* CSE Pamphlet No. 2. London: Stage One, 1976.
Harrison, J. 1973. The political economy of housework. *Bulletin of the Conference of Socialist Economists* 3(1): 35–52.
Himmelweit, S. 1984. The real dualism of sex and class. *Review of Radical Political Economics* 16(1): 167–183.
O'Brien, M. 1981. *The politics of reproduction*. London: Routledge & Kegan Paul.
Pahl, J. 1980. Patterns of money management within marriage. *Journal of Social Policy* 9(3): 313–335.
Samuelson, P.A. 1956. Social indifference curves. *Quarterly Journal of Economics* 70(1): 1–22.
Seccombe, W. 1975. Domestic labour – reply to critics. *New Left Review* 94: 85–96.

Donisthorpe, Wordsworth (1847–1914)

Peter Newman

Wordsworth Donisthorpe was born on 24 March 1847 in Harrogate, graduated from Trinity College, Cambridge in 1869 and was called to the Bar at the Inner Temple in 1879. Thereafter he lived and practised law in London. What is apparently the last of his many books and pamphlets was published in 1913, the year before he died, his habitual enthusiasm as yet undimmed.

He is of interest to economists because of his first book, *Principles of Plutology* (reviewed in the *Saturday Review*, 9 September 1876, pp. 331–2). In it his vigorous and eccentric style, reminiscent of that of Fleeming Jenkin's last two papers in economics (Colvin and Ewing 1887, Vol. II, pp. 122–54), is already there in full spate. While showing traces of Cairnes and Jevons, the book is for the most part subjectively original and objectively mediocre. Chapter IX on 'The Laws of Value' is an interesting exception. His clear understanding of the importance of the Law of One Price ('[This] proposition is the fundamental one', p. 133) is refreshing and unusual for its time, as is his Wicksteedian insistence on the reservation price of the seller, so that 'sellers and buyers are not two classes, but one class' (p. 132).

In the same chapter he discusses substitutes and complements ('co-elements'). Although attempting no rigorous definitions he does lay down various 'laws' concerning them, i.e. propositions of comparative statics. Thus, the Third Law reads: 'Other things equal, a rise in the value of a co-element is followed by a fall in the values of its co-elements, and a fall by a rise, but not necessarily at the same rate' (p. 153), while the Fourth Law is: 'Other things equal, a rise in the value of any commodity is followed by a rise in the value of its substitutes, and a fall by a fall, but not necessarily at the same rate' (ibid.). He points out that, depending on the circumstances, two commodities may stand in both relations to each other, e.g. wool and cotton as inputs to cloth are often co-elements while as individual consumer goods they are usually substitutes.

Perhaps disappointed at the book's reception (I know of no economist's reference to it) Donisthorpe soon left economics for political philosophy and became a leading pamphleteer for anarchic Individualism, a libertarian movement that perhaps owed as much to fear of Henry George as to admiration for Herbert Spencer. It took the complacent view that the state should interfere with individual activities only when deemed necessary to protect the rights of private property, a view which aroused such derisive epithets as 'tomtits of Anarchy' from its opponents (see *Westminster Gazette*, 3 and 11 August 1894).

However, a streak of Yorkshire shrewdness and wit kept Donisthorpe from becoming quite as doctrinaire and saintly an Individualist as, say, Auberon Herbert. For example, the last chapter of his *Law in a Free State* (1895) contains an uproarious but penetrating account of the problems posed to the Individualist polity, by what we would now call externalities of various kinds, economic, political, social and moral.

Selected Works

1876. *Principles of plutology.* London: Williams & Norgate.
1880. *The claims of labour; or, serfdom, wagedom and freedom.* London: Tinsley.
1889. *Individualism: A system of politics.* London: Macmillan.
1891. The limits of liberty. In *A plea for liberty,* ed. T. Mackay. New York: D. Appleton.
1893. *Love and law: An essay on marriage.* London: W. Reeves.
1895a. *Law in a free state.* London: Macmillan.
1895b. *A system of measures of length, area, bulk, weight, value, force &c.* London: Spottiswoode.
1898. *Down the stream of civilization.* London: G. Newnes.
1913. *Uropa. A new philosophically-constructed language based on Latin roots.* Guildford: W. Stent & Sons.

References

Colvin, S., and J.A. Ewing, eds. 1887. *Papers, literary, scientific &c., by the late Fleeming Jenkin, F.R.S., LL. D.* With a memoir by Robert Louis Stevenson. 2 vols. London: Longmans, Green.

Dorfman, Joseph (1904–1991)

Henry W. Spiegel

Keywords
Dorfman, J.; History of economic thought; Veblen, T.

JEL Classifications
B31

Historian of American economic thought, Dorfman was born in Russia in 1904 and educated at Reed College and at Columbia University, where he earned a Ph.D. degree in 1935 and taught from 1931 until his retirement 40 years later. Dorfman was a student of Clarence Ayres at Reed, and of Wesley C. Mitchell and John Maurice Clark at Columbia. Mitchell in turn had been a student of Thorstein Veblen. These four economists, all with institutional leanings, stand out among the formative influences that affected Dorfman's early career. He made Veblen the subject of his doctoral dissertation, which was published under the title *Thorstein Veblen and His America* in 1934. This was at the time the only book-length appraisal of a modern economist that gave close attention not only to the subject's writings but also to biographical detail, the contemporary climate of opinion, and the general social and cultural setting of the work.

This type of holistic approach is characteristic also of Dorfman's monumental *The Economic Mind in American Civilization*, a five-volume work that he published from 1946 to 1959. It is dedicated 'To the pioneering spirit of Thorstein Veblen and the first-born of his intellectual heirs, Wesley C. Mitchell'. The work is a detailed history of American economic thought from colonial times to 1933, the first of its kind and not likely to be replaced for many years. It is based on extensive research and in many instances provides the first comprehensive account of a writer's life and work. Dorfman sees a break of emphasis in the history of American economic thought at the time of the Civil War: it was commerce before, and industry later. He notes with respect the achievements of the past, and is a critical but tactful chronicler of past foibles. He was a pioneer in exploring not only the printed page but also archival material made up of 'papers', 'letters', and similarly elusive sources of information, the first writer to do so on a large and systematic scale in the history of economic thought.

Selected Works

1934. *Thorstein Veblen and his America.* New York: Viking Press.
1935. (With R.G. Tugwell.) *William beach Lawrence: Apostle of Ricardo.* New York, reprinted from *Columbia University Quarterly,* September, 1935.

1940. *The economic philosophy of Thomas Jefferson.* New York: Academy of Political Science.
1946–59. *The economic mind in American Civilization.* 5 vols. New York: Viking Press.
1954. Introduction to Adams, H.C., *Relation of the state to industrial action, and economics and jurisprudence. Two Essays,* ed. J. Dorfman. New York: Columbia University Press.
1960. (With R.G. Tugwell.) *Early American policy: Six Columbian contributors.* New York: Columbia University Press.
1967–9. Introduction to W.C. Mitchell, *Types of economic theory: From mercantilism to institutionalism*, ed. J. Dorfman. New York: A.M. Kelley.

Dornbusch, Rudiger (1942–2002)

Kenneth Rogoff

Abstract

Rudiger Dornbusch was one of the leading researchers in international macroeconomics in the late 20th century. He introduced the influential concept of exchange rate 'overshooting' to explain the excessive volatility of exchange rates after the break-up of the Bretton Woods system of fixed exchange rates in the early 1970s. Along with Stanley Fischer and Paul Samuelson, he revived the Ricardian theory of international trade whereby trade was driven by differences in technology; their simple tractable framework became similarly influential in the study of international trade.

Keywords

Comparative advantage; Dornbusch, R.; Exchange rate dynamics; Exchange rates (floating vs. fixed); Fischer, S.; Heckscher-Ohlin framework; New open economy macroeconomics; Non-traded goods; Overshooting; Ricardian trade theory; Samuelson, P.; Sticky prices; Value at risk (VAR)

JEL Classifications

B31

Rudiger Dornbusch was born in Germany on 8 June 1942. He received his Licence es Sciences Politiques from the University of Geneva in 1966, and his Ph.D. in Economics from the University of Chicago in 1971. He was an assistant professor at the Department of Economics at the University of Rochester from 1972 to 1974, an associate professor at the Graduate School of Business at Chicago University from 1974 to 1975, and a member of the MIT Department of Economics from 1975 to 1978. He became a Professor of Economics at MIT in 1978. From 1984 until his death from cancer on 25 July 2002, he was Ford International Professor of Economics at MIT.

Dornbusch was, by any measure, one of the giants of late 20th century international macroeconomics. His celebrated *Journal of Political Economy* paper 'Expectations and exchange rate dynamics' (1976), which introduced the concept of exchange rate 'overshooting', became the workhorse of international macroeconomics over the ensuing two decades. His *American Economic Review* paper (with Stanley Fischer and Paul Samuelson) 'Comparative advantage, trade and payments in a Ricardian model with a continuum of goods' (1977) introduced a simple tractable framework that became similarly influential in the study of international trade.

This entry begins by reviewing Dornbusch's two most important scientific contributions, and goes on to give a brief sketch of his broader influence on the profession through students (he served as an advisor on over 125 doctoral dissertations), through his leading intermediate textbook *Macroeconomics* (written with Stanley Fischer), and through his role as an important voice in the public policy debate.

Exchange Rate Overshooting

Dornbusch's overshooting model of exchange rates (1976) captured the imagination of policymakers and academics alike during the

early years of floating exchange rates. The model attracted enormous attention because, after the break-up of the Bretton Woods system of fixed exchange rates in the early 1970s, exchange rates seemed far too volatile relative to the underlying fundamentals. Although subsequent empirical work has undermined the model's original bold claim to explain floating exchange rates (see Meese and Rogoff 1983), the model is still viewed as relevant, especially during episodes of major shifts in monetary policy. In fact, an informal survey conducted by Alan Deardorff of eight top economics departments found that, as late as 1990, Dornbusch's overshooting model was the only paper taught in every one of their graduate international finance courses.

The idea of overshooting is so simple and elegant that the small-country version can be illustrated with just a couple of equations (the analysis here draws on Rogoff 2002). The assumption of 'uncovered interest parity' relates the home nominal interest rate to the exogenous foreign nominal interest rate and the expected rate of depreciation of the exchange rate:

$$i_t = i_t^* + E(e_{t+1} - e_t) \tag{1}$$

where i_t is the level home nominal interest rate and e_t is the logarithm of the exchange rate (the home currency price of foreign currency), so that $E(e_{t+1} - e_t)$ is the expected rate of change in the exchange rate. The second key relationship is a money demand equation that relates the real balances to the nominal interest rate.

$$m_t - p_t = -\lambda i_t + \eta y_t \tag{2}$$

where y denotes the log of output, m is the nominal money supply and p is the price level. Higher interest rates lower the demand for real balances, and an increase in output raises it. Dornbusch posed the question of what would happen if there were a one-time permanent increase in the money supply, m. If prices were fully flexible, it would be possible to maintain equilibrium in the above two equations by having prices and exchange rates all rise permanently in proportion to the increase in the money supply. In this case, money would be neutral and have no real effects.

In reality, however, while asset markets (including the exchange rate) adjust very quickly, goods markets adjust more slowly partly due to temporary price rigidities. Therefore, in this set-up money is neutral only in the long run (in which the price level rises proportionately to the money supply). But with goods markets clearing only slowly, what is the impact of a money shock on exchange rates and interest rates? Assume that output, y, is also fixed. If domestic prices are constant, then a rise in the money supply implies a rise in real balances, $m - p$. But this means that the home nominal interest rate i must fall, so there is a corresponding rise in the demand for real balances. Then, however, the uncovered interest parity equation (Eq. (1) above) implies that e_t *must fall, or depreciate, relative to expectations of* e_{t+1}.

That is, after any initial movement of the exchange rate in response to an unexpected shock, the currency must subsequently be expected to appreciate. But recall that in the long run, even with sticky prices, money is still neutral, so the exchange rate has to depreciate by the same amount as the rise in the domestic price level, thus producing no real effect.

How is all this possible? The answer, Dornbusch deduced, is that the initial money shock must cause the exchange rate to depreciate by more in the short run than it does in the long run. It 'overshoots'. Therefore, Dornbusch's model offered a highly plausible explanation of why exchange rates seem to be so volatile relative to fundamentals. At one level of abstraction, of course, 'overshooting' is an application of Paul Samuelson's 'Le Chatelier's principle' theorem: when prices in some markets are inflexible in the short run, prices in others may overreact in the short run. But Dornbusch's model did much more than innovatively contrast the fast adjustment of asset markets with the slow adjustment of goods markets (an insight that any realistic short-run dynamic macroeconomic model should take into account). It offered a concrete and coherent analysis of an extremely important

practical phenomenon. Over the decades since Dornbusch's article appeared, the term 'overshooting' has become deeply woven into the popular economic lexicon.

Modern research has advanced considerably beyond the overshooting model, of course, and the Mundell-Fleming-Dornbusch model has largely been supplanted by 'new open economy macroeconomics' (see Obstfeld and Rogoff 1996). And the notion of looking at money shocks via a money demand equation has increasingly been supplanted by frameworks which view the overnight interest rate as the key instrument of monetary policy. Nevertheless, these newer frameworks typically include sticky prices – perhaps the most fundamental, and controversial, element of Dornbusch's model – and hence can all replicate a similar phenomenon to 'overshooting.'

Although Dornbusch's overshooting paper was his best-known work, with over 900 citations in refereed journals, he published numerous other very well-known articles, including his 1973 *American Economic Review* paper that was among the first to incorporate non-traded goods in a monetary model (see also his elegant 1974a contribution to the collection edited by Robert Aliber), his 1983 *Journal of Political Economy* paper that illustrated how changes in the real interest rate could affect exchange rates and current accounts, and his 1987 *American Economic Review* paper that demonstrated a link between market structure and the adjustment of relative prices to exchange rate movements. Without doubt, however, his other extremely influential paper was not in international finance but in trade.

Ricardian Model of Trade

Dornbusch's 1977 *American Economic Review* paper with MIT colleagues.

Stanley ischer and Paul Samuelson almost single-handedly revived the analysis of Ricardian trade; a 'Ricardian' model of trade is one with only one factor of production (usually taken to be labour). Trade is driven by differences in technology. The Ricardian model is contrasted with the Hecksher–Ohlin framework, where countries have identical technologies but different relative endowments of the factors of production (labour and capital, in the simplest canonical case). Prior to Dornbusch–Fischer–Samuelson (DFS), the Ricardian approach had been dormant for years, having been largely supplanted by the Hecksher–Ohlin framework. The Ricardian model had lost out not so much because of poor empirical results but because it had come be viewed as intractable for all but illustrative purposes. By introducing a continuum of goods (rather than a discrete number), DFS were able to analyse elegantly a broad range of comparative static questions that had previously seemed unapproachable. DFS showed, for example, how to mobilize the combination of comparative advantage and trade costs to endogenize the dividing line between 'traded' and 'non-traded' goods, and how to analyse the classic 'transfer' problem where one country owes debt to another. Although at first only a trickle of papers followed DFS, the power of their continuum specification has led to a recent explosion of related research. DFS have become the starting point for a number of applied papers (see, for example, Copeland and Taylor 1994). In addition, DFS form the basis for a broad range of empirical papers (see, for example, Eaton and Kortum 2002; Kehoe and Ruhl 2002; Kraay and Ventura 2002; Kei-Mu Yi 2003; Ghironi and Melitz 2005; see also Feenstra and Hanson 1996). As the empirical work following DFS deepens, it is fair to say that trade economists have increasing faith in the fundamental underpinnings of the model.

Broader Contributions

Aside from his path-breaking research, Dornbusch made important contributions to economics in a number of other dimensions. His intermediate undergraduate textbook with Stanley Fischer, *Macroeconomics*, written in the mid-1970s, became a worldwide best-seller. The book was really the first to integrate modern supply-side economics into the standard demand-driven framework of the day. As such, students were able to gain a far deeper

understanding of problems such as the effects of oil price shocks.

Dornbusch was enormously influential as a graduate teacher at MIT. At his regular early-morning international economics 'breakfasts', Dornbusch would dissect recent models and serve up provocative questions in a fast-paced freewheeling style; many students remember these unique seminars as their most influential experiences as Ph.D. students. Dornbusch served as thesis advisor to scores of economists (as noted earlier, more than 125 in all), including Jeffrey Frankel, Paul Krugman, Maurice Obstfeld and Kenneth Rogoff. His dynamic, Socratic lecturing style also attracted students from outside MIT to his advanced graduate classes, including the likes of Jeffrey Sachs and Lawrence Summers. Many Dornbusch students went on to become finance ministers and heads of central banks throughout the world.

Through clear and incisive policy analysis embodied in editorials, speeches, and private meetings, Dornbusch exercised an enormous influence on global macroeconomic policy. He was a frequent guest of leading government officials throughout the world, who greatly valued and respected his advice. Arguably, no other recent economist has had so great an impact on the global macroeconomic policy debate, especially in emerging markets such as Brazil, Korea and Mexico, but also in more advanced countries such as Italy and Germany. Notably, in his later writing he succeeded in drawing ever more concrete insights from contemporary academic research, displaying a magnificent ability to translate complex theoretical models into ideas of immediate practical relevance. For example, his 1994 Brookings paper (with Alejandro Werner) argued that Mexico's pegged exchange rate had become overvalued to an extent that was unsustainable. Dornbusch's comments on markets prior to the currency collapse at the end of 1994 were highly influential. He also advanced a number of innovative ideas for dealing with international debt problems. His policy analysis was notable in that he managed to adopt strong views while continuing to be perceived as an independent and objective thinker. Over the last ten years of his life, Dornbusch became especially well-known for his monthly 'Economic Perspectives' newsletter, which covered with panache a broad range of topical global economic problems. One innovative idea, first developed in the newsletter and then formally published in his 'Primer on Emerging Market Crises' (2002) was to apply 'value at risk' analysis to the balance sheet of a country. In his primer, he wrote:

> . . .the right answer to crisis avoidance is controlling risk. The appropriate conceptual framework is *value at risk (VAR)* – a model-driven estimate of the maximum risk for a particular balance sheet situation over a specified horizon. There are surely genuine issues with the specifics of VAR surrounding modelling as has been widely discussed with respect to bank risk models used for meeting BIS requirements. But just as surely there is no issue whatsoever in recognizing that this general approach is the right one. If authorities everywhere enforced a culture of risk-oriented evaluation of balance sheets, extreme situations such as those of Asia in 1997 would disappear or, at the least, become a rare species. (2002, pp. 743–54)

In this short space it has not been possible to do full justice to the range and breadth of Dornbusch's contributions. But I hope the reader has gained some perspective on why he will have a lasting influence.

See Also

- ▶ Comparative Advantage
- ▶ Exchange Rate Volatility
- ▶ Extremal Quantiles and Value-at-Risk
- ▶ Neo-Ricardian Economics

Selected Works

1973. Devaluation, money and nontraded goods. *American Economic Review* 63, 871–880.

1974a. Real and monetary aspects of the effects of exchange rate regime changes. In *National Monetary Policies and the International Financial System*, ed. R. Aliber. Chicago: University of Chicago Press.

1974b. Tariffs and nontraded goods. *Journal of International Economics* 4, 177–185.

1976. Expectations and exchange rate dynamics. *Journal of Political* Economy 84, 1161–1176.

1977. (With S. Fischer and P. Samuelson.) Comparative advantage, trade and payments in a Ricardian model with a continuum of goods. *American Economic Review* 67, 823–839.
1980. *Open economy macroeconomics*. New York: Basic Books.
1983. Real interest rates, home goods and optimal external borrowing. *Journal of Political Economy* 91, 141–153.
1987. Exchange rates and prices. *American Economic Review* 77, 93–106.
1990. (With S. Fischer.) *Macroeconomics*, 5th edition. New York: McGraw-Hill.
1994. (With A. Werner.) Mexico: Stabilization, reform and no growth. *Brookings Papers on Economic Activity*, 1994: 1, 253–315.
2002. A primer on emerging market crises. In *Preventing currency crises in emerging markets*, ed. S. Edwards and J.A. Frankel. Chicago: The University of Chicago Press.

Bibliography

Copeland, B., and M. Scott Taylor. 1994. North-South trade and the environment. *Quarterly Journal of Economics* 109: 755–787.
Eaton, J., and S. Kortum. 2002. Technology, geography, and trade. *Econometrica* 70: 1741–1779.
Feenstra, R., and G. Hanson. 1996. Globalization, outsourcing, and wage inequality. *American Economic Review* 86: 240–245.
Ghironi, F., and M. Melitz. 2005. *International trade and macroeconomic dynamics with heterogeneous firms*. Mimeo: Harvard University.
Kehoe, T.J., and K.J. Ruhl. 2002. *How important is the new goods margin in international trade?* Mimeo: University of Minnesota.
Kei-Mu, Yi. 2003. Can vertical specialization explain the growth of world trade? *Journal of Political Economy* 111: 52–102.
Kraay, A., and J. Ventura. 2002. Trade integration and risk sharing. *European Economic Review* 46: 1023–1048.
Meese, R., and K. Rogoff. 1983. Empirical exchange rate models of the seventies: Do they fit out of sample? *Journal of International Economics* 14: 3–24.
Obstfeld, M., and K. Rogoff. 1996. *Foundations of international macroeconomics*. Cambridge, MA: MIT Press.
Rogoff, K. 2002. Dornbusch's overshooting model after 25 years: IMF Mundell-Fleming Lecture. *International Monetary Fund Staff Papers* 49 (Special Issue), 1–35 (including remarks by Rudiger Dornbusch).

Double-Entry Bookkeeping

Basil S. Yamey

Abstract

Double-entry bookkeeping is a system for arranging and organizing accounting information. It requires that each transaction (or other change) recorded in the accounting system must be recorded twice, and for the same money amount, once in debit form and once in credit form. Because it is concerned with the organization of information rather than with the scope and detail of that information, the system of double-entry bookkeeping is highly adaptable. It neither generates nor requires any particular set of valuation rules or profit concepts, and it is compatible with different treatments for changes in the value of money.

Keywords

Accounting; Assets and liabilities; Balance-sheet equation; Double-entry bookkeeping; National accounting; Pacioli, L.; Sombart, W.; Spengler, O.; Transaction analysis; van Gezel, W.

JEL Classifications

M4

Firms of all kinds need, in different degrees, to maintain records of their transactions with other firms and persons, of the debts they owe or are owed, and of their assets. The records they keep for this purpose constitute their accounting records. Traditionally they have consisted of account-books of various kinds, but they can take the form also of magnetic tapes and so on. If the records are kept on a systematic basis, one can speak of an accounting system. From the accounting records one can prepare a variety of accounting statements in which the detailed accounting information is rearranged, regrouped and presented in summary form. The balance

sheet and the profit-and-loss (or income) account or statement are important examples of such accounting statements.

Double-entry bookkeeping is a system or method for the arrangement and classification of accounting information. It developed in Italy, possibly in the second half of the 13th century. A description of the system was first published in Venice in 1494 as one part of a famous compendium of mathematical and commercial information: Luca Pacioli's *Summa de Arithmetica Geometria Proportioni et Proportionalità.* Knowledge of the double-entry system spread gradually from Italy to the rest of Europe by way of commercial contacts, schools and published treatises. It is not possible to establish how widely the system was used by merchants and others, say, in the 18th century. But by the late 19th century it had become the standard system for accounting records. Today it is used by virtually all corporate enterprises and many other firms as well as non-profit-making organizations in the West and also elsewhere. It has also proved suitable to serve as a useful scaffolding for the construction of the national income and related accounts for countries or regions.

Double-entry bookkeeping is no more than a system for arranging and organizing accounting information. It does not itself define the scope and detail of that information. Thus, for example, the double-entry system does not require that all transactions with third parties should be recorded, although it is the convention now to record all of them. What is more important, it does not prescribe which occurrences or changes that do not involve external transactions should be recorded in the accounts. Thus it does not prescribe whether changes in the value of the firm's assets should be recorded, how they should be determined, or how they should be recorded. Double entry neither generates nor requires any particular set of valuation rules or profit concepts. Different valuation bases or conventions, and different treatments for changes in the value of money, are all compatible with the use of the double-entry system. The system itself is highly adaptable, since it is concerned with arrangement and organization rather than with scope and content. Its adaptability has made it possible for it to serve as the basis for arranging the records needed by the relatively small-scale merchants in the early modern period of economic expansion as well as for those of the largest corporate enterprises operating today. But this does not mean that asset values were recorded and profits calculated in the same way by 17th-century merchants as they are by today's corporate enterprises. In fact, 17th-century merchants used several alternative bases for recording changes in asset values. And some of these would not be used by companies today.

Moreover, although all the companies within the same jurisdiction are subject to the same laws and the same institutional constraints (for example, those imposed by the stock-market authorities and those reflecting professional accounting standards), there is still scope for considerable variation in the determination and statement of accounting profits and asset values. However, because of developments in legislation and in the other constraining forces operating on corporate enterprises, it is no longer the case that a company chairman in the United Kingdom would be able to say (as Arthur Chamberlain, chairman of Tube Investments said in 1935) that he 'would almost undertake to draw up two balance-sheets for the same company, both coming within an auditor's statutory certificate, in which practically the only recognizable items would be the name and the capital authorised and issued'.

Double entry requires that each transaction (or other event) recorded in the accounting system must be recorded twice, and for the same money amount, once in debit form and once in credit form. In double entry, as Pacioli expressed it, 'all the entries placed in the ledger must be double, that is if you make a creditor (entry) you must make a debtor (entry)'. The debit and credit entries are made in the ledger, on the basis of the information entered in preliminary records. The ledger, which may for convenience be subdivided into a series of specialized ledgers, consists of a number of ledger accounts, pertaining, for example, to particular debtors or creditors, particular assets or particular categories of expenditure. It is the convention that the debit entry is made on the

left-hand (debit) side of the appropriate ledger account, and the corresponding off-setting credit entry on the right-hand (credit) side of the other appropriate ledger account.

The duality of entries for each transaction (or other recorded event) ties together the ledger accounts into an interlocking system of recorded information. Moreover, as each transaction gives rise to two equal but opposite entries, the system of accounts (if properly kept) is always in balance or equilibrium. The total of debit entries must be equal to the total of credit entries. Similarly, the total of the balances on all ledger accounts that have debit balances must be equal to the total of the balances on all the remaining ledger accounts that have credit balances. (If debit balances are taken as positive amounts and credit balances as negative amounts, the algebraic sum of the balances on all ledger accounts is zero.) The equality of debits and credits is the basis for the trial balance. This is a list of the balances on all open (that is, unbalanced) accounts in the ledger, distinguishing between debit and credit balances. If the trial balance does not balance, there is some error in the ledger. Postlethwayt in his Dictionary (1751) wrote of the 'agreeable satisfaction' of getting a trial balance to balance, and said that the trial balance will 'shew you that this [double entry], of all methods, is the most excellent'. The fact that a trial balance does not balance is proof that the ledger does contain some error. The converse is, of course, not correct.

Roger North, son of the prominent Turkey merchant Sir Dudley North, wrote in 1714 as follows: 'The making true Drs. (debtors) and Crs. (creditors) is the greatest Difficulty of Accompting, and perpetually exerciseth the Judgment; being an Act of the Mind, intent upon the Nature and Truth of Things.' Writers of instructional books on bookkeeping and accounts through the centuries have devised various lists, rules or approaches to help the accountant decide which debit and credit entries he should make for the various categories of transaction.

An early rule, widely used, was as follows (taken from a verse, 'Rules to be Observed', in a book of 1553 by James Peele):

> To make the thinges Received, or the receiver,
> Debter to the thinges delivered, or to the deliverer.

This rule is obviously readily applicable to many categories of transaction. If cash is received from a debtor, debit the cash account; and credit the debtor's account. If office furniture is bought on credit, debit the furniture account; and credit the supplier's account. If the owner withdraws cash from the business, debit the capital (that is, owner's) account; and credit the cash account. But it is evidently a straining of the language to say, when an amount is written off the book value of, say, a ship, in order to reflect diminution of value due to wear and tear, that the profit-and-loss account, which is to be debited, 'receives' something that has been 'delivered' to it by the ship account. Teachers and textbook writers not surprisingly looked for a rule that is robust enough to cover comfortably all transactions and events to be recorded, and to indicate unambiguously in each case where the debit and where the credit are to be placed.

The most common rule or approach adopted today in transaction analysis in the double-entry system derives from the so-called balance-sheet equation. The earliest formulation of this approach can be traced to the work of a Dutchman, Willem van Gezel, published in 1681.

The basic balance-sheet equation is:

$$\text{Owner's Equity (or the firm's net worth)} = \text{Assets} - \text{Liabilities} = \text{Net Assets}; \text{ or } \text{Owner's Equity} + \text{Liabilities} = \text{Assets}.$$

The ledger contains accounts for the various assets and liabilities; and there are accounts in it for the capital contributed or withdrawn by the owner(s) and for any increases (decreases) in 'net worth' resulting from the activities of the firm. In the double-entry system, increases in assets are indicated by debits to an asset account – the extent to which assets are subdivided into separate ledger accounts is for each firm to decide. Conversely, decreases in assets are recorded as credits to asset accounts. The total of a firm's assets is represented by the total of claims on those assets; namely, its liabilities (that is, its debts to third parties) and its owner's equity. The total of these claims must be a credit amount that equals the debit amount

representing the assets. An increase (decrease) in a claim is therefore represented by a credit (debit) in a liability account or an equity account. (Again, the extent to which claims are subdivided into various ledger accounts is a matter for each firm to decide. As regards the equity element, it is common for a ledger to contain separate accounts for each major category of business expenditure and income, a trading account, perhaps subdivided by type of activity, for showing the gross profit, and a profit-and-loss account to bring together the results from all the subordinate ledger accounts.)

Transaction analysis follows readily. The payment of salaries reduces the asset 'cash' and reduces the owner's equity, since the payment, taken by itself, represents a loss to the firm: hence, debit the salaries (eventually, profit-and-loss) account; and credit the cash account. The depreciation of an asset likewise reduces an asset and reduces the equity: debit the depreciation account (eventually profit-and-loss) account; and credit the ship account.

As has already been emphasized, the double-entry system does not itself dictate whether or in what circumstances increases or decreases in assets are to be recognized in the accounts. Neither does the system dictate the basis on which, or the circumstances in which, assets are to be revalued in the accounts. Decisions of these kinds are accounting decisions; and whenever such decisions are taken, the double-entry system of recording will accommodate them in accordance with its own logical structure. It follows from this that, although the value of the owner's equity in the ledger will always be equal to the value of the firm's net assets (that is, assets minus liabilities to those outside the firm) as stated in the accounts, those two values depend on the bases on which the values of assets are stated in the accounts.

Subject to this crucial qualification, it follows from the equilibrium feature of the double-entry system that the change (increase or decrease) in the value of the net assets of a firm over a period will be reflected as entries in the various ledger accounts that represent the owner's equity. Those entries in the various equity accounts that relate to the firm's operations, when they are brought together in the profit-and-loss account, yield a balance that is equal to the change in the value of the net assets over the period. It is the profit (loss) for the period. This profit is equal to the change in the value of the net assets over the period (allowance being made for any contributions or withdrawals of assets by the owner). It may be noted that the same profit figure would be established if one took the difference between the totals of two inventories of the firm's net assets taken, respectively, at the beginning and at the end of the period, provided that the same valuations were used and the same allowance made for the owner's contributions and withdrawals. The method of profit calculation by means of successive inventories of assets and liabilities was widely used in the past. The surviving 16th-century records of the large-scale commercial, financial and mining enterprise of the Fugger family of Augsburg provide examples of this procedure.

The equality – Profit (Loss) = Change in Net Assets – evidently holds only if all the changes recorded in asset and liability accounts (other than the owner's contributions or withdrawals) are also recorded in equity accounts that, in turn, are closed into the profit-and-loss account. In contemporary corporate financial accounting it is permissible to allow the counter-entries representing certain changes in asset values, depending upon the circumstances, to bypass the profit-and-loss account (for example, by recording these changes as debits or credits to one or other reserve account). This practice breaks the nexus between changes in net asset values and profits. It does, however, allow more 'realistic' values to be used in asset accounts where, otherwise, their use might produce 'distortions' in the profit figures that could mislead users such as investors and investment advisers. Both 'realistic' and 'distortions' are words that give rise to much debate in accounting circles. The double-entry recording system can accommodate the practice of bypassing the profit-and-loss account as comfortably as it can the alternative. The system itself imposes no discipline or constraint upon accountant or management – except the constraint that

for each transaction or change recorded in the firm's accounting system, equal but offsetting debit and credit entries have to be made in accounts in the ledger.

The German economic historian, Werner Sombart, claimed that 'capitalism without double-entry bookkeeping is simply inconceivable', and that double-entry was one of the most significant inventions or creations of the human spirit. In similar vein, Oswald Spengler asserted that the creator of double-entry bookkeeping could take his place worthily beside his contemporaries Columbus and Copernicus. These scholars evidently attributed to the double-entry system a role that goes well beyond what one might think appropriate to ascribe to a system of organizing and arranging accounting data. In a nutshell, Sombart argued that, historically, the double-entry system opened up possibilities and provided stimuli that enabled capitalism to develop fully. It clarified the acquisitive ends of commerce and provided the rational basis on which this acquisition could be carried on. It provided the basis for the continued rational pursuit of profits, and virtually compelled its users to pursue the acquisition of wealth. It also enabled the firm or enterprise to be separated from its owners, thus facilitating the development of corporate enterprises.

These views are in their details either untenable or grossly exaggerated. To note only a few points: the profits of an enterprise and its capital employed can be calculated without double-entry bookkeeping; joint-stock companies, such as the Dutch East India Company, have existed and flourished without double-entry bookkeeping; 16th- and 17th-century merchants, like the Fugger, who did not use the system do not seem to have been any less acquisitive, rational and successful than those who did use the system; and the adoption of the double-entry system could not have changed, or even have reinforced, the temperament, commercial acumen, motivation or goals of those who adopted it for organizing their accounting records.

To reject grandiose claims made for double-entry bookkeeping is not to deny the more workaday usefulness of the system. A method or system for recording and classifying accounting data that has been used increasingly over a period of six centuries must indeed have substantial practical merit. Double entry is a useful and versatile method for organizing accounting data, its value increasing with the volume and complexity of the data to be organized. In turn, the efficient organization of data helps management at various levels in many ways, more notably in large organizations. But its contribution to efficiency does not proceed along the lines emphasized by Sombart.

See Also

- ▶ Accounting and Economics
- ▶ Assets and Liabilities
- ▶ Sombart, Werner (1863–1941)

Bibliography

Yamey, B.S. 1964. Accounting and the rise of capitalism. *Journal of Accounting Research* 2: 117–136. for a discussion of Sombart's views on double-entry bookkeeping and capitalism.

Douglas, Clifford Hugh (1879–1952)

David Clark

Major Douglas, the founder of the Social Credit movement, was born in Stockport, Cheshire, in 1879. After a period at Pembroke College, Cambridge, he trained as an engineer and then served with the Royal Flying Corps. He died at Dundee, Scotland, in 1952.

Major Douglas is best known for his A + B theorem, which his followers used to impress laypersons and exasperate academic economists. It was based on the claim that all productive organizations make two kinds of payments: Group A payments, made up of wages, salaries and dividends; and Group B payments, made up of all other payments to banks and suppliers of materials. In his own words (Douglas 1924):

> Since all payments go into prices, the rate of flow of prices cannot be less than A plus B. Since A will not purchase A plus B, a proportion of the product at least equivalent to B must be distributed by a form of purchasing power which is not comprised in the description grouped under A.

By first reducing prices below cost to the individual consumer and then making up this difference between price and cost by a Treasury issue to the producer, Douglas argued that such an issue of 'Social Credit' would enable underconsumption to be eliminated without inflation. The anti-socialist Douglas appeared oblivious to the fact that his scheme would have required an army of inspectors to fix and supervise the huge number of individual price reductions involved.

Social Credit ideas had the largest following in the Dominion economies of Canada, Australia and New Zealand. The province of Alberta had a Social Credit government between 1935 and 1971 and British Columbia one from 1952 to 1972. In other countries, his followers ranged from the 'Red Dean' (the Very Reverend Hewlett Johnson, Dean of Canterbury) through to the neo-fascist author Ezra Pound.

See Also

- ▶ Gesell, Silvio (1862–1930)
- ▶ Monetary Cranks

Selected Works

1920. *Economic democracy.* London: Stanley Nott.
1921. *Credit power and democracy.* London: Cecil Palmer.
1924. *Social credit.* London: Eyre & Spottiswoode.
1931. *Warning democracy.* London: Stanley Nott.

References

Dobb, M.H. 1936. Social credit discredited: Being an examination in terms of political economy of the much advertized nostrums of Major Douglas, which are subject to a devastating analysis and found wanting as a solution to the troubles of our time. London, no publisher given.
Mairet, P. (ed.). 1934. *The Douglas manual*. London: Stanley Nott.
McConnell, W.K. 1932. *The Douglas credit scheme: A simple explanation and criticism*. Sydney: Angus & Robertson.

Douglas, Paul Howard (1892–1976)

Colin G. Clark

Keywords

Agricultural economics; Climacteric (of 1896–1914); Cobb, C. W.; Cobb–Douglas functions; Douglas, P. H.; Phelps Brown, H.; Real wage growth

JEL Classifications

B31

Born in 1892 in Salem, Massachusetts, Paul Douglas attended Bowdoin College in Maine (BA, 1913) and Columbia University (Ph.D., 1921). After holding a number of teaching posts between 1916 and 1920, he joined the faculty of the University of Chicago where he remained (apart from service in the Second World War) until 1948, when he became a United States Senator from Illinois. After his retirement from the Senate in 1966, he taught at the New School for Social Research for two years (1967–9).

Paul Douglas first became well known for his massive theoretical and factual studies (for example, 1930) of all the available information on wages in the United States from 1890. This work required laborious following up of old, obscure records, and repairing gaps in the available knowledge, such as domestic service wages. Douglas also collected information on prices so as to make an estimate of the movement of real wages.

In Britain there was almost complete cessation of the growth of real wages between 1896 and 1914. Understandably, it was a period of growing social tension. Sir Henry Phelps Brown called it

the 'climacteric'. We still do not really understand its cause; there was some sociological evidence about the deterioration of the quality of businessmen. D.H. Robertson found at least a partial explanation in economic causes, namely, that, of the two leading British export industries, cotton was produced under constant returns and coal under diminishing returns.

This problem remains of primary interest to economic historians, and naturally they enquire whether there is any evidence of a similar 'climacteric' in other countries. In Germany there was a slowing down of the rate of rise in real wages, but not very marked. Douglas's American data likewise do not show such a 'climacteric'. Recent research, however, has thrown some doubt not on Douglas's wage data, but on his price data; and perhaps there was some slowing down of the rate of growth of real wages.

Douglas became famous to the whole economic world through the 'Cobb–Douglas function' (for example, 1934). Working in conjunction with Charles W. Cobb, a mathematician from Amherst College, and using Massachusetts State annual factory returns, Douglas in 1928 established the following relation: Let product be *P,* labour input *L,* capital input *C,* and *k* a constant. Then $P = kL^aC^b$. (The same formula, with land in place of capital, had already been used by Wicksell – for example, 1900 – but he gave it neither theoretical nor empirical development.)

We may, if we wish, constrain *a* and *b* to add up to 1; but we get much the same results unconstrained. If *a* and *b* add up to more than 1 this is an indication of economies of scale (increasing returns) – a uniform increase in the quantities of inputs giving a more than proportionate increase in product.

Annual data, which many economists have been using, give results mainly dependent on fluctuations in the short-period business cycle – which is not what we want at all. It is only when we have data for such a long period as to make it possible to average out the business cycle that we can draw conclusions about productivity. This has been done by Solow in the United States, Aukrust in Norway, and Niitamo in Finland. In each case it was found, in the long run, that the product was rising much more rapidly than expected from inputs and their exponents. This difference is generally held to be due to technical advance, though some look for economies of scale. Some difficult but promising work by Denison further analyses the labour input by numerous categories, male and female, adult and juvenile, and various levels of education. These methods reduce the unknown factor – but it does not disappear.

Differentiating the Cobb–Douglas formula to obtain marginal productivities, then aggregate earnings of the factors should be proportional to a:b – assuming that each factor is remunerated according to its marginal productivity. When he first made this calculation (so he told me), Douglas fully expected the aggregate income of labour to be below that indicated by its marginal productivity. He was surprised, however, to find that it was almost exactly what was to be expected – about 75 per cent of the product.

The Cobb–Douglas formula has had abundant application in agricultural economics, especially for cross-section studies, where each farm may be considered an independent piece of evidence. Land is introduced as a factor, and also data for other inputs – fertilizers, insecticides, and so on – even (in one study in Sweden) the age of the farmer – a negative factor.

Douglas was very much a political economist. Organized labour in the United States did not attempt to form a political party of its own as in Britain, but instead played the two existing parties off against each other in demanding concessions. But in the 1920s this was not fully agreed. The other element in the population with a grievance against the current state of affairs was the farmers, and an attempt was made to form a Farmer–Labour political party. Douglas took an active part in these negotiations, and was national treasurer of the organization. But with the Roosevelt reforms of the 1930s the prospects of a Farmer–Labour party died away.

Chicago had acquired a worldwide reputation for corruption and crime; and the ruling Democratic Party considered that its 'image' would be improved by an upright professor of economics on the city council. Douglas assured me that some

improvement had taken place, though less than was hoped for. Later, the despotic Mayor Daley achieved a real reduction in crime. But once I asked Douglas whether, if I wished to set up a milk distribution business in Chicago, he could guarantee my safety. He replied that, 'regrettably', he could not.

Douglas was a Quaker, and in the First World War applied for exemption from military service on religious grounds. But in the Second World War he felt very differently. In spite of his age, he obtained a commission in the marines through President Roosevelt's personal intervention, and took part in the bloody landing on Iwojima. He sustained an injury to his hand which was with him for the rest of his life.

From city councillor he advanced to become Senator for Illinois. On the very day that he arrived in Washington he found a vanload of furniture which had been offered to him as a gift. He sent it back. This episode prompted him to write a little book, *Ethics in Government* (1952). He saw no harm in the small presents customarily exchanged among businessmen and politicians – calendars, cigars, and so on – but instructed his staff to return any present valued at over four dollars.

See Also

► Cobb–Douglas Functions

Selected Works

1928. (With C.W. Cobb.) A theory of production. *American Economic Review* 18(Suppl): 139–165.
1930. *Real wages in the United States, 1890–1926.* Boston/New York: Houghton Mifflin Company.
1934. *The theory of wages.* New York: Macmillan.
1936. *Social security in the United States: An analysis and appraisal of the federal social security act.* New York/London: Whittlesey House/McGraw-Hill.
1939a. The effect of wage increases upon employment. *American Economic Review* 29(Suppl): 138–157.
1939b. (With H.G. Lewis.) Some problems in the measurement of income elasticities. *Econometrica* 7: 208–220.
1939c. (With M. Bronfenbrenner.) Cross-section studies in the Cobb–Douglas function. *Journal of Political Economy* 47: 761–785.
1947. (With E.H. Schoenberg.) Studies in the supply curve of labour; the relation in 1929 between average earnings in American cities and the proportions seeking employment. *Journal of Political Economy* 45: 45–79.
1948. Are there laws of production? *American Economic Review* 38: 1–41.
1952. *Ethics in government. The Godkin lectures at Harvard University, 1951*. Cambridge, MA: Harvard University Press.
1972. *In the fullness of time: The memoirs of Paul H. Douglas.* New York: Harcourt Brace Jovanovich.
1976. The Cobb–Douglas production function once again: Its history, its testing and some new empirical values. *Journal of Political Economy* 84: 903–915.

Bibliography

Wicksell, K. 1900. Marginal productivity as the basis of distribution in economics. *Ekonomisk Tidskrift.* English trans. in *K. Wicksell*: *Selected papers in economic theory*, ed. E. Lindahl. London: Allen & Unwin, 1958.

Du Pont de Nemours, Pierre Samuel (1739–1817)

Peter Groenewegen

Keywords

Advisers; Assignats; Du Pont de Nemours, P. S.; Excise taxes; Mathematical economics; Mercier de la Rivière, P.-P.; Mirabeau, V. R., Marquis de; Physiocracy; Quesnay, F.; Silver; Turgot, A. R. J

JEL Classifications
B31

Economic writer and editor. Born in Paris, he trained for various occupations including medicine and watch making. A pamphlet on taxation (1763) brought him in contact with Mirabeau and Quesnay, under whose guidance he wrote a work on the grain trade (1764). He also befriended Turgot, with whom he diligently corresponded until Turgot's death. From 1766 to late 1768 he edited the *Journal de l'Agriculture* in the Physiocratic cause, then the *Ephémérides* until 1772. During this period he also published Quesnay's economics under the title *Physiocratie* (Du Pont 1767) and summarized Mercier (1767), adding material on the history of the new science (Du Pont 1768). From the early 1770s he developed a career as economic adviser through correspondence with the King of Sweden and the Margrave of Baden; the correspondence with the latter was subsequently published (Knies 1892). In 1774 he was appointed tutor to the Polish royal family. On becoming *contrôleur-général,* Turgot required his friend's assistance and Du Pont was back in Paris by early 1775. Financial compensation for loss of his royal tutorship enabled him to purchase landed property near Nemours. Turgot's dismissal from office in 1776 did not end Du Pont's career in giving official economic advice; a highlight of which is his influence on the 1786 Anglo–French Commercial Treaty. Du Pont was politically active in the French Revolution, serving from 1789 as Deputy for Nemours in the National Assembly and becoming its President during 1790; in 1794 to 1797 he was imprisoned for short periods. He migrated to the United States in 1799 but returned to Paris in 1802. From 1803 to 1810 he served in the Paris Chamber of Commerce, and in addition edited Turgot's works (Du Pont 1808–11). In 1815 he returned to the United States and settled in Delaware, the town where his son Irenée had started the gunpowder factory from which the Du Pont chemical conglomerate developed, and where he died in 1817. Du Pont is now mainly remembered as a major propagator of Physiocracy, an early historian of economics, a pioneer in the use of diagrams in economic argument and, most importantly, as the editor of Quesnay and Turgot, whose works he helped to preserve. An assessment of his work as economist needs to take all facets of his career into account, as the one full-length attempt at this (McLain 1977) has in fact done.

Virtually all Du Pont's economic work is characterized by dogmatic adherence to the Physiocracy developed by Quesnay and codified by Mercier de la Rivière. Turgot criticized this 'servitude to the ideas of the master' as totally inappropriate in matters of science (Schelle 1913–23, vol. 2, p. 677). Despite such criticism Du Pont allowed his dogmatism to colour excursions into the history of economics (Du Pont 1769) and, more importantly, his preparation of Turgot's works for the press (see Groenewegen 1977), particularly his editions of the *Reflections* (Turgot 1766). Two examples of his more novel contributions to economics can be given. One is his use of diagrams in explaining economic policy, which Theocharis (1961, p. 60) described as the first use of a diagram by a professional economist for 'illustrating an economic argument set out in essentially dynamic time', thereby making Du Pont (1774) 'the earliest French contribution of importance in mathematical economics'. The problem analysed is the price effects of an excise reduction, the benefits of which are argued to accrue ultimately to the landowning class. The excise reduction's initial income effect on manufacturers and merchants allows them either to reduce their own prices or to pay higher prices for raw materials. By assuming this increased competition for raw materials to raise their price in each period by three-fourths of the increase in the preceding period, Du Pont shows how a new equilibrium price will be reached which transfers the benefits from excise reduction to the rural sector. His proof relies on the properties of diminishing geometrical progressions which also formed the basis for much of the analysis of the *Tableau économique.* Du Pont's analysis of the inflationary consequences from issuing assignats is a second example. Although much of this is similar to Turgot's (1749) analysis, some of it is of interest in explaining Smith's version of the specie mechanism to which Du Pont (1790, p. 28) explicitly refers. Issuing paper money by assignats makes silver

superfluous as a circulating medium; this drives the metal out of the country because its only other use is to be sold abroad (Du Pont 1790, p. 42), a specie mechanism like Smith's (1776, pp. 293–4) that is independent of relative price movements. Both examples of his more original economics relate to matters of economic policy and add force to the claim by McLain (1977, p. 255) that Du Pont represents 'the first important case of a professional economist turned government policy-maker, a tradition in which he would be followed by [many] others . . .'.

Selected Works

1763. *Réflexions sur l'écrit intitulé: Richesses de l'état*. Paris.
1764. *De l'exportation et de l'importation des grains*. Soissons and Paris.
1767. *Physiocratie, ou Constitution Naturelle du Gouvernement le plus avantageux au genre humain*. Leyden and Paris.
1768. De l'origine et des progrès d'une science nouvelle. In *Physiocrates,* ed. E. Daire. Paris, 1846.
1769. Notice abrégée des différents écrits modernes qui ont concours en France à former la science de l'économie politique. In *Oeuvres Oeconomiques et Philosophiques de François Quesnay,* ed. A. Oncken. Frankfurt/Paris, 1888.
1774. *On economic curves,* ed. H.W. Spiegel. Baltimore: Johns Hopkins. Reprints of Economic Tracts, 1955.
1790. *The dangers of inflation.* Trans. E.E. Lincoln. Boston: Kress Library Publications, 1950.
1808–11. *Oeuvres de Turgot.* 9 vols. Paris.

Bibliography

Groenewegen, P.D. 1977. *The economics of A.R.J. Turgot*. The Hague: Nijhoff.
Knies, K. 1892. *Carl Friedrichs von Baden Briefticher Verkehr mit Mirabeau und Du Pont*. Heidelberg: Carl Winter.
McLain, J.J. 1977. *The economic writings of Du Pont de Nemours*. Newark/London: University of Delaware Press.
Mercier de la Rivière, P.P. 1767. *L'Ordre Naturel et Essentiel des Sociétiés politiques.* Paris.
Schelle, G. 1913–23. *Oeuvres de Turgot et documents le concernant.* 5 vols. Paris: F Alcan.
Smith, A. 1776. In *An inquiry into the nature and causes of the wealth of nations*, ed. R.H. Campbell and A.S. Skinner. Oxford: Clarendon. 1976.
Theocharis, R.D. 1961. *Early developments in mathematical economics*. London: Macmillan.
Turgot, A.R.J. 1749. Letter on paper money. In Groenewegen (1977).
Turgot, A.R.J. 1766. Reflections on the production and distribution of wealth. In Groenewegen (1977).

Dual Economies

David Vines and Andrew Zeitlin

Abstract

Dual economies have asymmetric sectors, the interaction between which influences the path of development. These are typically a rural, traditional, or agricultural sector on one hand, and an urban, modern, or industrial sector on the other. The relevant asymmetries are not merely technological but also include institutional, behavioural, and informational aspects. Modern treatments have grown out of the work of W. Arthur Lewis, whose model was based on the existence of surplus labour in agriculture. Subsequent authors have considered the implications of alternative assumptions for the development of a dual economy.

Keywords

Agriculture and economic development; Asymmetric information; Capital accumulation; Credit markets in developing countries; Development economics; Disguised unemployment; Dual economies; Engel's law; Harris–Todaro hypothesis; Hecksher–Ohlin trade theory; Immigration; Kuznets curve; Labour surplus economies; Lewis, W. A.; Malthus's theory of population; Neoclassical growth theory; Outsourcing; Pure surplus labour; Rural–urban migration; Scissors problem; Subsistence; Terms of trade; Urban unemployment; Wage differentials

JEL Classifications
O3

Dual economies have asymmetric sectors, the interaction between which influences the path of development. W. Arthur Lewis introduced this idea in his paper, 'Economic Development with Unlimited Supplies of Labour' (Lewis 1954), which earned him the Nobel Prize for Economics in 1979. That paper contains two theoretical models, both designed to explain the intrinsic problems of underdevelopment. When the prize was awarded, Ronald Findlay wrote that 'a large part of ... development economics ... can be seen as an extended commentary on the meaning and ramifications [of this article]' (Findlay 1980, p. 64). Here we focus primarily on the first of Lewis's two models of dualism – that of a single underdeveloped economy. We describe that model, trace the evolution of the ideas which grew from it, and discuss the continuing importance of these ideas in the study of economic development.

Long before Lewis wrote his article, there had been much thinking about 'dual' economies, conceived of as economies with both an industrial sector and an agricultural sector. Adam Smith and David Ricardo both focused on the interaction between these sectors during the Industrial Revolution; for Ricardo the outlook for industrial growth was 'dismal' because of diminishing returns in agriculture (see Hicks 1965; Pasinetti 1974). In the early 20th century, there was an extended discussion in the Soviet Union of the 'scissors problem', concerning the determination of the terms of trade between these two sectors. Evgeny Preobrazhensky (1924) argued that a decrease in the relative price of agricultural goods could be used to stimulate industrial investment; others replied that sufficient agricultural goods would not be available at lower relative prices and that these goods would need to be seized by force, something which the collectivization of agriculture made possible (see Sah and Stiglitz 1984). And during the Great Leap Forward in China in the 1950s, Chairman Mao attempted to confiscate an increasing quantity of primary goods from the Chinese countryside in order to facilitate the development of urban manufacturing. These policies led to famine and to the deaths of approximately 30 million people. Thus both theorists and policymakers have long recognized that, in an economy with two very different sectors, growth prospects hinge on how these sectors interact.

In his Nobel Prize autobiography, Lewis (1979) writes that his interest was in the 'fundamental forces determining the rate of economic growth'. But he was not satisfied with the neoclassical model of growth that was emerging at the time (Solow 1956; Swan 1956), out of the work of Roy F. Harrod (1939) and Evsey D. Domar (1945). That neoclassical framework aimed to provide a *general* theory of growth. But to Lewis it seemed inadequate because it did not deal with interactions between the industrial and the agricultural sectors: in Lewis's words, this model contained no discussion 'of what determines the relative price of steel and coffee [namely, of industrial goods and agricultural goods]. The approach through marginal utility made no sense to me. And the Heckscher–Ohlin framework could not be used, since that assumes that trading partners have the same production functions, whereas coffee cannot be grown in most of the steel-producing countries.' Furthermore, the neoclassical theory seemed inadequate to him for historical reasons: '[a]pparently, during the first fifty years of the industrial revolution, real wages in Britain remained more or less constant while profits and savings soared. This could [also] not be squared with the neoclassical framework, in which a rise in investment should raise wages and depress the rate of return on capital' (Lewis 1979).

Then, Lewis continues:

> One day in August, 1952, walking down the road in Bangkok, it came to me suddenly that both problems have the same solution. Throw away the neoclassical assumption that the quantity of labour is fixed. An 'unlimited supply of labour' will keep wages down, producing cheap coffee in the first case and high profits in the second case. The result is a dual (national or world) economy, where one part is a reservoir of cheap labour for the other. The unlimited supply of labour derives ultimately from population pressure, so it is a phase in the demographic cycle. (Lewis 1979, p. 397)

D

This key insight launched Lewis on the journey that led to his famous article. Spelling out the implications of his insight led him to use the term 'dualism' to describe economies in which there are differences between industrial and agricultural sectors that cannot be adequately explained by differences in production technologies or in factor endowments, in the manner normally used by economists.

The Lewis Model

Lewis identified three such differences between industry and agriculture, which we term 'asymmetries' in this article (following Kanbur and McIntosh 1987).

First, there are technological differences between the sectors. Labour is used in each sector. In agriculture it is combined with land in production, whereas industrial goods are produced by combining labour with reproducible capital. Moreover, industrial goods can be consumed or invested, whereas agricultural goods can only be consumed.

Second, there are organizational differences between the sectors. The large, rural agricultural sector functions on traditional lines and is primarily based on subsistence; industrial production happens in a modern, market-oriented sector, located in towns and cities. There is 'an unlimited supply of labour, available at [a] subsistence wage' (Lewis 1954, p. 139) to both sectors. Lewis interprets the word 'subsistence' broadly. The level of the wage is determined in some way by conventions in the underdeveloped agricultural sector. Lewis is non-committal as to whether wages in this sector are set according to actual subsistence needs, or living standards, or workers' average product. The central idea is that workers are paid above their *marginal* product. Labour can be transferred from agricultural sector to the industrial sector by the migration of workers to towns and cities. The overall stock of labour in the economy is normally fixed in supply (though Lewis, like Ricardo, did sometimes allow for Malthusian features). Workers in the cities are paid not much more than the subsistence wage, although there may be a gap, as discussed below.

Third, and finally, there are differences in the behaviour of the actors in the two sectors. Capitalists in the industrial sector save all their profits, because they are ambitious. Workers save nothing, in either sector, because they are poor (Lewis describes them as not belonging to the 'the saving class' – 1954, p. 157). And landlords in agriculture are assumed to consume all their income, which comes to them to the extent that agricultural workers receive a wage below their average product.

The general story is this: the profits in the modern, capitalist, sector create a growing supply of savings. This finances the formation of an increasing stock of capital, which is used to employ more and more labour in the urban workforce.

We can explain the story in detail, using a simplified version of the model. To do this we make four sets of extreme assumptions. (*a*) There is 'pure' surplus labour, by which we mean that the marginal product of workers withdrawn from agriculture is *zero*. Wages initially consist only of agricultural goods, the level of wages per worker is exogenous, and workers are indifferent between working in industry and in agriculture at the same wage. (*b*) When one individual worker leaves agriculture and no longer needs to be rewarded there, then all the increase in the agricultural surplus (that is, all the increase in the total of food produced minus the total of wages paid to agricultural workers) accrues to landlords and is spent by them on consumption of industrial goods. (*c*) Industrial capitalists employ labour up to the point at which the marginal physical product of labour is equal to the cost of the wage, measured in industrial goods. (*d*) All industrial profits are saved and then invested in industrial production.

Given these assumptions, there are two steps to the argument. First, given assumptions (*a*), (*c*) and (*d*), the rate of growth depends negatively on the relative price of agricultural goods in terms of industrial goods. This is because an increase in food prices raises the cost of the wage per worker in terms of steel, causing less labour-intensive methods of production to be adopted, that is,

causing production to become more capital intensive. As a result of this, any given amount of savings, and the accumulation of capital that it causes, will 'go less far' in employing labour in industry, and, as a result, industrial output will grow less rapidly. Second, assumptions (*a*) and (*b*) determine the relative price of agricultural goods, in the following way: the accumulation of capital in industry increases the demand for industrial workers, who must be transferred from agriculture. The relative price of agricultural goods will need to be high enough to induce the workers' landlords to offer up those agricultural goods that they would have paid to the transferred workers but now receive as surplus, so as to receive industrial goods for consumption in exchange. Such trade enables workers to be paid in industry, where they now work. As Lewis (1954, p. 188) says, 'the capitalists need the peasants' food, and ... the demand for food is inelastic'.

Clearly, the relative prices of industrial and agricultural goods, and the growth rate of the economy, are *jointly* determined in this process – as Lewis's intuition had suggested to him. And it will clearly be true that the relative price of agricultural goods will need to be less high – and so the rate of growth will be higher – the lower is the price of agricultural goods required for landlords to release their surplus in exchange for consumable industrial goods.

Note that the share of income that accrues to industrial capitalists will increase during the growth process, as the capitalist sector grows in size. This suggested to Lewis (1954, p. 155) that a growth process of this kind might help to solve what he called the 'central problem' of development: the need to raise the savings rate enough to enable rapid growth to take place. In this model it is necessary to transfer labour into industry, in order to increase the overall savings rate of the economy. This is due to the behavioural assumption that agricultural income is not saved; we revisit this assumption below. Interestingly – from today's point of view – Lewis thought that a savings rate of ten to twelve per cent might be sufficient to achieve the 'rapid capital accumulation' that he believed integral to the process of development (Lewis 1954, p. 155). Note also that increasing inequality is a frequent, if not necessary, correlate to this rising savings share, at least in the early stages of development (see, for example, Fei et al. 1979). This story thus also provides an explanation of the 'Kuznets curve'.

Generalizations

Lewis does sometimes enlist the extreme simplifications made above. They correspond most closely to those made by Gustav Ranis and John C. H. Fei (1961), who used them to explain, more formally than Lewis did, what they call the 'first phase' of economic development – a phase in which there is 'pure' surplus labour. But Lewis also hints at many ways in which these assumptions could be relaxed. Ranis and Fei, along with Dale W. Jorgenson (and many others), went on to consider the implications of dualism when there are sectoral asymmetries different from those outlined above. In what follows, we consider a number of these extensions.

The first, and most fundamental, generalization of Lewis's model was made by Ranis and Fei (1961), who demonstrated that the dualistic framework continued to give insight into the process of economic growth even when the condition of pure surplus labour does not hold. They initiated a large body work on this question by examining the microeconomic foundations of surplus labour and exploring what occurs when these conditions come to an end. This occurs when a sufficient number of workers have been removed from agriculture for the marginal productivity of the remaining agricultural workers to become positive. As a result, agricultural output declines as further workers leave. (This may happen even if there is technological progress in agriculture, providing that this progress is not sufficient to fully compensate for lost labour.) Consequently, the marginal agricultural surplus per worker, which accrues to landlords as each worker leaves – and which is traded by landlords for industrial goods – begins to decline, *even if* the wage per worker (measured in terms of agricultural goods) is exogenous. This means that the cost of labour to industry, measured in terms of industrial goods, will

begin to rise above the level described in the sketch above – thereby constraining the rate of growth. This is the 'first turning point' identified by Ranis and Fei. It corresponds to the onset of Ricardo's 'dismal' diminishing returns. Ranis and Fei label what happens beyond this point as the 'second phase' of economic development. In that phase the economy is characterized by 'disguised unemployment', since labour in agriculture is still paid more than its marginal product.

Lewis himself was accused of not allowing for this possibility, even though he had written that the existence of zero marginal product is 'not... of fundamental importance to our analysis' (Lewis 1954, p. 142). This accusation led to what Lewis later called an 'irrelevant and intemperate controversy' about the existence, or not, of 'pure' surplus labour (Lewis 1972, p. 77). Ranis (2003, p. 8) agrees with Lewis's self-defence: in a retrospective assessment, he describes the postulation of a 'pure' labour surplus as a red herring. Amartya Sen (1966) helpfully clarifies the debate about this issue.

Growth becomes more difficult in this second stage of development. Recall that Lewis argues that the real wages per worker, and the level of welfare per worker, do not fall as growth proceeds. But growth is driven by the transfer of labour from agriculture to industry, which, in this second phase, causes agricultural output to fall. As a consequence of this the relative price of agricultural goods rises, and real wages can remain constant only if workers are able to substitute towards industrial goods in such a way as to avoid any damage to their welfare.

The Agricultural Sector as a Constraint on Growth

To highlight the essential role of such substitution, Mukesh Eswaran and Ashok Kotwal (1993) assume an extreme version of Engel's law. Consumers are assumed to spend *all* their income on food until they reach a particular threshold level of consumption, when they become sated with food. Beyond this point all further increases in consumption are devoted to industrial goods. At the same time they assume that labour always has a positive marginal product in agriculture. Under these assumptions, if workers remain so poor that they are not sated with food, then the transfer of labour across sectors – and therefore accumulation of industrial capital – becomes impossible. The inability of the poor to 'eat shirts' – an extreme version of what Ranis (2003) describes as the 'product' dimension of dualism – becomes a constraint on whether savings can lead to development. (And this constraint will bind quite independently of how high the marginal *physical* product of labour is in industry.) Any attempt to increase savings rates, in the manner desired by Lewis, so as to draw labour out of agriculture, would fail in these circumstances. The withdrawal of labour would lead to a reduction in the supply of food per worker – the only thing that matters for workers' real wages – and so to a shortage of food. That shortage would turn the terms of trade against industry, depressing industrial profits and savings until the downward pressure on the supply of food had been removed, or until growth has ceased. As a result, all the gains from any increase in industrial production would accrue, in the form of lower prices, to those who consume industrial goods, rather than enabling growth, as in the Lewis model. It is thus clear that an important influence on whether development can proceed under dualism is the ability to shift workers' demands away from agricultural goods.

Of course, in a small, open, economy, the relative prices of tradables will be tied down, and the economy can respond to any developing shortage of food simply by exporting manufactures and importing food. That was Ricardo's insight, over 100 years earlier, about the gains to Britain from the abolition of the Corn Laws; Lewis's model of dualism in the world economy also incorporates such trade. But Lewis (1972, p. 94) cautions that there may be limits to this if export prices are not really exogenous, and if, instead, the county needs to cheapen its exports to pay for the imports of food – and other goods – that it will need as it grows. Perhaps partly because of this, Lewis (1954, p. 176) argues that a country which exhausts its surplus labour supply might instead export its savings, investing in industrial

development in countries where the surplus labour condition continues to hold, and so enabling the output of manufactured goods to grow without driving down the rate of profit. In addition, the country might import labour from these countries. In this way Lewis's early contributions anticipated, and fed into, debates about the roles of outsourcing and immigration in contemporary globalization.

Jorgenson (1961) further develops the study of the dynamics of a dualistic economy in this second phase of development – when there is a positive marginal product of labour in agriculture and disguised unemployment. He incorporates a Malthusian perspective, by supposing that population growth is increasing in the amount of food consumed per capita, up to a biological ceiling that corresponds to the food-consumption threshold of Eswaran and Kotwal. This has the consequence that too rapid a rate of growth of population can cause a Malthusian trap by preventing the emergence of any significant agricultural surplus. Growth of manufacturing activity, such as that analysed by Lewis, can then be sustained only if technological progress in agriculture enables food production to outstrip population growth. (Capital accumulation in agriculture could have a similar effect in a model more general than that used by Jorgenson.) Only then can an agricultural surplus emerge, and grow, and so only then can labour progressively move away from agriculture. If this does not happen, then any increases in profits, savings and capital accumulation in industry become self-defeating, since they turn the terms of trade against industry and so bring down profits and savings, and bring growth to an end, in the way described two paragraphs above.

As stressed by Avinash Dixit (1973, p. 346), such a model focuses on 'the constraint on growth imposed by the rate of release of labour from agriculture', whereas in Lewis's model the focus had been on the ability of capital accumulation in industry to soak up the surplus labour force in agriculture. Nevertheless, as Dixit notes, growth paths in the two models will produce similar outcomes. In particular, in both models one would observe an endogenous rise in the savings rate as development proceeds. And in both models, it *may* be the case that any attempt to foster growth in industry, by a 'big push' to save more, is self-defeating. (This can be true in Jorgenson's model, and as we saw above, it can also be true beyond the 'first stage' of growth in the Lewis model, if it is not possible to induce workers to substitute away from agricultural goods.) This is why Jorgenson thought of increases in savings rates as an *outcome* of development, not as a policy tool which can be used to *promote* development (Jorgenson 1961, p. 328).

It is worth contrasting this view of potential 'development traps' with that which had been put forward in the 1940s by Paul Rosenstein-Rodan (1943), who built on his experience of eastern Europe. Rosenstein-Rodan's viewpoint also came from thinking about the interaction between agriculture and industry; like Lewis, he argued that development could only come to an agricultural economy through a process of industrialization. This, he argued, is because only industrial capitalists could afford to pay for the large fixed costs that are necessary to enable them to produce goods in a modern way, with low marginal costs. But if most people live in an impoverished agricultural sector then this would constrain their incomes, and so would limit their demand for modern industrial goods. That might make it unprofitable to make the required investment, and so might thwart the process of development. Here, just as for Lewis, a shortage of savings can be *the* problem of development. But by contrast with Lewis, a big push might fix it, since, roughly speaking, if all capitalists invested at once and paid their workers higher wages, then the demand for industrial goods would grow, making the investment worthwhile. This insight gave birth to the other great analytical engine of development economics, subsequently formalized by Kevin M. Murphy, Andrei Shleifer and Robert W. Vishny (1989) and Kiminori Matsuyama (1991), and well explained by Paul Krugman (1993). Since the pecuniary externalities that allow an economy to escape from a development trap are accessible only in the 'modern' sector, asymmetries between sectors are also central to this view.

Further Aspects of Labour Transfer

The Lewis model was also generalized to explain the gap between the wage paid in the rural sector and that paid in the urban sector and to explore the consequences of such a gap. Lewis himself (1954, p. 150) acknowledged the existence of a wage gap, and suggested that it may result from the psychological costs of lifestyle changes, from the need to reward skills accumulated in the urban sector, or from the ability of workers in cities to bargain for higher wages. (This is particularly relevant when we recognize that the urban sector includes government employment and some services.) Subsequent authors took up this question, arguing, for example, that wage premia may arise because they lead to greater productivity through effects on health or employee motivation (for example, Dasgupta and Ray 1986; Shapiro and Stiglitz 1984).

The consequences of such a gap, for the process of labour transfer from agriculture to industry, were set out in the celebrated work of John Harris and Michael P. Todaro (1970). It may be that a wage floor in the urban formal sector prevents the market from clearing there. If a wage floor operates, then workers who choose to leave the rural sector face the prospect of receiving an urban wage which is above that of the rural sector, if they get employed, but also face some probability of becoming unemployed. In the simplest version of this model, equilibrium occurs when labour migration equalizes expected income across sectors – an outcome in which the rural wage equals a weighted average of the incomes received by employed and unemployed urban workers, weighted according to the probability of unemployment in the urban sector. Even without this extreme outcome, there are important policy implications in such a model. The more elastic is labour supply to the urban sector with respect to expected income there, the greater the amount of urban unemployment that will be induced by any policies that increase urban wages. This incorporation of urban unemployment into the model also enables one to begin to discuss the growth of a third sector: the production of services in cities (see Fields 1975). Roughly speaking, we can say that services get produced by (some of) those who migrate to cites, but do not get a job in manufacturing.

It is clear that the expansion of the industrial sector will ultimately take the economy beyond the second phase of economic development, in which there is disguised unemployment. This is because withdrawal of labour from agriculture will eventually reach the point at which the marginal product of the remaining labour rises to equality with the subsistence wage. Ranis and Fei call this a 'second turning point'. At this point the marginal worker, offered a subsistence wage, can now instead offer his or her labour to a higher bidder. From then on the wage (measured in agricultural goods) will begin to rise in both sectors as growth continues. We can say that the 'dualistic' structure of the economy then comes to an end, in that the rural economy becomes 'commercialized'. (Something similar, too, will happen in any services sector.) That leads one back to a labour-scarce economy, the analysis of which is better suited to neoclassical theory. A two-sector neoclassical growth model – something like the model of Hirofumi Uzawa (1961, 1963) – may be a better way to think about growth in these circumstances.

One key strand of the story of dualism that we have been telling is the assumption that capitalists save, but workers (and landlords) do not. Lewis's explanation of this asymmetry is largely behavioural. But such differences in savings rates between the traditional and the modern sectors might also be explained *institutionally*, by means of credit-market imperfections. If a technological asymmetry precludes investment in rural areas, and if limited financial development means that rural residents lack access to investment opportunities in manufacturing, then the agricultural surplus will not be used directly to finance investment. Moreover, typical characterizations of credit-market imperfections highlight the moral hazard problems that persist in rural areas because the poor there are unable to provide the kind of collateral required for formal-sector loans. (Small rural landholdings are of limited use as collateral.) Such lack of collateral stands as a barrier to borrowing, even though

loans might be used to facilitate growth by promoting education, or capital accumulation, or technical progress in agriculture. (See Ray 1998, for a summary of these arguments.) By contrast, Abhijit Banerjee and Andrew Newman (1998) provide an alternative perspective, emphasizing a sectoral asymmetry in the *informational* dimensions of credit-market imperfections, and showing how this can affect the willingness of individual workers to migrate in a dualistic economy. They present a model in which there is access to credit for consumption in rural areas. Given that workers have limited collateral wherever they live, a crucial determinant of their access to credit is the amount of information that lenders have about prospective borrowers. In contrast with the relative anonymity of urban life, small communities of the rural sector may provide superior information about borrowers, and thus foster lending. Banerjee and Newman show that dualism, characterized in terms of this differential severity of information asymmetries, might lead to a suboptimal allocation of labour across sectors. By financing consumption in the rural sector, rural credit might actually provide an incentive for labour to remain there; this incentive could offset the relatively high wages of the modern sector and could thereby impede the development process. Their paper suggests – at the least – that the lens of asymmetric information can shed useful light on the development of such economies.

Defining Characteristics of Economic Dualism

We conclude by noting that we have described a number of reasons for differences between the industrial and agricultural sectors of a developing economy. Just as in Lewis's original article, all these differences go beyond mere asymmetries in production technologies or factor endowments between the sectors. This is why, following Ravi Kanbur and James McIntosh (1987), we would not normally describe the two-sector growth models of Uzawa (1961, 1963) as models of dualism, even though in those models the two sectors have different factor intensities. Nor would we say that that the two-sector Hecksher–Ohlin model of international trade is a model of a dualistic economy – even when its two sectors have different factor intensities, and even when the two sectors are labelled 'agriculture' and 'industry'. Furthermore, although the specificity of factors to sectors appears central to Lewis's set-up (with land specific to agriculture and capital specific to industry), this feature does not seem to be sufficient to merit the label of 'dualism'. Thus, for example, we would not regard the short-run version of the Heckscher–Ohlin trade model presented by J. Peter Neary (1978), with factors specific in each of the two sectors, as portraying a dualistic economy.

Instead, we would argue that the defining characteristic of modern theories of economic dualism lies – just as it did in Lewis's article – in a focus on sectoral asymmetries that are not simply technological. For Lewis, and for Ranis and Fei, there were *organizational* differences between sectors – in that wages were assumed to be determined by institutional factors in the agricultural sector – and *behavioural* differences between sectors – in that those in the rural sector were assumed to be unwilling to save, while capitalists were assumed to save everything. A focus on these features might imply that 'pull' factors drive labour transfer, and hence economic growth, in a dualistic economy. But since Lewis, economists studying economic development have explored alternative asymmetries between sectors and have reached different conclusions. The model of Eswaran and Kotwal, in which the defining asymmetries are *product* asymmetries – an assumption that all income is spent on agricultural goods until some threshold – highlights the need for labour productivity increases in agriculture to avoid stagnation of real wages. This is a need that persists even in the presence of rising productivity in industry. Jorgenson, who coupled such a view with a demonstration that Malthusian pressures can prevent income from ever rising above this threshold, showed clearly that growth can be constrained unless the 'push' factor of growth in agricultural technology is strong enough. Baneijee and Newman, by contrast, have emphasized that

informational asymmetries between traditional and modern sectors can constrain the growth process.

We thus believe that, in the study of any particular economy, it is important to understand which asymmetries impose binding constraints on growth. Different constraints imply the need for different policies. But identifying the relevant asymmetries is even more important if we wish to remove these underlying constraints themselves. Joseph Stiglitz has proposed that we do just this, advocating what he calls 'growth strategies based on duality's elimination' (Stiglitz 1999, p. 56). Much empirical work is necessary if we are to understand what such strategies might require.

See Also

- ▶ Labour Surplus Economies
- ▶ Lewis, W. Arthur (1915–1991)

Bibliography

Banerjee, A.V., and A.F. Newman. 1998. Information, the dual economy, and development. *Review of Economic Studies* 65: 631–635.

Dasgupta, P., and D. Ray. 1986. Inequality as a determinant of malnutrition and unemployment: theory. *Economic Journal* 96: 1011–1034.

Dixit, A. 1973. Models of dual economies. In *Models of economic growth: Proceedings of a conference held by the international economic association at Jerusalem*, ed. J.A. Mirrlees and N.H. Stern. London: Macmillan.

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

Eckaus, R.S. 1955. The factor proportions problem in underdeveloped areas. *American Economic Review* 45: 539–565.

Eswaran, M., and A. Kotwal. 1993. A theory of real wage growth in LDCs. *Journal of Development Economics* 42: 243–269.

Fei, J.C.H., G. Ranis, and S.W.Y. Kuo. 1979. *Growth with equity: The Taiwan case*. Oxford: Oxford University Press.

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

Findlay, R. 1980. On W. Arthur Lewis' contributions to economics. *Scandinavian Journal of Economics* 82: 62–79.

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

Harrod, R.F. 1939. An essay in dynamic theory. *Economic Journal* 49: 13–33.

Hicks, J. 1965. *Capital and growth*. Oxford: Oxford University Press.

Higgins, B. 1956. The 'dualistic theory' of underdeveloped areas. *Economic Development and Cultural Change* 4: 99–115.

Jorgenson, D.W. 1961. The development of a dual economy. *Economic Journal* 71: 309–334.

Kanbur, R., and J. McIntosh. 1987. Dual economies. In *The new Palgrave: A dictionary of economics*, ed. J. Eatwell, M. Milgate, and P.K. Newman, vol. 1. London: Macmillan.

Krugman, P. 1993. Toward a counter-counterrevolution in development theory. In *Proceedings of the World Bank Annual conference on development economics, 1992: Supplement to the World Bank economic review and the World Bank research observer*, ed. L.H. Summers and S. Shah. Washington, DC: World Bank.

Lewis, W.A. 1953. *Report on industrialisation and the Gold Coast*. Accra: Government Printing Department.

Lewis, W.A. 1954. Economic development with unlimited supplies of labour. *Manchester School* 28: 139–191.

Lewis, W.A. 1955. *The theory of economic growth*. London: Allen & Unwin.

Lewis, W.A. 1972. Reflections on unlimited labour. In *International economics and development: Essays in honour of Raul Prebisch*, ed. L.E. di Marco. London: Academic.

Lewis, W.A. 1978. *Growth and fluctuations, 1870–1913*. London: Allen & Unwin.

Lewis, W.A. 1979. Autobiography. In Lindbeck (1992). Also online. Available at http://nobelprize.org/nobel_prizes/economics/laureates/1979/lewis-autobio.html. Accessed 9 Jan 2007.

Lindbeck, A., ed. 1992. *Nobel lectures: Economic sciences 1969–1980*. Singapore: World Scientific Publishing.

Matsuyama, K. 1991. Increasing returns, industrialization, and the indeterminacy of equilibrium. *Quarterly Journal of Economics* 106: 617–650.

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

Neary, J.P. 1978. Short-run capacity specificity and the pure theory of international trade. *Economic Journal* 88: 488–510.

Pasinetti, L. 1974. *Growth and income distribution – Essays in economic theory*. Cambridge: Cambridge University Press.

Preobrazhensky, E. 1924. *The new economics*. Trans. B. Pearce. Oxford: Clarendon Press, 1965.

Ranis, G. 2003. Is dualism worth revisiting? Discussion Paper No. 870. Economic Growth Center, Yale University.

Ranis, G., and J.C.H. Fei. 1961. A theory of economic development. *American Economic Review* 51: 533–565.

Ray, D. 1998. *Development economics*. Princeton: Princeton University Press.
Rosenstein-Rodan, P. 1943. Problems of industrialisation of eastern and south-eastern Europe. *Economic Journal* 53: 202–211.
Sah, R.K., and J.E. Stiglitz. 1984. The economics of price scissors. *American Economic Review* 74: 125–138.
Sen, A. 1966. Peasants and dualism with or without surplus labour. *Journal of Political Economy* 74: 425–450.
Shapiro, C., and J.E. Stiglitz. 1984. Equilibrium unemployment as a worker discipline device. *American Economic Review* 74: 433–444.
Solow, R.M. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70: 65–94.
Stiglitz, J.E. 1999. Duality and development: some reflections on economic policy. In *Development, duality, and the international economic regime: Essays in honor of Gustav Ranis*, ed. G. Saxonhouse and T.N. Srinivasan. Ann Arbor: University of Michigan Press.
Swan, T.W. 1956. Economic growth and capital accumulation. *Economic Record* 32: 334–361.
Uzawa, H. 1961. On a two-sector model of economic growth. *Review of Economic Studies* 29: 40–47.
Uzawa, H. 1963. On a two-sector model of economic growth II. *Review of Economic Studies* 30: 105–118.

Dual Track Liberalization

Yingyi Qian

Abstract

Dual track liberalization is a reform strategy in which a market track is introduced while the plan track is maintained at the same time. Dual track liberation is Pareto improving in the sense that it makes some people better off without making anybody worse off. Because prices are liberalized at the margin, dual track liberalization can also achieve efficiency. China used the dual track reform strategy in liberalizing many markets such as the markets of agricultural goods, industrial goods, consumer goods, foreign exchange, and labour, as well as in creating special economic zones.

Keywords

Allocative efficiency; China, economics in; Compensatory transfers; Corruption; Dual track liberalization; Foreign exchange control; Market liberalization; Pareto efficiency; Planning; Price control; Price liberalization; Rationing; Rent seeking; Special economic zones (China)

JEL Classification

P3

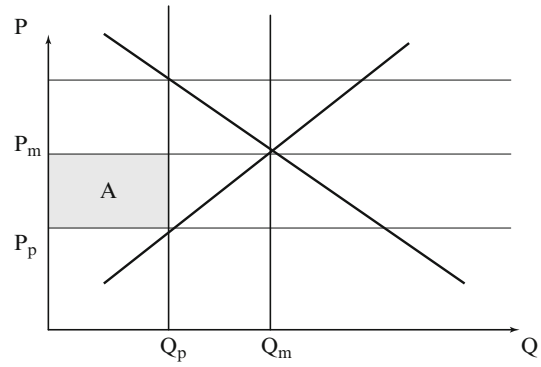
Dual track liberalization is a reform strategy of market liberalization in which a market track is introduced while the plan track is maintained at the same time. Under the plan track, economic agents are assigned rights to and obligations for a fixed quantity of goods and services at fixed planned prices as specified in the pre-existing plan. Under the market track, economic agents can participate in the market at free market prices, provided that they fulfil their obligations under the pre-existing plan. The essential feature of the dual track strategy to market liberalization is that prices are liberalized at the margin while inframarginal plan prices and quotas are maintained for some time before being phased out. Although the dual track reform strategy is widely adopted in China during its transition from plan to market, it is also used in other countries. For example, when introducing new legislation, a 'grandfathering' clause is often adopted to protect existing interests, which is a form of the dual track approach to reform.

Analysis of dual track liberalization follows two lines of approach. The first focuses on its Pareto-improvement property, that is, dual track liberation makes nobody worse off while it makes somebody better off – and therefore it has a political advantage in implementing reforms. Most efficiency-improving market liberalization reforms potentially create winners and losers, despite the fact that, in theory, efficiency gains should be large enough to allow the potential losers to be compensated. For example, the single track approach to liberalization (that is, where all

the prices are freed at once) in general cannot guarantee an outcome without losers. Dual track liberalization means that planned quantity continues to be delivered at plan price but any additional quantity can be sold freely in the market. With the dual track, the surpluses of the rationed users and the planned suppliers remain exactly the same. The purpose of maintaining the plan track is to provide implicit transfers to compensate potential losers from market liberalization by protecting status quo rents under the pre-existing plan. On the one hand, the introduction of the market track provides the opportunity for economic agents who participate in it to be better off. At the same time, the new users and suppliers outside the plan are also better off. Therefore, the intuitive appeal of dual track liberalization for reformers lies precisely in the fact that it represents a mechanism of the implementation of a reform without creating losers (Lau et al. 2000).

The second approach focuses on the efficiency property of dual track liberalization. Pareto-improvement property implies that it always improves efficiency. This is independent of other assumptions, for example, as to whether the market is competitive or not. In contrast, the single track approach to liberalization may improve efficiency under perfect competition, but may not improve efficiency if the market is monopolistic (Li 1999). The more subtle and deeper point is that the dual track approach to liberalization may achieve allocative efficiency, despite the fact that it appears inefficient, by maintaining the inefficient planned track. The fundamental reason is that the compensatory transfers, which are implicitly embodied in the planned track, are inframarginal, and thus the distortion can be avoided.

To see this we look at the special case where the pre-reform status quo features efficient rationing and efficient planned supply in the sense that the planned output is allocated to users with the highest willingness to pay and the planned supply is delivered by suppliers with the lowest marginal costs. Nevertheless, the price of the good is fixed at an artificially low level and the production quota is fixed below market equilibrium (Fig. 1). When the market track is introduced into this



Dual Track Liberalization, Fig. 1 The case of efficient supply and efficient rationing

setting, it is clear that the market equilibrium quantity and price would be identical to the case without the planned price and quota to start with. Therefore, dual track liberalization achieves efficiency. Notice that efficiency is achieved without making anyone worse off. Indeed, the rents enjoyed by the buyers under rationing (area A in Fig. 1) are preserved under dual track liberalization, but would be lost under single track liberalization.

In a more general case of inefficient rationing and/or inefficient planned supply, efficiency can still be achieved provided market liberalization is full, in the sense that market resales of plan-allocated goods and market purchases by planned suppliers for fulfilling planned delivery quotas are permitted after the fulfilment of the obligations of planned suppliers and rationed users under the plan. This removes any inefficiency associated with the original planned prices and quotas and makes imputed rents under planning inframarginal. This type of transaction takes many forms in practice, for example, subcontracting by inefficient planned suppliers to more efficient non-planned suppliers, and labour reallocation when workers in inefficient enterprises keep the housing while taking a new job in more efficient firms. In both examples, after fulfilling the obligations under the plan (planned delivery of supply and welfare support through housing subsidies), the market track functions to undo the inefficiency of the plan track.

This above partial equilibrium analysis can be generalized to a general equilibrium mode (Lau et al. 1997). Efficiency requires full market liberalization under which market resales, subcontracting, and market purchases for redelivery are all allowed. Indeed, the distinction between limited and full market liberalization is a major difference between Lau et al. (1997) and Byrd (1991), and others who have studied the dual track approach.

If such resales and purchases are not allowed or cannot be achieved, then dual track liberalization is limited and efficiency in general cannot be achieved, although it can be improved. Of course, in the special case discussed above with efficient supply and efficient rationing, dual track with limited market liberalization is the same as dual track with full market liberalization. In general, dual track limited market liberalization need not be the same as dual track full market liberalization.

Sometimes dual track liberalization of the market takes the following sequential form: in a first stage, limited market liberalization is implemented, and then in a second stage full market liberalization is implemented. In the first stage, going from a centrally planned economy to limited market liberalization, Pareto improvement is clearly attained, but efficiency cannot be guaranteed. Specifically, limited market liberalization generally leads to inefficient overproduction due to market entry. In the second stage, when full liberalization is introduced, efficiency is attained but Pareto improvement may not be. This is because the second-stage full market liberalization implies efficiency, and thus there must be a production contraction and some people have to reduce production and are made worse off. Therefore, the sequential dual track liberalization may result in some opposition to further reforms after the first and before the second stage, while the dual track full market liberalization that is implemented in one stroke will not. Nevertheless, it is also clear that, even under the sequential dual track liberalization, there are no losers at the end of the second stage compared with the status quo before the reform.

The dual track approach to market liberalization is an example of reform making the best use of existing information and institutions. First, it utilizes efficiently the existing information embedded in the original plan (that is, existing rents distribution) and its implementation does not require additional information. Second, it also enforces the plan through the existing plan institutions and does not need additional institutions. Enforcement of the plan track is crucial for preserving pre-existing rents. However, contrary to common understanding of the relationship between state power and reform, state enforcement power is needed here not to implement an unpopular reform, but to carry out one that creates only winners, without losers.

Economists sometimes find dual track liberalization puzzling and counter-intuitive, for several reasons. First, economists are used to the law of one price: in a competitive setting, multiple prices entail inefficiency. However, in dual price liberalization, the planned price comes together with planned quantity, when they are fixed, they do not entail inefficiency, at least not additional inefficiency. Second, dual track resembles price control, which is associated with inefficiency and rent seeking. But dual track is not price control; on the contrary, it is a move towards price liberalization. An important difference between the plan track under dual track and price control is that the plan track embodies both fixed prices and fixed quantities; it is a package of price and quantity control, not just price control. Under pure price control, the government fixes only prices, but not quantities. Third, to reformers, dual track seems a partial reform and not a complete reform. This is true under dual track with limited market liberalization, but not true with full market liberalization. Although dual track with limited market liberalization does not achieve efficiency, it improves efficiency and makes nobody worse off.

Dual track liberalization requires enforcement of the rights and obligations under the plan track. In fact, enforcement of the plan track alone would prevent any decline in aggregate output. Can the plan track be enforced? With a collapsing government, it cannot. But enforcing the pre-existing plan is informationally much less demanding for the government than drawing up a new plan. Under central planning, the information

requirement for drawing up a plan is huge. Enforcing a pre-existing plan is different. In fact, the dual track approach uses minimal additional information as compared with other possible compensation schemes that may be used with other approaches to reform. Compliance with the plan by economic agents depends on their expectations of the credibility of state enforcement. If state enforcement is not credible, then the economic agents will have no incentive to fulfil their plan obligations. If people think that they are not going to receive the plan- mandated deliveries at plan prices, they will not make the plan-mandated sales at the fixed plan prices. In that case, dual track liberalization degenerates to single track liberalization.

Lack of enforcement of the plan track may result in supply diversion as analysed by Murphy et al. (1992). These authors studied a partial reform model with the following two crucial assumptions: (i) suppliers are free to sell to all users, and (ii) buyers who are not covered by the plan can freely purchase inputs at any price, but buyers who are covered by the plan are not allowed to purchase inputs above the plan price. This partial reform model differs from the dual track liberalization model in an important respect: there is no plan delivery quota enforced on the suppliers.

In their model, partial reform may lead to inefficient supply diversion to such an extent that the outcome can be worse than that without reform. Therefore, the partial reform is not only not Pareto improving, but also total welfare reducing. Consider the case where the initial condition is also characterized by efficient rationing and efficient supply as shown in Fig. 1, where the planned price Pp is below the market clearing level P_m. Then, after the partial liberalization as defined above, suppliers can sell the good freely to the highest bidders. While the firms under the plan are forced to buy the good at price P_p, the firms outside the plan are free to buy the good at any price. Then they will bid the good for price $P_p + \varepsilon$ where ε is a positive but small number. Because the firms under the plan are constrained to pay P_p, an amount will be diverted from them to those not covered by the plan. Because the willingness to pay from those not covered by the plan is lower than those covered by the plan (by the assumption of efficient rationing), this kind of partial reform induces a net efficiency loss. While the sector not covered by the plan gains, the sector covered by the plan loses, and the total welfare effect is unambiguously negative. Although the assumption of efficient rationing and efficient supply under central planning is too strong, the result of inefficient supply diversion under partial reform remains valid with weaker conditions about initial rationing and supply.

So which model is more relevant? It depends on the quota enforcement capability of the government. A good enforcement capability makes the dual track liberalization model of Lau, Qian and Roland more relevant, while a poor enforcement would make the partial reform model of Murphy, Shleifer, and Vishny more relevant. The dual track liberalization model is motivated mainly by the practice in China, where enforcement has been reasonably good, while the partial reform model is mainly motivated by the experiences of the last years of the Soviet Union, when the state enforcement power diminished quickly.

In China's context, lack of quota enforcement sometimes takes the following form. The government may be unable to freeze the plan by creating new quotas with (below market equilibrium) planned price and giving windfall rents to some people who are politically connected. This may lead to corruption: firms find it easier to make profits by lobbying the government for allocating more input goods delivery at low planned prices, without the corresponding obligations to deliver low price outputs as under central planning. They then sell the goods at the market price to receive the windfall gains. This type of corruption is often attributed to the dual track approach to liberalization. Indeed, without the coexistence of the planned prices and market prices, the above form of corruption is not possible. By eliminating the two prices, such form of corruption would disappear. However, the essence of the problem is the failure in the enforcement of the original planned track. If the planned track is strictly enforced, no new quotas should be created. (On the other hand, full market liberalization

allows for market arbitrage, which may increase the welfare of those who were allocated with goods at below-market prices. This is essential for achieving efficiency. The difference is that the potential rents are inherited from the previous regime in this case, not from a new creation.)

Dual Track Liberalization in Practice

Studies of dual track liberalization focus mostly on China, although other cases, such as that of Mauritius, are also mentioned. The origin of the dual track can be traced to the 1950s when China had two prices for grain, the official price and negotiated price. However, dual track approach to market liberalization as a reform strategy was used only after 1979, first in the agricultural goods markets, and then in other markets (Byrd 1991; Naughton 1995; Lau et al. 2000).

Agriculture Goods The agricultural reform in China started with a dual track approach to market liberalization. Under that reform, the commune (and later the household) was assigned the responsibility to sell a fixed quantity of output to the state procurement agency as previously mandated under the plan at predetermined plan prices and to pay a fixed tax (often in kind) to the government. It also had the right (and obligation) to receive a fixed quantity of inputs, principally chemical fertilizers, from state-owned suppliers at predetermined plan prices. Subject to fulfilling these conditions, the commune was free to produce and sell whatever it considered profitable, and retain any profit. Moreover, the commune could purchase from the market grain (or other) output for resale to the state in fulfilment of its responsibility. There was thus a full market liberalization.

Between 1978 and 1988 state procurement of domestically produced grain remained essentially fixed, with 47.8 million tons in 1978 and 50.5 million tons in 1988. During that same period, total grain output increased by almost one-third. But the dual track approach to liberalization applied to agricultural products other than grain: between 1978 and 1990, the share of transactions at plan prices in all agricultural goods fell from 94 to 31%, when the agricultural output in China doubled. There was a huge supply response to the introduction of the market track.

Industrial Goods The most noticeable and often cited application of the dual track approach to liberalization is to industrial goods (Byrd 1991; McMillan and Naughton 1992). The Chinese government issued a document in May 1984 stipulating that there would be two forms of production in state-owned enterprises: planned and non-planned. Correspondingly, there were two types of material supplies for enterprises, namely, state allocation and free purchase. Prices of goods in the former were fixed by the state and prices of goods above quota quantity could be sold in the market at price within a range up to 20% higher or lower than of the planned price. In February 1985, the 20% price cap was removed and the dual track for industrial goods was formally in place (Wu and Zhao 1987). As a result, the share of transactions at plan prices, in terms of output value, fell from 100% before the reform to 45% in 1990.

Coal and steel are the two important industrial commodities most tightly controlled under central planning, and both coal and steel markets were liberalized through the dual track approach. For coal, China's principal energy source, the planned delivery led to some slight increases in absolute terms during the 1980s, but the market track increased dramatically from 293 million tons to 628 million tons over the same period – the supply came mainly from small rural mines run by Township–Village Enterprises. As a result, the share of the plan allocation declined from 53% in 1981 to 42% in 1990. For steel, the plan track was quite stable in absolute terms during the 1980s, but the share of plan allocation fell from 52% in 1981 to 30% in 1990. In the cases of both coal and steel, because the plan track was essentially frozen, the economy was able to 'grow out of the plan' on the basis of the expansion of the market track (Naughton 1995).

Consumer Goods Prior to the economic reform of 1979, most essential consumer goods and

services for urban residents, such as grain, cooking oil, meat, electricity, housing, and the monthly transport pass, were rationed with coupons at values lower than corresponding free market prices. With dual track liberalization, urban residents continued to have the right to purchase grain, meat, electricity and housing at the same pre-reform prices and within the limits of the pre-reform rationed quantities, but, at the same time, they were also free to buy consumer goods from the free market at generally higher prices. The proportion of transactions at plan prices declined from 97% in 1978 to only 30% in 1990.

Foreign Exchange Under central planning, foreign exchange transactions were strictly controlled by the government at the official exchange rate. Exporters were required to surrender to the state all foreign exchange they earned at the official exchange rate, and importers were allocated with planned quotas of foreign exchange, also at the official exchange rate. Foreign visitors to China were required to use 'foreign exchange certificates', which were available at the official exchange rate. Starting from May 1988, China allowed trading of foreign exchange at Foreign Exchange Adjustment Centres (more commonly referred to as 'swap centres') at the rate determined by market supply and demand, called 'swap rate'. This was the beginning of the dual track in the foreign exchange market. The swap rate was, not surprisingly, significantly higher than the official rate. The supply of foreign exchange in the swap markets was provided by exporters through the foreign exchange they were allowed to retain from net increases in their export earnings in relation to the base period. By the end of 1993, transactions at official exchange rates accounted only for about 20% of the total; the rest were at the market rate.

Labour As in many other centrally planned economies, the labour market in China was also distorted: most labour was allocated to unproductive, state-owned enterprises and few to the non-state sector. Dual track liberalization in the labour market takes two forms. In the first, the non-state sector (the liberalized sector) pays market wages and decides on hiring and firing. Between 1978 and 1994, employment in the non-state sector increased by 318.8%, while employment in the state sector (including civil servants in government agencies and non-profit organizations) increased by only 50.5%. Second, even within the state sector there are also two tracks. Beginning in 1980, while pre-existing employees maintained their permanent employment status, most new hires in the state sector were made under the more flexible contract system and often at lower effective wage rates. Employment in the plan track was virtually stationary – it declined from 87.14 million in 1983, on the eve of the introduction of economic reform in industry, to 83.61 million in 1994.

Special Economic Zones Dual track liberalization can also have a geographical dimension: special economic zones are such examples. Although similar zones for processing exports can be seen in other Asian economies, special economic zones had a more profound effect in China because the whole country was still under central planning when they were created. Therefore, the purpose of special economic zones was more than for exporting; it was a strategy for market reform.

In 1980, China established four 'special economic zones', Shenzhen, Zhuhai and Shantou in Guangdong province and Xiamen in Fujian province. Most transactions relating to activities inside the zones were on the market track, including prices of input and output goods and wages of labour – at a time when the rest of the economy was still operating under central planning. The special economic zones were insulated from the rest of the economy to minimize the impact on and interaction with the rest of the economic system. Initially, firms inside the special economic zones had to import all their inputs and export all their outputs – thus creating no disruption to the domestic aggregate supply and demand. The principal purpose of this approach was to minimize the impact of new economic activities on the old-style domestic state-owned enterprises. Thus, once again, there were two tracks and the reform was Pareto improving.

In order for the special economic zones to work, merely creating them was not enough. One of the crucial conditions was the insulation of the non-liberalized sector from the liberalized sector so that the latter's existing rents could be maintained while the other sector was liberalized. Therefore, creation of special economic zones is a type of limited market liberalization. It is Pareto improving and efficiency enhancing, but cannot be fully efficient.

Phasing Out the Plan Track

With rapid growth, the plan track will become a matter of little consequence to most potential losers, which in turn reduces the cost required for compensating them. In China, the plan track in product markets was largely phased out during the 1990s. By 1996, the plan track was reduced to 16.6% in agricultural goods, 14.7% in industrial producer goods, and only 7.2% in total retail sales of consumer goods. However, this phasing-out of the plan track was generally accompanied by compensation. For example, urban food coupons (grain, meat, oil, and so on) were removed in the early 1990s with lump-sum compensation. But the cost of compensation was much smaller in relative terms as compared to the potential cost of compensation in the early 1980s. The dual track exchange rate ended on 1 January 1994, when the two exchange rates – the official rate and the swap rate – were merged into a single, market rate. In this last step of foreign exchange reform, those organizations that used to receive cheap foreign exchange were provided with annual lump-sum subsidies for a period of three years, which was sufficient for them to purchase the pre-reform allocation of foreign exchange. Because at that time the share of centrally allocated foreign exchange had already fallen to less than 20% of the total, the cost of compensation was not too large.

See Also

► China, Economics in

Bibliography

Byrd, W.A. 1991. *The market mechanism and economic reforms in China*. Armonk: M.E. Sharpe.

Lau, L.J., Y. Qian, and G. Roland. 1997. Pareto-improving economic reforms through dual track liberalization. *Economics Letters* 55: 285–292.

Lau, L.J., Y. Qian, and G. Roland. 2000. Reform without losers: An interpretation of China's dual track approach to transition. *Journal of Political Economy* 108: 120–143.

Li, W. 1999. A tale of two reforms. *RAND Journal of Economics* 30: 120–136.

McMillan, J., and B. Naughton. 1992. How to reform a planned economy: Lessons from China. *Oxford Review of Economic Policy* 8: 130–143.

Murphy, K.M., A. Shleifer, and R.W. Vishny. 1992. The transition to a market economy: Pitfalls of partial reform. *Quarterly Journal of Economics* 107: 887–906.

Naughton, B. 1995. *Growing out of the plan*. Cambridge: Cambridge University Press.

Wu, J., and R. Zhao. 1987. The dual pricing system in China's industry. *Journal of Comparative Economics* 11: 309–318.

Duality

Lawrence E. Blume

Abstract

This article surveys duality in producer theory, consumer theory and welfare economics. As opposed to the usual analysis through first-order conditions for optimization, the various dualities are derived here from convex duality theory, using Fenchel transforms and subdifferentials.

Keywords

Antonelli, G.B.; Bergson–Samuelson social welfare function; Convex programming; Convexity; Cost functions; Cyclic monotonicity; Duality; Envelope th; Fenchel transform; Firm, theory of the; Hicksian-compensated demand; Hotelling, H.; Hotelling's lemma; Hyperplanes; Indirect utility function; Lagrange multipliers; Marginal revolution; Monotonicity; Profit functions; Quasi-

equilibrium; Saddlepoints; Separation th; Shephard, R.W.; Shephard's lemma

JEL Classifications
D0

Introduction

The word 'duality' is often used to invoke a contrast between two related concepts, as when the informal, peasant, or agricultural sector of an economy is labelled as dual to the formal, or profit-maximizing, sector. In microeconomic analysis, however, 'duality' refers to connections between quantities and prices which arise as a consequence of the hypotheses of optimization and convexity. Connected to this duality are the relationship between utility and expenditure functions (and profit and production functions), primal and dual linear programs, shadow prices, and a variety of other economic concepts. In most textbooks, the duality between, say, utility and expenditure functions arises from a sleight of hand with the first-order conditions for optimization. These dual relationships, however, are not naturally a product of the calculus; they are rooted in convex analysis and, in particular, in different ways of describing a convex set. This article will lay out some basic duality theory from the point of view of convex analysis, as a remedy for the microeconomic theory textbooks the reader may have suffered.

Mathematical Background

Duality in microeconomics is properly understood as a consequence of convexity assumptions, such as laws of diminishing marginal returns. In microeconomic models, many sets of interest are closed convex sets. The mathematics here is surveyed in convex programming. The urtext for this material is Rockafellar (1970).

Closed convex sets can be described in two ways: by listing their elements, the 'primal' description of the set, and by listing the closed half-spaces that contain it. A closed (upper) half-space in $\mathbf{R^n}$ is a set of the form $h_{pa} = \{x : p \cdot x \geq a\}$, where p is another n-dimensional vector, a is a number and $p \cdot x$ is the inner product. The vector p is the *normal vector* to the half-spaces h_{pa}. Geometrically speaking, this is the set of points lying on or above the line $p \cdot x = a$. The famous separation theorem for convex sets implies that every closed convex set is the intersection of the half-spaces containing it.

Suppose that C is a closed convex set, and that p is a vector in $\mathbf{R^n}$. How do we find all the numbers a such that $C \subset h_{pa}$? If there is an $x \in C$ such that $p \cdot x < a$, then a is too big. So the natural candidate is $w = \inf_{x \in C} p \cdot x$. If $a > w$ there will be an $x \in C$ such that $p \cdot x < a$ on the other hand, if $a < w$, then $p \cdot x > a$ for all $x \in C$. So the half-spaces h_{pa} for $a \leq w$ are the closed half-spaces containing C.

This construction can be applied to functions. A concave function on $\mathbf{R^n}$ is an $[-\infty, \infty)$ valued function f such that the *hypograph* of f the set hypo $f = \{(x, a) \in \mathbf{R^{n+1}} : a \leq f(x)\}$, is convex. If hypo f is closed, f is said to be *upper semi-continuous* (usc). The *domain* dom f of concave f is the set of vectors in $\mathbf{R^n}$ for which f is finite-valued. Concave (and convex) functions are very well-behaved on the *relative interiors* of their effective domains. The relative interior ri C of a convex set C is the interior relative to the smallest affine set containing C (see convex programming), and on ri dom f, f (concave or convex) is continuous.

Suppose that f is usc. The minimal level a such that $h_{(p, -1)a}$ the hyperplane in $\mathbf{R^{n+1}}$ with normal vector $(p, -1)$, contains hypo f is $f^*(p) = \inf_x p \cdot x - f(x)$. Why the normal vector $(p, -1)$? Because the graph of the affine function $x \mapsto f^*(p) + px$ is a tangent line to f, the graph of f lies everywhere beneath it, and no other line with the same slope and a smaller intercept has this property. The function $f^*(p)$ is the (concave) *Fenchel transform* or *conjugate* of f, and is traditionally denoted f^*. The construction of the preceding paragraph can be done just this way: the *concave indicator function* of a convex set C is the function $\delta_c(x)$ which is 0 on C and $-\infty$ otherwise, and $\delta^*_c(p) = \inf_{x \in C} p \cdot x$. For any function f, not necessarily usc or

concave, the Fenchel transform f^* is usc and concave. If f is in fact both usc and concave, then $f^{**}=f$. This fact is known as the conjugate duality theorem. Convex functions with range $(-\infty, \infty]$ are treated identically. The function f is convex if and only if $-f$ is concave, but the definitions are handled slightly differently in order to preserve the intuition just described. The set epi $f = \{x, a) : a > f(x)\}$, and the convex Fenchel transform is defined differently: $f^*(p) = \sup_x p \cdot x - f(x)$. The convex indicator function of a convex set C is the function $\delta^C(x)$ which is 0 on C and +otherwise; its (convex) conjugate is $\delta^{C*}(p) = \sup_x p \cdot x$. These facts are discussed in convex programming.

If concave functions have tangent lines, then they must have something like gradients. A vector p is a *sub gradient* of f at x if $f(x) + p \cdot (y - x) \leq f(y)$. If f has a unique sub gradient at x, then f is differentiable at x and $p = \nabla f(x)$, and conversely. But the subgradient need not be unique: the set $\partial f(x)$ of sub gradients at x is the *sub differential* of f at x. The *domain* of f, dom f is the set of x such that $f(x) > -\infty$. The sub differential is non-empty for all x in its *relative interior*. It follows from the definition of concavity (and is proved in convex programming that the subdifferential correspondence is *monotonic:* if $p \in \partial f(x)$ and $q \in \partial f(y)$, *then* $(p - q) \cdot (x - y) \leq 0$. If f is convex, then the inequality is reversed, and $(p - q) \cdot (x - y) \geq 0$. Finally, suppose f is usc and concave. Then so is its conjugate f^*, and their sub differentials have an inverse relationship: $p \in \partial f(x)$ if and only if $x \in \partial f^*(p)$.

Cost, Profit and Production

In the theory of the firm, profit functions and cost functions are alternative ways of describing the firms' technology choices. A technology is described by a set of vectors F in $\mathbf{R}^{\mathbf{N}}$. Each vector $Z \in F$ is an input–output vector. We adopt the convention that negative coefficients correspond to input quantities and positive quantities correspond to outputs. Suppose that the first L goods are inputs and the last $M = N-L$ are outputs, so that $F \subset \mathrm{R}^{\mathrm{L}}_{-} \times \mathrm{R}^{\mathrm{M}}_{+}$. It is convenient to assume free disposal, so that if $(x, y) \in F$, and both $x' \leq x$ and $y' \leq y$ (more input and less output), then $(x', y') \in F$. Two important dual representations of the technology are the cost and profit functions. The profit function is $\pi(p, w) = \sup_{(x,y) \in F} p \cdot y + w \cdot x$ for $p \in$ and $w \in \mathbf{R}^{\mathrm{L}}$, which is the conjugate of the convex indicator function of F. The cost function too can be obtained through conjugacy. The set $F(y) = \{x : (x, y) \in F\}$ is the set of all input bundles that produce y. Then $C(y, w) = -\sup_{x \in F(y)} w \cdot x$, that is, $C(y, \cdot) = -\delta^{F(y)*}$.

Immediately the properties of the Fenchel transform imply that $\pi(p,w)$ is convex in its arguments and $C(y,w)$ is concave in w, the profit function is lsc and the cost function is usc. (This implies that both functions are continuous on the relative interior of their effective domains.) Cost and profit functions are also linear homogeneous. Doubling all prices doubles both costs and revenues. Cost is also monotonic. If $w_l' < w_l$ for every input l, then $C(y, w') \leq C(y, w)$ and if $w_l' < w_l$ for all l, then $C(y, w') < C(y, w)$.

The point of duality is that, if the technology is closed and convex, then cost profit functions each characterize the technology F. The conjugate duality theorem (see convex programming) implies that $\pi^*(x, y) = \delta^{F**}(x, y) = \delta^F(x, y)$, the convex indicator function of F:

$$\sup_{(p,w) \in \mathbf{R}^{\mathbf{N}}} p \cdot x + w \cdot y - \pi(p, w) = \begin{cases} 0 & \text{if} (x, y) \in F, \\ +\infty & \text{otherwise.} \end{cases}$$

If F *is* closed and convex, then each $F(y)$ is convex. If F is closed then $F(y)$ will also be closed. Then $\delta^{F(y)}$is concave and usc, so

$$\sup_{w \in \mathbf{R}^{\mathbf{L}}_{+}} w \cdot x + C(y, w) = \sup_{w \in \mathbf{R}^{\mathbf{L}}_{+}} w \cdot x \delta^{F(y)*} (w) = \delta^{F(y)}(x).$$

Hotelling's lemma is a famous result of duality theory. It says that the net supply function of good i is the derivative of the profit function with respect to the price of good i. The usual proof is via the envelope theorem: the marginal change in

profits from a change in price p is the quantity of good i times the change in the price plus the price of all goods times the changes in their respective quantities. But the quantity changes are second-order because the quantities solve the profit maximization first-order conditions, that price times the marginal change in quantities in technologically feasible directions is 0. Every advanced microeconomics text proves this. A result like this is true whenever the technology is convex, even if the technology is not smooth.

The convex version of Hotelling's lemma is a consequence of the inversion property of sub differentials for concave and convex f that $p \in \partial f(x)$ if and only if $x \in \partial f^*(p)$. See convex programming for a brief discussion.

Hotelling's lemma $(x,y) \in \partial \pi(p,w)$ if and only if (x, y) is profit-maximizing at prices (p, w).

Hotelling's lemma is quickly argued. If $(x, y) \in \partial\pi(p, w) = \partial\delta^{F*}(p, w)$, then $(p, w) \in \partial\delta^{F**}(x, y) = \partial\delta^{F}(x, y)$. Then $\delta^F(x, y) + (p, w) \cdot ((x', y') - (x, y)) \leq \delta^F(x', y')$ for all (x', y'). This implies that $x \in F$ and furthermore that $(p, w) \cdot ((x', y') - (x, y)) \leq 0$ for all $(x, y) \in F$ in other words, that (x, y) is profit-maximizing at prices (p,w). Conversely, suppose that (x,y) is profit maximizing at δ prices (p,w). Then (p,w) satisfies the sub gradient inequality of δ^F at (x,y), and so $(p, w) \in \partial\delta^F$. Consequently,

$$(x, y) \in \partial\delta^{F*}(p, w) \equiv \partial\pi(p, w).$$

The textbook treatment of duality observes that, if net supply is the first derivative of the profit function, then the own-price derivative of net supply must be the second own-partial derivative of profit with respect to price, and convexity of the profit function implies that this partial derivative should be positive, so net supply is increasing in price. The same fact follows in the convex framework from the monotonicity properties of the sub gradients. Suppose that $(w,$ p) and (w',p') are two price vectors, and suppose that (x, y) and (x', y') are two profit-maximizing production plans corresponding to the two price vectors. Then $(w - w', p - p')(x - x', y - y') \geq 0$. If the two price vectors are identical for all prices but, say, $p_k \neq p'_k$, then $(p_k - p'_k)(y_k - y'_k) \geq 0$, and net supply is non-decreasing in price. As with net supplies, some comparative statics of conditional factor demand with respect to input price changes follows from the monotonicity property of subgradients.

Another implication of profit function convexity and (twice continuous) differentiability is symmetry of the derivatives of net supply:

$$\frac{\partial y_k}{\partial p_1} = \frac{\partial^2 \pi}{\partial p_k \partial p_1} = \frac{\partial^2 \pi}{\partial p_1 \partial p_k} = \frac{\partial y_1.}{\partial p_k}$$

The convex analysis version of this is that for any finite sequences of goods i,. . .,*l*,

$$p_i \cdot \left(y_j - y_i\right) + p_j \cdot \left(y_k - y_j\right) + \cdots + p_l \cdot (y_i - y_l) \leq 0.$$

This requirement, which has a corresponding expression in terms of differences in prices, is called cyci/c monoion/c/iy. All subdifferential correspondences are cyclicly monotone. The connection with symmetry is not obvious, but it helps to know that Rockafellar (1974) leaves as an exercise (and so do we) that cyclic monotonicity is a property of a linear transformation corresponding to an $n \times n$ matrix M if and only if M is symmetric and positive semi-definite. Monotonicity is cyclic monotonicity for sequences of length 2.

The other famous result in duality theory for production is Shephard's lemma, which does for cost functions what Hotelling's lemma does for profit functions: conditional input demands are the derivatives of the cost functions. This is demonstrated in the same way, since the cost function and the indicator function for the set of inputs from which y is produceable are both convex and have closed hypographs.

Utility and Expenditure Functions

A quasi-concave utility function U defined on the commodity space $\mathbf{R}^{\mathbf{n}}_{+}$ has upper contours sets, the sets R_u of consumptions bundles which have utility at least u, which are convex. If u is usc, these sets are closed as well.

The *expenditure function* gives for each utility level u and price vector p the minimum cost of realizing utility u at prices $p : e\{p, u) = \inf (p \cdot x : u(x) \geq u\}$. If the infimum is actually realized at a consumption bundle x, then x is the *Hicksian* or *compensated real income demand.*

In terms of convex analysis, $e(p, u)$ is the conjugate of the concave indicator function $\varphi_u(x)$ of the set $R(u) = \{x : U(x) \geq u\}$, that is, $e(p, u) = \varphi_u^*(p)$. Thus $e(p, u)$ will be usc and concave in p for each u. The expenditure function is also linearly homogeneous in prices. If prices double, then the least cost of achieving u will double as well.

The duality of utility and expenditure functions is that each can be derived from the other; they are alternative characterizations of preference. Since the concave indicator function $\varphi_u(x)$ is closed and convex, $e(\cdot, u)^* = \varphi_u(x)$. For fixed u, the Fenchel transform of the expenditure function is the concave indicator function of $R(u)$; $\inf_p p \cdot x - e(p, u)$ is 0 if $U(x) \geq u$ and $-\infty$ otherwise. If $x \in R(u)$, then the cost of x at any price p can be no less than the minimum cost necessary to achieve utility u. The gap between the cost of x and the cost of utility level u is made by taking ever smaller prices, and so its minimum is 0. Suppose that x is not in $R(u)$. The separation theorem for convex sets says there is a price p such that $p \cdot x < \inf_{y \in R(u)} p \cdot y$; there is a price at which x is cheaper than the cost of u. Now, by taking ever larger multiples of p, the magnitude of the gap can be made arbitrarily large, and so the value of the conjugate is $-\infty$. Thus the conjugate is the concave indicator function of $R(u)$.

Among the most useful consequence of the duality between utility and expenditure functions is the relationship between derivatives of the expenditure function and the Hicksian, or compensated, demand. Hicksian demand. The compensated demand at prices p and utility u are those consumption bundles in $R(u)$ which minimize expenditure at prices p. This result is just Shephard's lemma for expenditure functions:

Hicks Compensated Demand Consumption bundle x is a Hicks compensated consumption bundle at prices p if and only if $x \in \partial_p e(p, u)$. Furthermore, if x is demanded at prices p and utility u, and y is demanded at prices q and the same utility u, then $(p - q) \cdot (x - y) \leq 0$.

The downward-sloping property just restates the monotonicity property of the subdifferential correspondence. For the special case of changes in a single price, the statement is that demand is non-increasing in its own price.

Equilibrium and Optimality

The equivalence between Pareto optima and competitive equilibria can also be viewed as an expression of duality. When preferences have concave utility representations, *quasi-equilibrium* emerges from Lagrangean duality. Quasi-equilibrium entails feasibility, profit maximization, and expenditure minimization rather than utility maximization. That is, each trader's consumption allocation is expenditure minimizing for the level of utility it achieves. The now traditional route of Arrow (1952) and Debreu (1951) to the Second Welfare Theorem first demonstrates that a Pareto-optimal allocation can be regarded as a quasi-equilibrium for an appropriate set of prices. Under some additional conditions, the quasi-equilibrium is in fact a competitive equilibrium, wherein utility maximization on an appropriate budget set replaces expenditure minimization. Our concern here is with the first step on this path.

Suppose that each of I individuals has preferences represented by a concave utility function on $\mathbf{R}_+^{\mathbf{N}}$, and that production is represented, as in section "Cost, Profit and Production", by a closed and convex set F of feasible production plans. Suppose that $0 \in F$ (it is possible to produce nothing) and that the aggregate endowment e is strictly positive. Assume, too, that there is free disposal in production. Every Pareto optimum is the maximum of a Bergson-Samuelson social welfare function of the form $\sum_i \lambda_i u_i$ defined on the set of all consumption allocations. An allocation is a vector (x, *y)* where $x \in \mathbf{R}_+^{\mathbf{NI}}$ is a consumption allocation, a consumption bundle for each individual, and y is a production plan. The allocation is *feasible* if $y \in F$ and $y + e - \sum_i x_i \geq 0$. A Lagrangean for this convex program is

$$= \begin{cases} \sum_i u_i(x_i) + p.\left(y + e - \sum_i x_i\right) & \text{if } x \in \mathbf{R}^{\mathbf{NI}}_+, y \in F \text{ and } p \in \mathrm{R}^{\mathbf{L}}_+, \\ +\infty & \text{if } x \in \mathbf{R}^{\mathbf{NI}}_+, y \in F \text{ and } p \notin \mathbf{R}^{\mathbf{L}}_+, \\ -\infty & \text{otherwise,} \end{cases}$$

where p is the vector of Lagrange multipliers for the L goods constraints.

The possibility of 0 production and the strict positivity of the aggregate endowment guarantee that the set of feasible solutions satisfies Slater's condition, and so a saddlepoint (x^*, y^*, p^*) exists; that is, $\sup_{x\,,\,y} L(x, y, p^*) \leq L(x^*, y^*, p^*) \leq L(x^*, y^*, p)$ for all $x \in \mathbf{R}^{\mathbf{NI}}$, $y \in F$ and $p \in \mathbf{R}^{\mathbf{L}}$. Then (x^*, y^*) is Pareto optimal and p^* solves the dual problem $\min_p \sup_{xy}$, $L(x, y, p)$. The interpretation of (x^*, y^*, p^*) as a quasi-equilibrium comes from examining the dual problem. The dual problem can be rewritten as

$$\begin{aligned} \inf_{p \in \mathbf{R}^{\mathbf{L}}_+} \quad \sup_{x \in \mathbf{R}^{\mathbf{NI}}_+, y \in F} \mathrm{L}(x, y, p) &= \inf_{p \in \mathbf{R}^{\mathbf{L}}_+} \sup_{x \in \mathbf{R}^{\mathbf{NI}}_+, y \in F} \sum_i u_i\,(x_i) + \mathrm{p} \cdot \left(y + e - \sum_i x_i\right) \\ &= \inf_{p \in \mathbf{R}^{\mathbf{L}}_+} \sum_i \sup_{x_i \in \mathbf{R}^{\mathbf{L}}_+} \{\lambda_i u_i(\mathrm{x}_i) - p \cdot x_i\} \sup_{+y \in F} p \cdot y. \end{aligned} \tag{1}$$

In the dual problem, the Lagrange multipliers can be thought of as goods prices. The Second Welfare Theorem interprets the optimal allocation as an equilibrium allocation using the Lagrange multipliers as equilibrium prices. To see this, look at the second line of (1). At prices p, a production plan is chosen from y to maximize profits $p \cdot y$, so the value of this term is $\pi(p)$. Each consumer is asked to solve

$$\begin{aligned} &\max_i \lambda_i u_i(x_0) - p \cdot x = -\min\ p \cdot x - \lambda_i u_i(x_i) \\ &\quad = \lambda_i u_i^* - \min\ p \cdot x - \lambda_i\left(u_i(x_i) - u_i^*\right) \end{aligned}$$

where $u_i^* = u_i(x_i^*)$. The term being minimized is the Lagrangean for the problem of expenditure minimization, and so x_i^* is the Hicksian demand for consumer i at prices p and utility level $u_i^* = u_i(x_i^*)$. Finally, the optimal allocation is feasible, and so (x^*, y^*, p^*) is a quasi-equilibrium.

Given the observation about expenditure minimization, the saddle value of the Lagrangean is

$$\sum_i \lambda_i u_i^* - e_i\left(p^*, u_i^*\right) + \pi(p^*)$$

The planner chooses prices to minimize net surplus, which is the sum of profits from production and the excess of total Bergson-Samuelson welfare less the cost of the consumption allocation.

Historical Notes

Duality ideas appeared very early in the marginal revolution. Antonelli, for instance, introduced the indirect utility function in 1886. The modern literature begins with Hotelling (1932), who provided us with Hotelling's lemma and cyclic monotonicity. Shephard (1953) was the first modern treatment of duality, making use of notions such as the support function and the separating hyperplane theorem.

The results on consumer and producer theory are surveyed more extensively in Diewert (1981), who also provides a guide to the early literature. In its focus on Fenchel duality, this review has not even touched on the duality between direct and indirect aggregators, such as utility and indirect utility, and topics that would naturally accompany this subject such as Roy's identity. Again, this is admirably surveyed in Diewert (1981).

See Also

- ▶ Convex Programming
- ▶ Convexity
- ▶ Duality
- ▶ Lagrange Multipliers

- ▶ Pareto Efficiency
- ▶ Quasi-Concavity

Bibliography

Arrow, K.J. 1952. An extension of the basic theorems of classical welfare economics. In *Proceedings of the second berkeley symposium on mathematical statistics and probability*, ed. J. Neyman. Berkeley: University of California Press.

Debreu, G. 1951. The coefficient of resource utilization. *Econometrica* 19: 273–292.

Diewert, W.E. 1981. The measurement of deadweight loss revisited. *Econometrica* 49: 1225–1244.

Hotelling, H. 1932. Edgeworth's taxation paradox and the nature of demand and supply. *Journal of Political Economy* 40: 577–616.

Rockafellar, R.T. 1970. *Convex analysis*. Princeton: Princeton University Press.

Rockafellar, R.T. 1974. *Conjugate duality and optimization*. Philadelphia: SIAM.

Shephard, R.W. 1953. *Cost and production functions*. Princeton: Princeton University Press.

Dühring, Eugen Karl (1833–1921)

Tom Bottomore

Keywords

Bernstein, E.; Dühring, E. K.; Engels, F.; Marx, K. H.; Positivism; Private property; Schumpeter, F. A

JEL Classifications

B31

Dühring was born on 12 January 1833 in Berlin and died on 21 September 1921 at Nowawes bei Potsdam. The son of a Prussian state official, Dühring studied law, philosophy and economics at the University of Berlin and practised law until blindness obliged him to abandon this career. He then became a *Privatdozent* at the University of Berlin, where he taught philosophy and economics from 1863 to 1877, and began to write voluminously on a wide range of subjects, from the natural sciences to philosophy, social theory and socialism, his aim being to construct a system of social reform based upon positive science. His system was expounded in a series of books on capital and labour (1865), the principles of political economy (1866), a critical history of philosophy (1869), a critical history of political economy and socialism (1871), and courses in political economy and philosophy (1873, 1875). Duhring was an adherent of positivism, concerned in his philosophical works to expound a 'strictly scientific world outlook', in opposition particularly to the Hegelian dialectic. His economic writings emphasize the role of political factors in the development of capitalism, and he argued that social injustice is not caused primarily by the economic system, but by social and political circumstances, the remedy being to control the misuse of private property and capital (not abolish them) through workers' organizations and state intervention.

Schumpeter (1954, pp. 509–10), praised Dühring's history of mechanics (1873), which was awarded an academic prize, suggested that he would retain a prominent place in the history of anti-metaphysical and positivist currents of thought, and noted that he made an important criticism of Marxist theory in his argument that political causes had played a major part in constituting the property relations of capitalist society. In other respects, however, Schumpeter considered that Dühring had made no significant contribution to economic theory.

Engels, in his well-known book (originally published as a series of articles), *Herr Eugen Dühring's Revolution in Science* [*Anti-Dühring*] (1877–8), which has done more than anything else to keep Dühring's name alive, took a much more critical view, deriding his work as a prime example of the 'higher nonsense' which infected German academic life. His philosophical views were dismissed by Engels as 'vulgar materialism' and compared unfavourably with the 'revolutionary side' of Hegel's dialectics; and in the chapter of *Anti-Dühring* devoted to the history of political economy (largely written by Marx, but not published in full until the

third edition of the book in 1894), Dühring was castigated for his superficiality and theoretical misconceptions. It was, however, the concern with Dühring's programme of social reform, and its possible baleful effect on the developing labour movement (Eduard Bernstein, for example, was initially impressed by Dühring's *Cursus* of 1873, though soon repelled by his anti-Semitism) that originally provoked Engels's articles, and was countered in the final section of the book (frequently reprinted later as a separate text under the title *Socialism, Utopian and Scientific*) by an exposition of Marxist socialism which became enormously influential.

It seems doubtful that Dühring occupies more than a minor place in the history of economic and social thought, except for this encounter with Marx and Engels, though Schumpeter (1954, p. 509) called him a 'significant thinker' and the entry in the *Encyclopedia of the Social Sciences* (1931, vol. 5, p. 273) described his writings as 'among the important intellectual achievements of the nineteenth century'.

Selected Works

1871. *Kritische Geschichte der Nationalökomie und des Sozialismus.* Berlin: T. Grieben.
1873. *Cursus der National- und Sozialökonomie einschliesslich der Hauptpunkte der Finanzpolitik*. Berlin: T. Grieben.
1875. *Cursus der Philosophie als streng wissenschaftlicher Weltanschauung und Lebensgestaltung.* Leipzig: E. Koschny.

Bibliography

Albrecht, G. 1927. *Eugen Dühring: ein Beitrag zur Geschichte der Sozialwissenschaften*. Jena: G. Fischer.
Engels, F. 1877–8. *Anti-Dühring. Herr Eugen Dühring's revolution in science*. Moscow: Progress Publishers, 1947.
Schumpeter, J.A. 1954. *A history of economic analysis*. London: Allen & Unwin.

Dummy Variables

Pietro Balestra

Abstract

The dummy-variable method is a useful device for introducing, into a regression analysis, information contained in qualitative or categorical variables, that is, in variables that are not conventionally measured on a numerical scale, such as race, sex, marital status, occupation, or level of education. It is a means for considering a specific scheme of parameter variation, in which the variability of the coefficients is linked to the causal effect of some precisely identified qualitative variable. But when the qualitative effects are generic, as in the cross-section time-series model, an interpretation in terms of random effects may seem more appealing.

Keywords

Covariance model; Cross-section time-series model; Dummy variables; Engel curve; Error component model; Qualitative variables; Random coefficient model

JEL Classifications

C1

In economics, as well as in other disciplines, qualitative factors often play an important role. For instance, the achievement of a student in school may be determined, among other factors, by his father's profession, which is a qualitative variable having as many attributes (characteristics) as there are professions. In medicine, to take another example, the response of a patient to a drug may be influenced by the patient's sex and the patient's smoking habits, which may be represented by two qualitative variables, each one having two attributes. The dummy-variable method is a simple

and useful device for introducing, into a regression analysis, information contained in qualitative or categorical variables; that is, in variables that are not conventionally measured on a numerical scale. Such qualitative variables may include race, sex, marital status, occupation, level of education, region, seasonal effects, and so on. In some applications, the dummy- variable procedure may also be fruitfully applied to a quantitative variable such as age, the influence of which is frequently U-shaped. A system of dummy variables defined by age classes conforms to any curvature and consequently may lead to more significant results.

The working of the dummy-variable method is best illustrated by an example. Suppose we wish to fit an Engel curve for travel expenditure, based on a sample of n individuals. For each individual i, we have quantitative information on his travel expenditures (y_i) and on his disposable income (x_i), both variables being expressed in logarithms. A natural specification of the Engel curve is:

$$y_i = a + bx_i + u_i$$

where a and b are unknown regression parameters and u_i is a non-observable random term. Under the usual classical assumptions (which we shall adopt throughout this presentation), ordinary least-squares produce the best estimates for a and b.

Suppose now that we have additional information concerning the education level of each individual in the sample (presence or absence of college education). If we believe that the education level affects the travel habits of individuals, we should explicitly account for such an effect in the regression equation. Here, the education level is a qualitative variable with two attributes: college education; no college education. To each attribute, we can associate a dummy variable which takes the following form:

$$d_{1i=}\begin{cases} 1 \text{ if college education} \\ 0 \text{ if no college education} \end{cases}$$

$$d_{2i=}\begin{cases} 1 \text{ if college education} \\ 0 \text{ if no college education} \end{cases}$$

Inserting these two dummy variables in the Engel curve, we obtain the following expanded regression:

Specification I

$$y_i = a_1 d_{1i} + a_2 d_{2i} + bxi + u_i$$

which may be estimated by ordinary least-squares. Alternatively, noting that $d_{1i} + d_{2i} = 1$ for all i, we can write:

Specification II

$$y_i = a_2 + (a_1 - a_2)d_{1i} + b_{xi} + u_i$$

which, again, may be estimated by ordinary least-squares.

It is easy to see how the procedure can be extended to take care of a finer classification of education levels. Suppose, for instance, that we actually have s education levels (s attributes). All we require is that the attributes be exhaustive and mutually exclusive. We then have the two following equivalent specifications:

Specification I

$$y_i = a_1 d_{1i} + a_2 d_{2i} + \ldots + a_s d_{si} + b_{xi} + u_i$$

Specification II

$$y_i = a_s + (a_1 - a_s)d_{1i} + \cdots + (a_{s-1} - a_s)d_{s-1,i} + bx_i + u_i.$$

Obviously, the two specifications produce the same results but give rise to different inputs. Specification I includes all the s dummy variables but no constant term. In this case, the coefficient of d_{ji} gives the specific effect of attribute j. Specification II includes $s - 1$ dummy variables and an overall constant term. The constant term represents the

specific effect of the omitted attribute, and the coefficients of the different d_{ji} represent the contrast (difference) of the effect of the jth attribute with respect to the effect of the omitted attribute. (Note that it is not possible to include all dummy variables plus an overall constant term, because of perfect collinearity.)

It is important to stress that by the introduction of additive dummy variables, it is implicitly assumed that the qualitative variable affects only the intercept but not the slope of the regression equation. In our example, the elasticity parameter, b, is the same for all individuals; only the intercepts differ from individual to individual depending on their education level. If we are interested in individual variation in slope, we can apply the same technique, as long as at least one explanatory variable has a constant coefficient over all individuals. Take the initial case of only two attributes. If the elasticity parameter varies according to the level of education, we have the following specification:

$$y_i = a_1 d_{1i} + a_2 d_{2i} + b_1 d_{1i} x_i + b_2 d_{2i} x_i + u_i.$$

Simple algebra shows that ordinary least-squares estimation of this model amounts to performing two separate regressions, one for each class of individuals. If, however, the model contained an additional explanatory variable, say z_i, with constant coefficient c, by simply adding the term cz_i to the above equation, we would simultaneously allow for variation in the intercept and variation in the slope (for x).

The dummy variable model also provides a conceptual framework for testing the significance of the qualitative variable in an easy way. Suppose we wish to test the hypothesis of no influence of the level of education on travel expenditures. The hypothesis is true if the s coefficients a_i are all equal; that is, if the s 1 differences $a_j - a_s$, $j = 1$ ... , $s - 1$, are all zero. The test therefore boils down to a simple test of significance of the s–1coefficients of the dummy variables in Specification II. If S=2, the t-test applied to the single coefficient of $d1i$ is appropriate. If $s > 2$, we may conveniently compute the following quantity:

$$\frac{(SS_c - SS)/(s-1)}{SS/(n-s-1)}$$

which is distributed as an F-variable with $s-1$ and $n-s-1$ degrees of freedom. In the above expression, SS is the sum of squared residuals for the model with the dummy variables (either Specification I or II), and SS_c is the sum of squared residuals for the model with no dummy variables but with an overall constant term.

In some economic applications the main parameter of interest is the slope parameter, the coefficients of the dummy variables being nuisance parameters. When, as in the present context, only one qualitative variable (with s attributes) appears in the regression equation, an easy computational device is available which eliminates the problem of estimating the coefficients of the dummy variables. To this end, it suffices to estimate, by ordinary least-squares, the simple regression equation:

Specification III

$$y_i^* = bx_i^* + u_i^*.$$

where the quantitative variables (both explained and explanatory) for each individual are expressed as deviations from the mean over all individuals possessing the same attribute. For the dichotomous case presented in the beginning, for an individual with college education, we subtract the mean over all individuals with college education and likewise for an individual with no college education. Note, however, that the true number of degrees of freedom is not $n-1$ but $n-1-s$. The same procedure also applies when the model contains other quantitative explanatory variables. The interested reader may consult Balestra (1982) for the conditions under which this simple transformation is valid in the context of generalized regression.

The case of multiple qualitative variables (of the explanatory type) can be handled in a similar fashion. However, some precaution

must be taken to avoid perfect collinearity of the dummy variables. The easiest and most informative way to do this is to include, in the regression equation, an overall constant term and to add for each qualitative variable as many dummy variables as there are attributes minus one. Take the case of our Engel curve and suppose that, in addition to the education level (only two levels for simplicity), the place of residence also plays a role. Let us distinguish two types of place of residence: urban and rural. Again, we associate to these two attributes two dummy variables, say e_{1i} and e_{2i}. A correct specification of the model which allows for both qualitative effects is:

$$y_i = a_1 + a_2 d_{1i} + a_3 e_{1i} + b_{x1} + u_i.$$

Given the individual's characteristics, the measure of the qualitative effects is straightforward, as shown in the following table:

	Urban	Rural
College education	$a_1 + a_2 + a_3$	$a_1 + a_2$
No college education	$a_1 + a_3$	a_1

The specification given above for the multiple qualitative variable model corresponds to Specification II of the single qualitative variable model. Unfortunately, when there are two or more qualitative variables there is no easy transformation analogous to the one incorporated in Specification III, except under certain extraordinary circumstances (Balestra 1982).

One such circumstance arises in connection with cross-section time-series models. Suppose that we have n individuals observed over t periods of time. If we believe in the presence of both an individual effect and a time effect, we may add to our model two sets of dummy variables, one corresponding to the individual effects and the other corresponding to the time effects. This is the so-called covariance model. The number of parameters to be estimated is possibly quite large when n or t or both are big. To avoid this, we may estimate a transformed model (with no dummies and no constant term) in which each quantitative variable (both explained and explanatory) for individual i and time period j is transformed by subtracting from it both the mean of the ith individual and the mean of the jth time period and by adding to it the overall mean. Note that, by this transformation, we lose $n+t-1$ degrees of freedom.

To conclude, the purpose of the preceding expository presentation has been to show that the dummy-variable method is a powerful and, at the same time, simple tool for the introduction of qualitative effects in regression analysis. It has found and will undoubtedly find numerous applications in empirical economic research.

Broadly speaking, it may be viewed as a means for considering a specific scheme of parameter variation, in which the variability of the coefficients is linked to the causal effect of some precisely identified qualitative variable. But it is not, by any means, the only scheme available. For instance, when the qualitative effects are generic, as in the cross-section time-series model, one may question the validity of representing such effects by fixed parameters. An interpretation in terms of random effects may seem more appealing. This type of consideration has led to the development of other schemes of parameter variation such as the error component model and the random coefficient model.

A final remark is in order. In the present discussion, qualitative variables of the explanatory type only have been considered. When the qualitative variable is the explained (or dependent) variable, the problem of these *limited* dependent variables is far more complex, both conceptually and computationally.

Bibliography

Balestra, P. 1982. Dummy variables in regression analysis. In *Advances in economic theory*, ed. Mauro Baranzini. Oxford: Blackwell.

Goldberger, A.S. 1960. *Econometric theory*, 218–227. New York: Wiley.

Maddala, G.S. 1977. *Econometrics*, chap. 9. New York: McGraw-Hill.

Suits, D.B. 1957. Use of dummy variables in regression equations. *Journal of the American Statistical Association* 52: 548–551.

Dumping

Wilfred J. Ethier

The term 'dumping' has been used for centuries in a general way to refer to export sales at a price low enough to cause significant harm to some interests in the importing country. Beginning early in this century, many countries instituted anti-dumping laws, and this has required a more precise definition of the term. The most common definition, both in the law and among professional economists, is export sales at a price below that at which similar goods are sold in the domestic market of the exporting country, taking into account differences in quality, attendant services and the like. However, an alternative definition, export sales at a price below the cost of production, is also incorporated into many of the laws, and this alternative has in recent years become of increasing practical importance.

Anti-dumping laws typically define the practice, prohibit it, provide for a penalty in cases where it nonetheless occurs, and establish an administrative procedure for determining in specific cases whether it has occurred and what penalty to impose. The penalty is usually in the form of an import levy related to the 'dumping margin', or difference between the export price and the source-country domestic price (or cost of production).

Such anti-dumping duties, though inherently at odds with most-favoured-nation treatment, are internationally accepted. The General Agreement on Tariffs and Trade (GATT) does not outlaw dumping, but it does countenance anti-dumping laws. The Tokyo Round of trade negotiations produced a code of conduct for antidumping legislation.

Numerous instances of alleged dumping have characterized recent tariff debates within and among the industrial countries. Recent changes in the administration of US anti-dumping and countervailing-duty statutes will likely further increase their use. More generally, the marked postwar reduction in tariffs on manufactured goods within the GATT framework on international responsibility, together with a secular convergence of cost structures in the industrial economies, prompt the conjecture that anti-dumping and anti-subsidy statutes will be a principal battle-ground for the 'new protectionism' concerning trade in manufactures among the developed economies. If so, the theory of dumping must become a major part of the positive theory of protection relevant to such trade.

The early literature generally defined dumping as price discrimination between national markets. This was the definition adhered to by Viner (1923) in his classic treatment and followed by most major authors (see Yntema 1928; Robinson 1933; Haberler 1937). Indeed, much of the early theory of price discrimination was developed in this context. Two problems arise when the phenomenon is viewed in this light. The first is why the firm is able to discriminate. This requires the firm to have some control over price; that is, imperfect competition is central. It also requires the firm to be able to segment markets on a national basis: tariffs or other trade barriers can serve this purpose. The second problem is why the export price should be lower than the domestic price rather than vice versa. One possible response is that such *reverse dumping* is indeed as common as dumping but is simply not a policy issue. For example, the sale of luxury German automobiles in the US at prices much higher than those in Germany brings forth not a whimper of an official threat from Washington, while the sales of low-priced European automobiles in the US was the occasion for a celebrated action some years ago. An alternative response is to hunt for circumstances that allow dumping to be more than a mere accident. One possibility is that the trade pattern is unidirectional and given by other considerations. Exporting firms thus compete only among themselves at home, but also with foreign firms in the export market. Thus, even if the *market* elasticity is the same in both countries, the elasticity facing each *firm* will be higher in the importing country, because more firms compete there. Thus, other things equal, exporters will charge lower prices abroad than at home (see Eichengreen and van der Ven 1984).

A second possibility involves transport costs between markets. Other things equal, such costs result in a firm having a smaller share, in equilibrium, in its export market than in its domestic

market. This again creates a presumption that the elasticity facing the firm is higher for foreign sales than for domestic sales. Furthermore, this reasoning does not require a unidirectional trade pattern; it is quite consistent with *reciprocal dumping,* or cross-hauling, with each country dumping in the other (see Brander and Krugman 1983).

Most of the formal theory of dumping essentially consists only of the theory of monopolistic price discrimination between two markets. But by contrast the 'sales at a price below cost of production' criterion has gradually become relatively more important in recent years, both in practice and in revisions of anti-dumping laws. Economists have long recognized that the two criteria are not inconsistent. A price-discriminating firm might well price its exports below average cost in a slump as long as export revenues at least cover the variable cost of producing those exports (*sporadic* or *cyclical* dumping). Or the firm might permanently sell its exports below average cost if those exports allow it to realize sufficient economies of scale. *Predatory* dumping, to drive rivals from the market, has long received much public attention, but economists have typically minimized its importance (see Viner 1923). How anti-dumping laws might in fact be applied in all these situations has, not surprisingly, concerned economists for years. More recently, attention has encompassed cases where export price does not cover even marginal cost. This might well occur *ex post* if the exporting firm must commit itself before demand conditions in the export market are fully known. Or the firm might do so deliberately if, instead of maximizing profit, it wishes to maximize sales subject to a profit constraint. More interesting, and closer to recent work in industrial organization, is the possibility that export sales, even at a low price, might make it easier for the firm to maintain excess capacity for the purpose of deterring entry by potential rivals (see Davies and McGuinness 1982).

Even though economists no longer confine themselves to price discrimination, dumping has been treated (aside from sporadic instances) usually either as profit (or sales) maximization by a discriminating monopolist or as an oligopolistic tactic to eliminate competition, to deter entry, or to enforce a cartel. Industries with dumping (or allegations of dumping) are most often characterized by large fixed costs, factor-market rigidities, susceptibility to demand fluctuations and downward price rigidity. Though by no means inconsistent with oligopolistic rivalry in segmented markets, these characteristics involve much more. Thus our theory largely excludes those considerations fundamental to most contemporary problems: imperfectly adjusting factor markets in the presence of changing conditions of product demand. The earlier literature did consider the related problem of dumping to stabilize production over the business cycle. Viner (1923, p. 28), although conjecturing that 'it is probable that this is the most prevalent form of dumping', basically treated it as only a distinct motive. The interdependence between such dumping and factor-market equilibrium within the relevant industries of trading countries is critical. The ability to dump aboard during periods of slack demand allows a firm to offer its workers greater job security, and thereby allows that firm to pay lower wages over time than it would have to do if it did not offer that security. Thus the possibility of dumping influences the normal trading equilibrium. Furthermore, that equilibrium clearly must be sensitive to employment practices in different countries and to the relations between the business cycles of the various countries (see Ethier 1982).

In addition to the deficient treatment of the fundamental issues involving factor markets, we have no theory of anti-dumping laws. The basic theory of tariffs applies of course to anti-dumping duties put in place and left there. But an anti-dumping *law* is a threat to impose (with considerably less than certainty) a duty in response to certain behaviour on the part of exporters and so will influence that behaviour even if not actually imposed. International trade theory has not yet addressed this issue. This is just one aspect, though an especially important one, of our prominent lack of a contemporary theory of protection. These two omissions (factor markets and anti-dumping laws) are serious, but the first is being addressed, and we have the technical equipment to deal adequately with both of them, so one would expect the deficiencies to be mended soon.

See Also

- ▶ Discriminating Monopoly
- ▶ Price Discrimination
- ▶ Quotas and Tariffs

Bibliography

Brander, J., P. Krugman, and 3/4. 1983. A 'reciprocal dumping' model of international trade. *Journal of International Economics* 15(3/4): 313–321.

Davies, S.W., and A.J. McGuinness. 1982. Dumping at less than marginal cost. *Journal of International Economics* 12(1/2): 169–182.

Eichengreen, B., and H. van der Ven. 1984. US antidumping policies: The case of steel. In *The structure and evolution of recent US trade policy*, ed. R.E. Baldwin and A.O. Kreuger. Chicago: University of Chicago Press, for the National Bureau of Economic Research.

Ethier, W.J. 1982. Dumping. *Journal of Political Economy* 90(3): 487–506.

Haberler, G. 1937. *The theory of international trade with its applications to commercial policy*. New York: Macmillan.

Robinson, J. 1933. *The economics of imperfect competition*. London: Macmillan.

Viner, J. 1923. *Dumping: A problem in international trade*. Chicago: University of Chicago Press.

Viner, J. 1931. Dumping. In *Encyclopedia of the social sciences*, ed. E.R.A. Seligman and A. Johnson. New York: Macmillan.

Yntema, T.O. 1928. The influence of dumping on monopoly price. *Journal of Political Economy* 36: 686–698.

Dunbar, Charles Franklin (1830–1900)

A. W. Coats

Dunbar's career illustrates the narrow gap between practical and academic economics in his lifetime, for he demonstrated that scholarly instincts, common sense and knowledge of current affairs could overcome deficiencies in formal academic training. Exactly 20 years after graduating from Harvard in 1851 he returned as Professor of Political Economy, having previously worked in a mercantile business, qualified and practised as a lawyer, and written articles on political questions for the *Boston Daily Advertiser*, of which he was sole proprietor and editor from 1865 to 1869. After President Eliot's invitation to Harvard, Dunbar spent two years travelling and studying in Europe, and subsequently served as Head of the Department of Political Economy for nearly 30 years, Dean of the College (1876–82) and the Faculty of Arts and Sciences (1890–95), and as editor from 1886 to 1896 of *The Quarterly Journal of Economics*, the first English-language scholarly periodical in the subject. His election as second President of the American Economic Association in 1892, following Francis A. Walker, testifies to his standing in the emerging economics profession. While he published comparatively little, his works on currency, finance and banking were widely respected, and his essays on the history, condition and methods of economics were wise and balanced at a time of intense controversy.

Selected Works

Dunbar, Charles Franklin. 1891a. *Chapters on the theory and history of banking*. New York: G.P. Putnam's Sons. 2nd enlarged. edn, ed. O.M.W. Sprague, 1901; 3rd enlarged edn, 1917.

Dunbar, Charles Franklin. 1891b. *Laws of the United States relating to currency, finance, and banking from 1789 to 1891*. Boston: Ginn & Co. Revised. edn, 1897.

Sprague, O.M.W. (ed.). 1904. *Economic essays*. New York: Macmillan Co.

Dunlop, John Thomas (1914–2003)

Richard B. Freeman

Keywords

Business cycles; Dispute resolution; Dunlop, J.; Industrial relations; Labour's share of income; Marginal productivity theory; Mediation; Wage contours; Wage determination

JEL Classifications
B31

John Dunlop was an extraordinary labour economist, Professor and Dean of the Faculty at Harvard University, Secretary of Labor of the United States, and mentor to students and practitioners in the world of labour. He was extraordinary because he was more than an economist and because he was driven by a moral vision of what economists and academics should do to make the world better. Labour economists and policymakers paid close attention to Dunlop's thoughts because he combined academic research with unparalleled practical experience in solving problems and building institutions. His academic writings, which include several classic articles as well as major books, reflect Dunlop's participation in events and direct observations of social behaviour.

Dunlop first attracted academic attention with his 1938 *Economic Journal* article on the movement of real and money wages over the business cycle, which forced Keynes to admit that the *General Theory* was wrong on this issue: real wages fall in recessions not in booms, contrary to simple marginal productivity analysis. Quite an achievement for a 24-year-old economist. Dunlop followed this with *Wage Determination Under Trade Unions* (1944), in which he modelled unions as optimizing organizations; with analyses of the cyclic variation of labour's share, with the concept of 'wage contours' that captured the notion that product markets influenced wages, and with numerous analysis of wage determination, labour relations, mediation and dispute resolution. Dunlop's book *Industrial Relations Systems* (1958) sought to develop a broader perspective on how labour relations fit into economics.

In the 1980s, concerned that labour economists were limited in their conceptual vision by narrow optimizing models and in their empirical analysis by extant government data-sets, Dunlop carped at them for failing to see what he could see in the labour market. Dunlop saw the labour market as pre-eminently a social institution to resolve labour problems, which should be analysed as such rather than as a bourse. His mode of analysis was that of a naturalist, who looks at the world with his own eyes and experience, with direct knowledge of the institutions and practitioners, without trying to force observation into a narrow conceptual framework.

Dunlop's career spanned a wide variety of activities. Earning his AB (1935) and Ph.D. (1939) from Berkeley, he rose to become professor of economics at Harvard and Dean of the University (1970–1973), when he helped stabilize the university during a period of student disorders, and Lamont University Professor (1970–2003). He worked for the National War Labor Board (1943–1954); served as member or chair on various national panels with responsibility for resolving labour disputes; led labour-management committees in areas ranging from missile sites to apparel, the public sector, and health; served as Director of the Cost of Living Council (1973–1974), and as Secretary of Labor of the United States (1975–1976). From 1993 to 1994 he chaired the Commission on the Future of Worker–Management Relations, popularly known as the Dunlop Commission, which was given the charge 'to recommend ways to improve labor–management cooperation and productivity'. The politics and economics of the time were not right, however, for bringing management and labour to a consensus on modernizing labour relations, so that much of the Commission's recommendations went unheeded.

Dunlop approached his work –advising presidents and cabinet officials and telling academics about the real world and practitioners about academic theory – with one goal: to help solve problems. The moral principle that guided him – that academics should use their knowledge and skill to help solve problems faced by real people, by workers and firms, and governments – represents social science at its best.

Selected Works

1938. The movement of real and money wage rates. *Economic Journal* 48: 413–434.

1944. *Wage determination under trade unions.* New York: Macmillan.

1948a. Productivity and the wage structure. In *Income, employment and public policy: Essays in honor of Alvin H. Hansen.* New York: W.W. Norton.

1948b. The development of labor organizations: A theoretical framework. In *Insights into labor issues*, ed. R. Lester, and J. Shister. New York: Macmillan.

1957. The task of contemporary theory. In *New concepts in wage determination,* ed. G. Taylor, and F. Pierson. New York: McGraw-Hill.

1958. *Industrial relation systems*. New York: Henry Holt & Co.

1984. *Dispute resolution.* Dover: Auburn House Publishing Co.

Bibliography

Segal, M. 1986. Post-institutionalism in labor economics: The forties and fifties revisited. *Industrial and Labor Relations Review* 39: 388–403.

Dunoyer, Barthélémy Charles Pierre Joseph (1786–1862)

R. F. Hébert

French economist and publicist, born at Carennac (southwest France) on 20 May 1786, died at Paris on 4 December 1862. Dunoyer studied law in Paris, where he befriended Charles Comte, who shared his liberalism and joined him in founding and editing *Le Censeur,* a journal of institutional and legal reform. The journal was discontinued in 1820 due to increasingly repressive press laws. Subsequently Comte went to Switzerland, whereas Dunoyer stayed in Paris and devoted himself exclusively to economics. He became professor of political economy at the Athenée, later publishing his lectures under the title *L'industrie et la morale considerées dans leurs rapports avec la liberté.* In 1832 he was elected to the French Institute, and in 1845 he became president of the Société d'Economie Politique. He spent two decades in public life, entering the government in 1830 under the bourgeois monarchy of Louis-Philippe, and withdrawing after the coup d'état of 1851. His articles appeared frequently in the *Journal d'Economie Politique* and in other French journals.

Dunoyer added nothing new to economic theory but he was part of a group of French radicals who helped create a powerful means of social analysis by fusing liberal historicopolitical thought with the economic orthodoxy of J.B. Say. Inspired by Turgot and Condorcet, Dunoyer and his cohorts advanced an evolutionary theory of history that identified progress with the gradual disintegration of authority and its replacement by the quiescent, voluntary relationships of the marketplace. These writers anticipated the flowering of *industriélisme,* of a kind apart from Saint-Simon's insofar as it envisioned government as a mere subsidiary institution, charged mainly with the functions of preserving order and ministering to the needs of production. Having thus nested their basic anarchism in an evolutionary concept of social development, the group fit surprisingly well into the republican and constitutional framework of the July Monarchy.

The product of Dunoyer's mature thought was his three-volume work, *De la liberté du travail.* Despite its brilliance and good sense, it is more a history of civilization than a sustained economic treatise. Dunoyer anticipated Herbert Spencer by developing the idea that society is an organic composition of institutions and individuals with specific functions. Although he regarded government's role as minimal, the presence of government professionals finds justification in his conception of 'immaterial wealth' (i.e. services). Although he followed classical economics in most things, Dunoyer rejected Say's Law, holding that a general glut could arise due to the ignorance or error of entrepreneurs, or to the unequal distribution of wealth. Unlike Sismondi, however, whose ideas had a certain allure, he spurned government palliatives, trusting the growth of industry to gradually reduce entrepreneurial error and to smooth

the distribution of income. Dunoyer also denied the classical theory of rent, because he admitted only one factor of production, labour. On population matters he was an unregenerate Malthusian, which tempered his basic faith in progress with a hint of pessimism.

Selected Works

1825. *L'industrie et la morale considerée dans leurs rapports avec la liberté*. Paris: A. Sautelet. Revised, enlarged and reprinted as Nouveau traité d'économie sociale, 2 vols. Paris: A. Sautelet, 1830.
1840. *Esprit et méthodes comparés de l'Angleterre et de la France dans les entreprises de travaux publics et en particulier des chemins de fer*. Paris: Carilian-Goeury et V. Dalmont.
1845. *De la liberté du travail*, 3 vols. Paris: Guillaumin.

References

Allix, E. 1911. La déformation de l'économie politique libérale aprés J.B. Say: Charles Dunoyer. *Revue d'Histoire Economique et Sociale* 4: 115–147.
Villey-Desmeserts, E.L. 1899. *L'oeuvreéconomique de Charles Dunoyer*. Paris: L. Larose.
Weinburg, M. 1978. The social analysis of three early 19th-century French Liberals: Say. *Comte and Dunoyer. Journal of Libertarian Studies* 2: 45–63.

Duopoly

James W. Friedman

A duopoly is a market in which two firms sell a product to a large number of consumers. Each consumer is too small to affect the market price for the product: that is, on the buyers' side, the market is competitive. Therefore, in its essence duopoly is a two player variable sum game. Each of the two duopolists is a rational decision-maker whose actions will affect both himself and his rival. Although the interests of the duopolists are intertwined, they are not wholly coincident nor wholly in conflict. In contrast to the agents in competitive markets, the duopolists must each concern themselves with what the other duopolist is likely to do.

The situation facing the duopolists is non-cooperative in the sense that they are barred from making binding agreements with one another. The relevance of this depends crucially on whether the model is a static market (i.e. a one-time-only, or one-shot market) or a market consisting of many time periods.

The first study of duopoly is the great contribution of Cournot (1838) in which the decision problem of the firms is posed for a homogeneous products market in a static setting. The equilibrium concept proposed by Cournot, variously called the *Cournot equilibrium* or the *Cournot–Nash equilibrium,* has become a cornerstone of non-cooperative game theory. To sketch his model let x and y be the output levels of firms A and B, let $f(x + y)$ be the inverse demand function for the market, let $C(x)$ and $\Gamma(y)$ be the two firms' total cost functions, and let their respective profit functions be $\pi^A = xf(x + y) - C(x)$ and $\pi^B = yf(x + y) - \Gamma(y)$.

Cournot proposed as an equilibrium a pair of output levels (x, y) such that neither firm could have obtained higher profit by having chosen some other output. Thus π^A is maximized with respect to x (with y given), while, simultaneously, π^B is maximized with respect to y (with x given). If $x^c > 0$ and $y^c > 0$ the Cournot equilibrium is a solution to the simultaneous equations

$$\frac{\partial \pi^A}{\partial x} = f(x+y) + xf'(x+y) - C'(x) = 0$$

$$\frac{\partial \pi^B}{\partial y} = f(x+y) + yf'(x+y) - \Gamma'(x) = 0$$

The Cournot equilibrium defines consistency conditions. If firm A contemplates (x^c, y^c) as an outcome, and believes firm B is contemplating the same output pair, then firm A will see (a) that it cannot do better than to choose x^c (given the

expectation that firm B will choose y^c) and (b) that, should firm B go through the same thought process, it will reach a parallel conclusion.

To translate this model into the language of game theory, player A chooses a *strategy* x from the set of all allowed output levels, say all $x \geq 0$. This set, $[0, \infty]$, is called the *strategy space* or *strategy set* of the player. Similarly for player B. The players' *payoff functions* are their respective profit functions. Thus the payoff function of a player gives his payoff as a function of the strategies of all players in the game. At a non-cooperative equilibrium (see Nash 1951; Owen 1968, or Friedman 1986) no player could obtain a higher payoff through the use of a different strategy, given the strategies of the other players. Note, finally, that the actual behaviour of one duopolist cannot affect the actual behaviour of the other in this static setting, because they choose their output levels simultaneously. They do take one another into account in making decisions by analysing the game using *both* payoff functions.

Cournot's contribution went largely unnoticed for nearly half a century, after which it was scathingly reviewed by Bertrand (1883). Bertrand berates Cournot on two grounds. First he says that the firms will collude to achieve monopoly-like profits. This possibility is acknowledged by Cournot who made a conscious choice to explore behaviour in the absence of collusion. Bertrand's point was echoed later by Chamberlin (1933), although neither of them showed how the duopolists could be expected to maintain a collusive agreement nor did they solve the problem of the distribution of profits between the firms. These issues are addressed below in connection with recent developments.

Bertrand's second criticism is that price, not output, should be the firm's decision variable. Then, using Cournot's *mineral spring* example in which $C(x) = \Gamma\ (y) = 0$, he sketches the Bertrand equilibrium, arguing that consumers will buy from the firm charging the lower price, and showing that the only prices that can be in equilibrium are zero for both firms. Bertrand's equilibrium concept is precisely that of Cournot, transferred to the price choosing variant of Cournot's model. Bertrand's analysis was taken up, elaborated and extended by Edgeworth (1897). He supposed that the firms have production capacity limits, each of which is less than the market demand at zero price. Consequently no pair of prices is an equilibrium.

While the logic of Bertrand and Edgeworth is correct, the economic relevance is dubious. Real world firms choose both prices and output levels; however, the discontinuity of one firm's sales with respect to another firm's decision variable is *not* an obvious feature of economic life. Consequently, the Cournot formulation seems preferable. A way to reconcile price choosing firms with an absence of demand discontinuities is via differentiated products models.

Edgeworth and many of his contemporaries thought there was no worthwhile content in Cournot's duopoly theory. Edgeworth (1925, p. 111), writing forty years after Bertrand, said 'Now the demolition of Cournot's theory is generally accepted. Professor Amoroso is singular in his fidelity to Cournot'. Amoroso's good judgement was shared by Wicksell (1925). It is now generally accepted that Cournot was the first to perceive clearly and enunciate the game theoretic concept of *non-cooperative equilibrium,* which received a general statement from Nash (1951) and is the cornerstone of one of the main parts of game theory.

The next influential innovation is due to Bowley (1924) who invented the conjectural variation (which later received this name from Frisch 1933). He wrote the two firms' first order conditions for equilibrium as $\partial\pi^A/\partial x + (\partial\pi^A/\partial y)\ (\partial y/\partial x) = 0$ for firm A and $\partial\pi^B/\partial y + (\partial\pi^B/\partial x)\ (\partial x/\partial y) = 0$ for firm B. The $\partial y/\partial x$ in firm A's condition indicates the way that A thinks B's output choice will vary according to the way that A varies his own output choice. A parallel meaning attaches to $\partial x/\partial y$ in B's first order condition. The presence of these conjectural variation terms is indefensible in a static model, but it shows the underlying concern that writers had with dynamic models, while, at the same time, limiting their formal analysis to static models. Given that the model is static with the two firms simultaneously selecting outputs, *and doing so only once,* there can be no conjectural variation. B's output choice will depend on what B expects A to do, but that expectation will

not vary as A changes his mind about what output to select. B's expectation depends on B's thought processes and the information B has about the structure of the model, and does not depend on A's actual thought processes.

Dynamic elements of reaction of one firm to the choice of another go back to Cournot who performed a 'stability' analysis. He solved $\partial\pi^A/\partial x = 0$ to obtain $x = \upsilon\ (y)$ and $\partial\pi^B/\partial y = 0$ to obtain $y = w(x)$. He looked for conditions under which, starting from an arbitrary x^0, the sequence (x_n, y^{n+1}) for $n = 0, 2, 4, \ldots$ would converge to (x^c, y^c), the Cournot equilibrium. Bertrand and Edgeworth also wrote of actions and reactions, and Bowley introduces a new reactive element with his conjectural variation terms. Later Stackelberg (1934) posed the leader–follower duopoly in which one firm, say A, chooses x and, after that choice is communicated to B, y is chosen. B will always choose y according to $y = w(x)$ and this is known to A who maximizes $\pi^A = xf[x + w(x)] - C(x)$ with respect to x. Note that a conjectural variation term for A makes a legitimate appearance because B's decision is, in fact, a function of A's choice. Wicksell (1925) and Bowley (1928) anticipate Stackelberg's leader–follower equilibrium in their discussions of bilateral monopoly. All of these treatments strongly suggest an explicitly multiperiod formulation under which each firm maximizes a discounted profit stream and behaves according to a *reaction function* under which a firm's output choice in time t is selected as a function of the other firm's output choice in time $t - 1$. The last twenty years have seen such analysis, beginning with Friedman (1968).

The next major step in duopoly was the recognition that, in many industries, the firms sell very similar, non-identical, products. In such a market, it is equally easy to represent the firms as price choosers or as quantity choosers. In either case, equilibrium can readily involve the firms selling at different prices. The pioneers here are Hotelling (1929) and Chamberlin (1933). To sketch a differentiated products duopoly, let the firm's prices be p and r, and let their demand functions be $x = \phi(p, r)$ and $y = \psi(p, r)$, respectively. The two firm's are assumed to produce gross (but imperfect) substitutes, so $\phi_r(p, r) > 0$ and $\psi_p(p, r) > 0$, but the own-price derivatives (ϕ_p and ψ_r) are negative and both firms' total revenues are bounded. Profit functions are $\pi^A = p\phi\ (p, r) - C[\phi\ (p, r)]$ and $\pi^B = r\psi\ (p, r) - \Gamma\ [\psi\ (p, r)]$. A non-cooperative (Cournot–Nash) equilibrium occurs at a price pair (p^c, r^c) for which π^A is maximized with respect to p (given $r = r^c$) and π^B is maximized with respect to r (given $p = p^c$).

Many writers have maintained that the Cournot equilibrium should not be expected to occur in practice because it does not lie on the firms' profit possibility frontier. In addition to Bertrand and Chamberlin there is a famous passage in Smith (1776) maintaining that people in the same line of business will attempt to collude whenever they get together. In response to such observations several points can be made. Smith's passage is a comment in passing that is not made within an analytical framework, so it cannot be closely judged. Bertrand and Chamberlin are discussing specific models within which their remarks do not hold up well, because the consistency condition embodied in the Cournot equilibrium is quite compelling and would be violated by collusive behaviour. Any agreement between the two firms–in a static setting where binding agreements cannot be made– will break down because at least one firm will note that, given the agreed decision for the rival, it can do better by violating its agreement. But both firms can perceive the incentives of either one of them, thus the only acceptable agreement in such circumstances is for a price pair (or output pair, if the firms are output choosers) such that, given the price of its rival, neither firm can gain by deviating from its agreement. Such a *self-enforcing agreement* is merely a non-cooperative equilibrium. We are back at Cournot.

However, this is far from the last word on collusion in the absence of legally binding agreements. Bertrand, Chamberlin, and others who have made, or agreed with, their assertion probably are motivated by a belief that voluntary collusion sometimes occurs in actual markets. They may be correct in their empirical observation; however, it remains true that voluntary collusion is not convincing in the traditional one-shot

models. Therefore, the clear suggestion is that one-shot models are simply inadequate for analysing voluntary collusion. Suppose, then, that the model is changed to have an infinite horizon with each firm having a discount parameter of α Then, letting t denote time, player A seeks to maximize

$$\sum_{t=0}^{\infty} \alpha^t \pi^A = \sum_{t=0}^{\infty} \alpha^t [p_t \phi(p_t, r_t) - C[\phi(p_t, r_t)]$$

and the objective function of player B is

$$\sum_{t=0}^{\infty} \alpha^t \pi^A = \sum_{t=0}^{\infty} \alpha^t [r_t \psi(p_t, r_t) - \Gamma[\psi(p_t, r_t)]$$

Strategy becomes much more complex than in the static model, because there will be an infinite succession of price choices by each firm and, prior to making a price choice any time after $t = 0$, the firm will know what past prices have been selected by its rival. For each t, a firm can choose its price according to a function (i.e. a rule) that depends on *all* past price choices of both of them. The rule for one period can be different from the rule for another. A strategy for a firm is a collection of such rules, one for each period t.

In this model it may be possible to find a non-cooperative equilibrium that yields an outcome on the profit possibility frontier. Such an equilibrium is based on three critical prices for each firm. First there is (p^c, r^c), the Cournot price pair. Second there is (p^*, r^*), chosen so that profits at (p^*, r^*) are on the profit possibility frontier and are higher for each firm than at (p^c, r^c). Third define p' as the price for A that maximizes $\pi^A = p\phi\ (p, r^*) - C[\phi\ (p, r^*)]$, and define r' in a parallel way for B. Now consider the following strategy for firm A: $p_0 = p^*$, $p_t = p^*$ for $t > 0$ if $(p_k, r_k) = (p^*, r^*)$ for $k = 0, \ldots, t-1$, and $p_t = p^c$ otherwise. Imagine a parallel strategy for B. These strategies amount to a firm saying 'I will begin by cooperating and will continue to cooperate as long as we both have cooperated in the past. If ever a lapse from cooperation occurs, I will revert to static Cournot behaviour'.

Whether this pair of strategies is a non-cooperative equilibrium depends on the sizes of α and the profits at (p^*, r^*), (p', r^*), (p^*, r'), and (p^c, r^c). A's choice boils down to comparing (i) receiving the profit associated with (p^*, r^*) in all periods or (ii) obtaining the larger profit associated with (p', r^*) for just one period and the reduced profit associated with (p^c, r^c) in all subsequent periods. If α is near enough to one, both firms will prefer alternative (i). Thus both firms can be better off following the 'cooperative' strategy. Note, however, that this cooperative outcome is the result of following non-cooperative equilibrium strategies. The strategy pair is chosen so that no single firm can increase its payoff by altering its strategy, given the strategy of the other firm. The strategies are designed so that deviating from cooperative behaviour is followed by punishment, and the punishment is carefully crafted so that it will be in the interests of all players to carry it out when the strategies call for it. This latter property, that the threats of punishment are credible because they are incentive compatible, is called *subgame perfection*. On the concept of subgame perfect non-cooperative equilibria, see Selten (1975) or Friedman (1986).

The work of Hotelling and Chamberlin raises an important issue that has received some recent attention: firms not only choose prices (or output levels), they decide on the design of their products. In deciding on how to design a product, the firm needs to know how design is related to cost of production and how it is related to consumers' tastes. The latter has been modelled by Lancaster (1979) in terms of inherent charactistics. The underlying notion is that consumers value certain attributes of goods that are analogous to the nutrients in foods. A particular product (e.g. a chair of a given design) is a specific bundle of characteristics. The product of a rival seller is a somewhat different bundle of characteristics. A difficulty with this approach in the most general form that Lancaster discusses is that it is difficult to define these characteristics. Less abstract versions are used in oligopoly models where, following Hotelling, physical location, is used as the only characteristic chosen by firms. Any single measurable attribute, such as sweetness of a bottled drink, can also be used.

Other topics treated in the duopoly theory literature include capital stock decisions, advertising, and entry. They can be found in Friedman (1983), along with a fuller account of the topics sketched above.

See Also

- ▶ Bertrand, Joseph Louis François (1822–1900)
- ▶ Cournot, Antoine Augustin (1801–1877)
- ▶ Nash Equilibrium
- ▶ Oligopoly
- ▶ Strategic Behaviour and Market Structure

Bibliography

Bertrand, J. 1883. Review of Cournot 1838. *Journal des Savants* 499–508.

Bowley, A. 1924. *The mathematical groundwork of economics*. New York: Kelley, 1965.

Bowley, A. 1928. Bilateral monopoly. *Economic Journal* 38: 651–659.

Chamberlin, E. 1933. *The theory of monopolistic competition*, 7th ed. Cambridge: Harvard, 1956.

Cournot, A. 1838. *Recherches sur les principes mathématiques de la théorie des richesses.* Trans. N.T. Bacon. New York: Macmillan, 1927.

Edgeworth, F. 1897. The pure theory of monopoly. *Papers Relating to Political Economy* I: 111–142.

Edgeworth, F. 1925. *Papers relating to political economy.* New York: Burt Franklin, 1970.

Friedman, J. 1968. Reaction functions and the theory of duopoly. *Review of Economic Studies* 35: 257–272.

Friedman, J. 1983. *Oligopoly theory.* Cambridge: Cambridge University Press.

Friedman, J. 1986. *Game theory with applications to economics*. New York: Oxford University Press.

Frisch, R. 1933. Monopole – polypole – la notion de force dans l'économie. *Festschrift til Harald Westergaard.* Supplement to *Nationalekonomisk Tidsskrift.*

Hotelling, H. 1929. Stability in competition. *Economic Journal* 39: 41–57.

Lancaster, K. 1979. *Variety, equity, and efficiency.* New York: Columbia University Press.

Nash, J. 1951. Noncooperative games. *Annals of Mathematics* 45: 286–295.

Owen, G. 1968. *Game theory*, 2nd ed. New York: Academic, 1982.

Selten, R. 1975. Reexamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Games Theory* 4: 25–55.

Smith, A. 1776. In *An inquiry into the nature and causes of the wealth of nations*, ed. R.H. Campbell, A.S. Skinner, and W.M. Todd. Oxford: Clarendon Press, 1976.

von Stackelberg, H. 1934. *Marktform und Gleichgewicht*. Vienna: Julius Springer.

Wicksell, K. 1925. Mathematical economics. In K. Wicksell, *Selected papers on economic theory,* ed. Erik Lindahl. Cambridge, MA: Harvard University Press, 1958.

Dupuit, Arsene-Jules-Emile Juvenal (1804–1866)

Robert B. Ekelund Jr.

Keywords

Antitrust enforcement; Consumer surplus; Cost-benefit analysis; Deadweight loss; Diminishing marginal utility; Dupuit, A.-J.-E. J.; Edgeworth, F.Y.; Hotelling, H.; Law of demand; Marginal cost pricing; Menger, C.; Monopoly profit maximization; Price discrimination; Public works

JEL Classifications

B31

French engineer and economic theorist, born at Fossano, Piedmont, Italy on 18 May 1804, when this region was part of the French empire; died 5 September 1866 in Paris. After his parents returned to Paris in 1814, Dupuit continued his education in the secondary schools at Versailles, at Louis-le-Grand and at Saint-Louis, where he finished brilliantly by winning a physics prize in a large group of competitors. Accepted to the Ecole des Ponts et Chaussées in 1824, Dupuit soon distinguished himself as an engineer and, in 1827, was put in charge of an engineering district in the department of Sarthe, where he concentrated on roadway and navigation work. Dupuit's numerous and trenchant engineering studies on such topics as friction and highway deterioration, floods and hydraulics, and municipal water systems made him one of the most creative civil engineers of his day. Decorated for such contributions by the Legion of Honour in 1843, Dupuit ultimately became director-chief engineer in

Paris in 1850 and Inspector-General of the Corps of Civil Engineers in 1855.

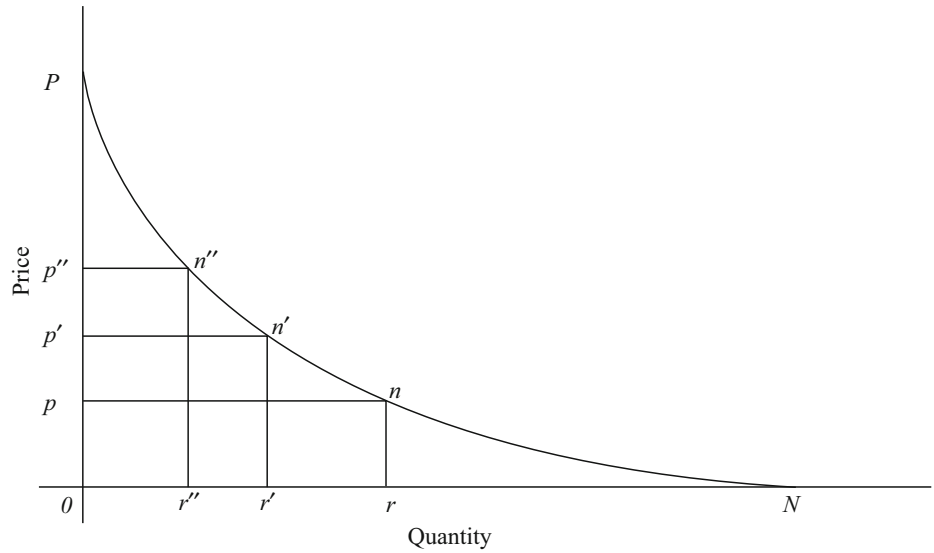
No less profound were Dupuit's contributions to general economic analysis and to the economic evaluation of public works (cost-benefit analysis). In fact, Dupuit was the most illustrious contributor in the long French tradition of study, teaching and writing on economic topics at the Ecole des Ponts et Chaussées, whose professors and students included Isnard, Henri Navier, Charles Minard, Emile Cheysson and Charles Ellet.

Led by a desire to evaluate the economic or *net* benefits of public provision, Dupuit directed his considerable analytical gifts to the utility foundation of demand and to its relevance to the welfare benefits of public works. In three substantial papers appearing in the *Annales des Ponts et Chaussées* (1844, 1849) and the *Journal des économistes* (1853), Dupuit became the first non-adventitious expositor of the theory of *marginal* utility, of (a variant of) marginal cost pricing, of simple and discriminating monopoly theory, and of pricing principles of the firm where location is a factor in expressing demand.

The font of Dupuit's contribution is the construction of a marginal utility curve and the identification of it with the demand curve or *courbe de consommation* (see Fig. 1).

Arguing in the manner of Carl Menger, who later elaborated on the point, Dupuit showed that the marginal utility that an individual obtained from a homogeneous stock of goods is determined by the use to which the last units of the stock are put. In doing so, he clearly pointed out that the marginal utility of a stock or some particular good diminishes with increases in quantity and that each consumer attaches a different marginal utility to the same good according to the quantity consumed. The importance of Dupuit's invention rests in the fact that the psychological concept of diminishing marginal utility, and its ramifications, were carried over to the law of demand. With some, but not all, of the reservations and qualifications of Alfred Marshall, Dupuit *identified* the marginal utility curve with the demand curve, adding up the utility curves of individuals to obtain the market demand curve. Dupuit (1844, p. 106) described his construction (see Fig. 1), which applied to all goods, public and private, as follows:

> If . . . along a line *Op* the lengths *Op, Op'*, *Op"* . . . represent various prices for an article, and that . . . *pn*, *p'n'*, *p"n"* . . . represent the number of articles consumed corresponding to these prices, then it is possible to construct a curve *Nn'n"P* which we shall call the curve of consumption. *ON* represents the quantity consumed when the price is zero, and *OP* the price at which consumption falls to zero.



Dupuit, Arsene-Jules-Emile Juvenal (1804–1866), Fig. 1

The identification of marginal utility and demand, of course, sets up the demand curve as a welfare tool and Dupuit made specific calculations. A measure of the welfare produced by the good (*utilité absolue*) at quantity *Or* is the definite integral of the demand curve between *O* and *r*. Given that *Op* is the (average) cost of producing quantity *Or*, consumers earn a surplus (*utilité relative*) equal to absolute utility (*OrnP*) less costs of production (*Ornp*). (Relative utility (*pnP*) is none other than Marshall's consumers' surplus without all the reservations that Marshall attached to the concept.) Importantly, Dupuit identified area *rNn* as lost utility (*utilité perdue*). Under competitive conditions this loss was inevitable due to the opportunity cost of resources. Under a monopoly structure, for example, if, in Fig. 1, *Op* were a monopoly price with zero production costs assumed, *utilité perdue* would be a loss to society – the 'deadweight' loss associated with excise taxes, tariffs or monopoly. Further, Dupuit advanced the theorem that the loss in utility was proportional to the *square* of the tax of price above marginal cost. This theorem, with attendant analysis, formed the base for large areas of neoclassical welfare economics, including the taxation studies of F.Y. Edgeworth and the marginal cost pricing argument of Harold Hotelling.

From this theoretical base, Dupuit investigated an impressive number of pricing systems and market models (1849). While Dupuit was an ardent and stubborn defender of laissez faire in most markets (1861), he was equally concerned that public works, provided or regulated by government as a *last* resort, should produce the maximum amount of utility possible. Thus tools such as marginal cost pricing find their theoretical foundations in the writings of Dupuit. Although Dupuit did not provide an explicit formulation of the principle, one of his bridge pricing examples and other statements strongly suggest the possibilities of such a technique to maximize welfare, but as a *long-run* proposition.

Dupuit analysed, independently of Cournot, who was apparently unknown to him, the profit-maximizing behaviour of the simple monopolist. He saw monopoly at the apex of a range of problems regarding the production of total welfare, being unconcerned about the 'distribution' of welfare between producers and consumers. His point was that the amount of 'absolute utility' (or what could be called net benefit) was lessened by monopoly profit maximization. This led him to defend the private practice of price discrimination and to produce an economic theory of discrimination. Price discrimination could exist, in Dupuit's view, with differences in 'buyer estimates', with the ability to segment markets either naturally or artificially, and with some degree of monopoly power. The motive was profit maximization, and although Dupuit discussed the effects of discrimination on price and revenue, he was primarily interested in the fact, as was Joan Robinson later, that discrimination could affect the size of the welfare benefit. This view was expanded to include the impact of price discrimination of welfare when buyers were spatially distributed (1849, 1854).

In the matter of policy, Dupuit recommended that tools be carefully fit to specific problems. If industries were to be collectivized or regulated by government, Dupuit proposed the maximization of net benefit under the constraint of covering total costs of production. The recovery of total cost might be achieved through regulated or constrained price discrimination or through a cost-based single price technique. However, Dupuit can hardly be credited with espousing an enlarged role for government or government intervention. A firm adherent of Smith's dictums concerning minimal government, Dupuit believed that free and open competition, along with vigorous antitrust or anticartel enforcement, would ensure optimal provisions in most cases, including transportation. Indeed, in the process of analysing the welfare principles of public works pricing, Dupuit discovered (in an uncommonly complete manner) some of the critical welfare-maximizing properties of a generalized competitive system.

See Also

- ▶ Consumer Surplus
- ▶ Public Utility Pricing and Finance

Selected Works

1844. On the measurement of the utility of public works. Trans. R.H. Barback from the *Annales des Ponts et Chaussées,* in *International Economic Papers,* No. 2, London: Macmillan, 1952.
1849. On tolls and transport charges. Trans. E. Henderson from the *Annales des Ponts et Chaussées,* in *International Economic Papers,* No. 11, London: Macmillan, 1962.
1853. On utility and its measure – on public utility. *Journal des économistes* 36, 1–27.
1854. Péages. In *Dictionnaire de l'Economie Politique,* vol. 2. Paris: Guillaumin.
1861. *La Liberté Commerciale.* Paris: Guillaumin.
1934. De l'Utilité et sa Mésure: *écrits choisis et republiés, ed. M. de Bernardi.*
Turin: La Riforma Sociale.

Bibliography

Ekelund, R.B. Jr. 1968. Jules Dupuit and the early theory of marginal cost pricing. *Journal of Political Economy* 76: 462–471.
Ekelund, R.B. Jr. 1970. Price discrimination and product differentiation in economic theory: An early analysis. *Quarterly Journal of Economics* 84: 268–278.
Ekelund, R.B. Jr., and Yeung-Nan Shieh. 1986. Dupuit, spatial economics, and optimal resource allocation: A French tradition. *Economica* 53: 483–496.

Durable Goods Markets and Aftermarkets

Michael Waldman

Abstract

There is an extensive literature on durable-goods markets that starts with the work of Akerlof, Coase, and Swan in the early 1970s. In this entry I survey the literature by starting with the three theoretical building blocks of time inconsistency, adverse selection, and substitutability between new and used units. I then focus on our understanding of three important real-world issues. These are whether firms choose optimal durability levels, whether firms have incentives to eliminate second-hand markets, and reasons for leasing. The article also provides an extensive discussion of aftermarket monopolization.

Keywords

Adverse selection; Aftermarket monopolization; Aftermarkets; Akerlof, G.; Asymmetric information; Bundling; Coase, R.; Commitment; Complementary goods; Deadweight loss; Durable goods markets; Durable-goods monopoly problem; Hold-up theories; Information costs; Leasing; Market power; Optimal durability; Price discrimination; Product-line pricing problem; Second-hand markets; Time inconsistency; Tying

JEL Classifications

L13

Durable goods are goods whose useful lifetime spans multiple periods.

This article surveys the extensive literature on durable-goods markets and aftermarkets. I begin with the main theoretical ideas, then turn to specific real-world issues such as durability choice and leasing, discuss aftermarket monopolization and then end with a brief conclusion. (A more in-depth survey appears in Waldman 2003.)

Three Theoretical Building Blocks

Much of our understanding of durable-goods markets derives from three theoretical contributions. The first is Coase's (1972) insight concerning time inconsistency. To see the basic logic, consider Bulow's (1982) formalization: a durable-goods monopolist sells its output in each of two periods and cannot commit in the first period to second-period actions. Bulow shows that, because in the second period the firm does not internalize how its actions affect the value of used units, its output is higher than under commitment. First-period

purchasers anticipate this, pay less for new units and thus lower overall monopoly profitability.

Coase's insight has spawned a large literature. One branch of this literature focuses on the Coase conjecture, that is, the idea that in an infinite-period setting time inconsistency causes price to drop immediately to marginal cost. A second branch identifies tactics such as leasing that firms can employ to reduce or possibly avoid time inconsistency. Finally, a third branch applies time inconsistency to other issues, including new-product introductions and repurchase prices.

The second major theoretical contribution is Akerlof's (1970) adverse-selection argument. This paper helped start the asymmetric-information revolution, but was not initially thought of as an important contribution to durable-goods theory. However, the paper's main example concerns second-hand markets. In Akerlof's model buyers have higher valuations than sellers, so efficiency requires that all units be traded. Further, each seller is privately informed of his own unit's quality. The result is a single price that reflects average quality, and sellers with high-quality units keep them because prices do not reflect actual quality, that is, trade is below the efficient level. (In Akerlof's analysis there is no trade, but this result is not robust.)

A small empirical literature looks for evidence of adverse selection in durable- goods markets. Most of these papers find some support. For example, Bond (1982) considers the used pickup truck market and finds support for adverse selection for older trucks, while Genesove (1993) finds some supporting evidence in used-car dealer auctions. More recently, Gilligan (2004) finds supporting evidence in business aircraft.

In terms of durable-goods theory, Akerlof's contribution was ignored for almost 30 years. Starting with Hendel and Lizzeri (1999a), however, a number of papers have extended Akerlof's analysis. There are three basic findings. First, Akerlof's main results continue to hold when new units are incorporated into the analysis. Second, because adverse selection in the used-unit market reduces the willingness to pay of new-unit buyers, firms will market new units in a manner that reduces adverse selection. Third, as discussed in detail later, new-unit leasing can be important for reducing adverse selection.

The third major theoretical contribution is that there is a close analogy between the product-line pricing problem and the durable-goods monopoly problem. This analogy is described in Waldman (1996). Consider Mussa and Rosen (1978), which analyses the product-line pricing problem of a non-durable-goods monopolist. The monopolist sells units of varying qualities to consumers who have heterogeneous valuations on quality. Because the substitutability between units links the various prices, the monopolist lowers below efficient levels the quality level sold to all but the highest-valuation group.

Now consider a durable-goods monopolist who controls the quality of a unit at every age. Further, assume heterogeneity in consumers' valuations for quality and a frictionless second-hand market. Then, if the firm can commit, quality choices are as above. That is, new-unit quality is efficient. But, because of the linkages between the various prices, all used-unit qualities are below efficient levels. As discussed later, a number of recent papers use this result to analyse various real-world issues concerning durable goods.

Three Real-World Issues

Optimal Durability Choice

A much debated issue is whether a durable-goods monopolist chooses socially optimal durability. Swan (1970, 1971) considers models that satisfy the once standard assumption that a unit is a bundle of 'service units', so some number of used units is a perfect substitute for a new unit. Swan's steady-state analysis shows durability choice to be socially optimal because the firm produces the steady-state flow of service units at minimum cost. (Swan's analysis corrected the conclusions of earlier papers that had concluded that in such settings the monopolist would choose inefficiently low durability levels.)

A large literature investigates the robustness of Swan's conclusions. There are two major findings. The first employs time inconsistency. Bulow (1986) moves away from Swan's

assumption of steady-state behaviour by considering a model similar to his earlier one, but now allows endogenous durability choice. He shows that time inconsistency provides a rationale for a durable-goods monopolist to choose less than the socially-optimal durability level. The logic is that durability is what leads to time inconsistency, so reducing durability below the efficient level reduces time inconsistency and thus increases profitability.

The second major finding appears in Waldman (1996) and Hendel and Lizzeri (1999b), which drop the service units assumption and instead assume that new and used units vary in quality and that durability choice controls the speed of quality deterioration. The earlier discussion immediately translates into an incentive for the firm to choose less than the socially optimal durability level. That is, in this setting the incentive for the monopolist to sell output whose used-unit quality is below the efficient level translates into durability below the efficient level. (In Hendel and Lizzeri's analysis durability choice can be above, below, or equal to the first-best level, but it is always below the second-best level defined by actual outputs.)

Eliminating Second-Hand Markets

Do durable-goods producers with market power have incentives to eliminate secondhand markets? For example, do textbook publishers introduce new editions in order to kill off the market for used books? Until recently, the standard argument, found, for example, in Swan (1980), was that, since the new-unit price reflects prices the product will sell for on the second-hand market in subsequent periods, the producer has no such incentive.

Two recent arguments show that this result is, in fact, quite limited. The first, which builds on the discussion above, appears in Waldman (1996, 1997) and Hendel and Lizzeri (1999b). The idea is that, because substitutability between new and used units means the price of a used unit on the second-hand market limits the amount the firm can charge for new units, the firm sometimes eliminates the second-hand market or similarly reduces used-unit availability in order to raise the new-unit price. In particular, this is more likely when consumers of used units have low valuations for the firm's product. This is both because little revenue is lost by not serving such consumers and because serving them means a low used-unit price and thus a lower new-unit price. (A number of earlier papers find similar results starting with demand functions rather than utility maximization.)

The second argument, found initially in Waldman (1993), employs time inconsistency. As discussed, the early literature on time inconsistency focused on output choice. My 1993 paper shows time inconsistency also applies to actions such as new-product introductions that make used units unavailable because they become obsolete. The difference between this argument and the one above concerns commitment. Above it is assumed the firm can commit, so the firm eliminates the second-hand market only when it is profitable to do so. In contrast, here commitment is not assumed, so the firm may eliminate the second-hand market even though this lowers overall profitability.

A related empirical analysis appears in Iizuka (2004), which shows that the market share of used textbooks is an important determinant of whether or not a publisher introduces a new edition. This is consistent with new editions being used at least partly to eliminate second-hand markets, although Iizuka does not distinguish between the two possibilities described above for why a firm might want to do this. In future research, it might be possible to identify which argument is at work by focusing on how the decision to introduce new editions affects overall profitability.

Reasons for Leasing

A number of reasons have been identified for why durable-goods producers frequently lease. (A reason I do not discuss is that there are sometimes tax advantages associated with leasing.) One reason, initially discussed in Coase (1972) and Bulow (1982), is that time inconsistency lowers profitability when a firm sells output because it chooses actions in later periods that inefficiently lower the value of used units. When the firm leases, however, it retains ownership of

those units so the incentive to take inefficient actions disappears.

A second reason is also related to a previous discussion. As discussed in Waldman (1997) and Hendel and Lizzeri (1999b), when used-unit prices serve as important constraints on the new-unit price, leasing can be used to eliminate secondhand markets or at least reduce used-unit availability. The logic is that leasing allows a firm to eliminate the second-hand market by allowing the firm to retire returned used units. My 1997 paper shows this formally and argues that it is consistent with classic cases concerning the use of a lease-only policy such as United Shoe in the shoe machinery market, IBM in the computer market, and Xerox in the copier market. (One might argue that leasing is not needed because a firm can sell and then use high repurchase prices to purchase and retire used units. My 1997 paper shows this strategy is inferior to leasing because of time inconsistency.)

Finally, leasing is a response to adverse selection. This argument appears in Hendel and Lizzeri (2002) and Johnson and Waldman (2003). These papers show that, whether the new-unit market is monopolistic or competitive, in a world of asymmetric information leasing in the new-unit market can arise because it means used units are returned to the seller(s), which, in turn, avoids or at least reduces adverse selection in the used-unit market. The two papers develop different variants of the argument and show it is consistent with various empirical findings concerning the automobile market.

Aftermarket Monopolization

Aftermarket monopolization is behaviour that stops alternative producers from selling aftermarket products to the firm's customers. The focus on this subject started after the US Supreme Court's 1992 decision in the case *Eastman Kodak Company v. Image Technical Services.* Aftermarkets are common with durable goods, where aftermarkets refer to markets for complementary products such as maintenance and upgrades. I consider three possibilities: (*a*) hold-up rationales; (*b*) price discrimination and efficiency rationales; and (*c*) other strategic rationales.

Hold-Up

There are two distinct hold-up arguments, each of which focuses on aftermarket monopolization by competitive producers. In both, the firm prohibits other firms from selling the aftermarket product – for example, maintenance – and then exploits the locked-in positions of its customers in pricing the product. The result is a standard deadweight loss due to the high aftermarket price, although no transfer between the consumers and the firm since competition in the primary market means firms earn zero profits overall. In the 'costly-information' version, consumers ignore the aftermarket price when purchasing the primary product. In the 'lack-of- commitment' version, developed in Borenstein et al. (1995), consumers correctly anticipate the aftermarket price but, because firms cannot commit, time inconsistency causes firms to monopolize the aftermarket and inefficiently raise the aftermarket price after consumers are locked in. (A third holdup theory is the 'surprise' theory. In this argument consumers are surprised by the aftermarket monopolization. Some discussions of this theory describe a transfer between the consumers and the firm, but it is unclear why competition does not result in zero profits, in which case the surprise and costly-information theories are equivalent.)

Price Discrimination and Efficiency Rationales

For various reasons, such as that many buyers in the relevant industries are sophisticated firms for which the costly-information argument is implausible, attention has shifted towards other arguments many of which have either neutral or positive social-welfare implications.

One such argument is the price discrimination argument that appears in Chen and Ross (1993) and Klein (1993). Suppose the primary-good producer has market power. Then the firm may monopolize the aftermarket in order to raise the aftermarket price and in this way price discriminate by charging a high aggregate price to the high-volume/high-valuation consumers. From a social-welfare standpoint, this argument has

neutral implications since an improved ability to price discriminate can either raise or lower social welfare. (Klein argues that this argument applies even when firms are competitive, although not perfectly competitive.)

A plausible efficiency rationale follows from Schmalensee's (1974) argument that, given a durable-goods monopolist (which means new units priced above marginal cost) and a competitive maintenance market (which means maintenance is priced at cost), consumers will sometimes inefficiently maintain rather than replace used units. Tirole (1988) shows this can lead to aftermarket monopolization in a durable-goods monopoly setting because having a monopoly in both markets allows the firm to avoid the inefficiency and thus increases its profits.

More recently, Morita and Waldman (2005) and Carlton and Waldman (2006) show that the argument extends to aftermarkets other than maintenance, and to competitive durable-goods markets given switching costs. In the latter case the inefficient substitution problem arises even with competition because switching costs create market power at the time of the maintenance/replacement decision. Interestingly, because competitive sellers earn zero profits in equilibrium, when aftermarket monopolization eliminates the distortion, both social welfare and consumer welfare increase.

Strategic Rationales

There is an extensive literature on strategic rationales for the tying of complementary products. Since the tying of primary and aftermarket products is one potential way to achieve aftermarket monopolization, much of this literature is relevant to aftermarket monopolization.

Whinston (1990) shows that, if the primary good is not essential, tying may force the exit of an alternative producer of the complementary good and in this way increase the firm's profits by monopolizing the segment of the complementary-good market for which the primary good is not required.

In contrast, in Carlton and Waldman (2002) tying is sometimes used to preserve a monopoly in the primary-good market. They consider two-period settings in which a single potential entrant can enter the complementary market in either period but the primary market only in the second. In the presence of fixed costs of entry or network externalities, the primary-good monopolist sometimes ties in order to preserve its primary-good monopoly in the second period. For example, with entry costs tying stops the alternative producer from entering the complementary market in the first period. In turn, because of a possible inability to cover entry costs, the outcome can be no entry in either market in either period.

A third argument appears in Carlton and Waldman (2005). Whinston shows that in one-period settings there is never an incentive to tie if the monopolist's primary product is essential. Carlton and I show that in durable-goods settings, given the presence of complementary-good upgrades and switching costs, tying can be optimal even when the primary product is essential. The basic logic is that some profits are realized in later periods in the sale or lease of the upgraded complementary good, and the only way the monopolist can ensure it captures those profits is by tying and becoming the sole producer of the complementary good.

Conclusion

Starting in the early 1970s with the work of Akerlof, Coase and Swan, significant progress has been made in our understanding of durable-goods markets. In this entry I have surveyed this literature as well as the literature on the related issue of aftermarkets. Although I have referred throughout to various empirical papers, durable-goods markets is a topic for which theory is far ahead of empirical investigation. In the future I expect to see work that extends the theory in various important ways, but also empirical work that tests the validity of the various theoretical approaches that have been explored since the early 1970s.

See Also

► Adverse Selection
► Akerlof, George Arthur (Born 1940)

▶ Bundling and Tying
▶ Coase, Ronald Harry (Born 1910)
▶ Resale Markets

Bibliography

Akerlof, G. 1970. The market for 'lemons': Quality uncertainty and the market mechanism. *Quarterly Journal of Economics* 84: 488–500.
Bond, E. 1982. A direct test of the 'lemons' model: The market for used pick-up trucks. *American Economic Review* 72: 836–840.
Borenstein, S., J. Mackie-Mason, and J. Netz. 1995. Antitrust policy in aftermarkets. *Antitrust Law Journal* 63: 455–482.
Bulow, J. 1982. Durable goods monopolists. *Journal of Political Economy* 90: 314–332.
Bulow, J. 1986. An economic theory of planned obsolescence. *Quarterly Journal of Economics* 101: 729–749.
Carlton, D., and M. Waldman. 2002. The strategic use of tying to preserve and create market power in evolving industries. *RAND Journal of Economics* 33: 194–220.
Carlton, D., and M. Waldman. 2005. *Tying, upgrades, and switching costs in durable goods markets*. Mimeo: Cornell University.
Carlton, D., and M. Waldman. 2006. *Competition, monopoly, and aftermarkets*. Mimeo: Cornell University.
Chen, Z., and T. Ross. 1993. Refusals to deal, price discrimination and independent service organizations. *Journal of Economics and Management Strategy* 2: 593–614.
Coase, R. 1972. Durability and monopoly. *Journal of Law and Economics* 15: 143–149.
Genesove, D. 1993. Adverse selection in the wholesale used car market. *Journal of Political Economy* 101: 644–665.
Gilligan, T. 2004. Lemons and leases in the used business aircraft market. *Journal of Political Economy* 112: 1157–1180.
Hendel, I., and A. Lizzeri. 1999a. Adverse selection in durable goods markets. *American Economic Review* 89: 1097–1115.
Hendel, I., and A. Lizzeri. 1999b. Interfering with secondary markets. *RAND Journal of Economics* 30: 1–21.
Hendel, I., and A. Lizzeri. 2002. The role of leasing under adverse selection. *Journal of Political Economy* 110: 113–143.
Iizuka, T. 2004. *An empirical analysis of planned obsolescence*. Mimeo: Vanderbilt University.
Johnson, J., and M. Waldman. 2003. Leasing, lemons, and buy-backs. *RAND Journal of Economics* 34: 247–265.
Klein, B. 1993. Market power in antitrust: Economic analysis after *Kodak*. *Supreme Court Economic Review* 3: 43–92.
Morita, H., and M. Waldman. 2005. *Competition, monopoly maintenance, and consumer switching costs*. Mimeo: Cornell University.
Mussa, M., and S. Rosen. 1978. Monopoly and product quality. *Journal of Economic Theory* 18: 301–317.
Schmalensee, R. 1974. Market structure, durability, and maintenance effort. *Review of Economic Studies* 41: 277–287.
Swan, P. 1970. Durability of consumption goods. *American Economic Review* 60: 884–894.
Swan, P. 1971. The durability of goods and regulation of monopoly. *Bell Journal of Economics* 2: 347–357.
Swan, P. 1980. Alcoa: The influence of recycling on monopoly power. *Journal of Political Economy* 88: 76–99.
Tirole, J. 1988. *The theory of industrial organization*. Cambridge: MIT Press.
Waldman, M. 1993. A new perspective on planned obsolescence. *Quarterly Journal of Economics* 108: 273–283.
Waldman, M. 1996. Durable goods pricing when quality matters. *Journal of Business* 69: 489–510.
Waldman, M. 1997. Eliminating the market for secondhand goods. *Journal of Law and Economics* 40: 61–92.
Waldman, M. 2003. Durable goods theory for real world markets. *Journal of Economic Perspectives* 17(1): 131–154.
Whinston, M. 1990. Tying, foreclosure, and exclusion. *American Economic Review* 80: 837–859.

Durand, David (Born 1912)

J. Fred Weston

Durand was born in Ithaca, New York. He received his PhD at Columbia University in 1941. He was a member of the Research Staff of the National Bureau of Economic Research from 1946 to 1955 when he became a Professor at the Massachusetts Institute of Technology, where he remained throughout his career.

The early contributions of David Durand were in statistical methodology. His election as a Fellow of the American Statistical Association was based on his work in developing and applying statistical analysis in the field of finance, including the construction of historical series on the term structure of interest rates.

In a National Bureau Conference publication (1952), Durand authored a chapter which laid the foundation for later developments on cost of capital theory and measurement. Durand followed the

premise that security appraisal is the key to measuring the cost of capital. He developed two alternative methods of appraisal, the Net Income (NI) method and the Net Operating Income (NOI) method. In the NI method, the cost of debt interest is deducted from net operating income and net income is capitalized at a constant rate; the value of the firm increases with higher debt leverage until both the cost of debt and of equity rise substantially. Under the NOI method a constant capitalization rate is applied to the net operating income so that the total value of all bonds and stocks is invariant so the degree of leverage employed. Durand leaned toward the NOI method, but recognized circumstances under which each model had applicability. Subsequent literature predominantly favoured the NOI approach until the 1980s, when Durand's views have essentially prevailed.

Durand also made contributions to the theory of capital budgeting. He emphasized an eclectic approach including the internal rate of return method and the net present value procedure. He also emphasized considering some measure of time to indicate how fast an investment project will liquidate itself. He argued that just as portfolio managers select bonds partly on the basis of term to maturity, financial managers responsible for capital budgeting should select investment projects partly on the basis of their weighted discounted payout period or duration.

Selected Works

1941. *Risk elements in consumer installment financing.* New York: National Bureau of Economic Research.

1942. *Basic yields of corporate bonds: 1900–1942.* Technical Paper No. 3, New York: National Bureau of Economic Research.

1948. An appraisal of the errors involved in estimating the size distribution of a given aggregate income. *Review of Economics and Statistics* 30: 63–8.

1952. Costs of debt and equity funds for business: Trends and problems of measurement. In *Conference on research in business finance,* 215–47. New York: National Bureau of Economic Research.

1954. Joint confidence regions for multiple regression coefficients. *Journal of the American Statistical Association* 49: 130–46.

1957. Growth stocks and the Petersburg paradox. *Journal of Finance* 12: 348–63.

1971. *Stable chaos: An introduction to statistical control.* Morristown: D.H. Mark Publication of General Learning Press.

1974. Payout period, time spread and duration: Aids to judgment in capital budgeting. *Journal of Bank Research* 5: 20–34.

1981. Comprehensiveness in capital budgeting. *Financial Management* 10(5): 7–13.

Durbin, Evan Frank Mottram (1906–1948)

Elizabeth Durbin

Born in Devon, Durbin was a Scholar of New College, Oxford, won the Senior and Junior Webb Medley Scholarships, first class Honours in Politics, Philosophy and Economics and the Ricardo Fellowship to University College, London. Hired as an economics lecturer at the London School of Economics in 1930, he was later promoted to senior lecturer. During the war he was a personal assistant to Clement Attlee, the Deputy Prime Minister, and in 1945 he was elected Labour Member of Parliament for Edmonton. He served as Parliamentary Private Secretary to Hugh Dalton at the Treasury, and was appointed junior Minister of Works in March 1947. He was drowned in Cornwall in 1948. Durbin is best remembered for his book *The Politics of Democratic Socialism* (1940), an influential statement of the revisionist case in Britain, of which his close friend and professional colleague, Hugh Gaitskell, later commented: 'it marked the transition from the pioneering stage to that of responsibility and power.'

As a professional economist, Durbin published two books on macroeconomic theory and policy, and a number of articles on economic planning. In the intellectual turmoil of the early 1930s, he was searching for a theory to explain the trade cycle, because he believed that its control was essential to the socialist alternative to capitalism. He was strongly influenced by Hayek's cyclical theory of sectoral imbalance, although he argued that the crisis was precipitated by 'an excessive supply of money' in the consumers' sector, not capital scarcity in the producers' sector (Hayek's view). In further work, he introduced the role of the money market, an important advance on Hayek's model, which foreshadowed Keynes's use of uncertainty in *The General Theory*. Later scholars have also recognized his contributions to identifying the crucial growth problem of maintaining sufficient savings without causing sectoral imbalance. Together with Hugh Gaitskell, J.E. Meade and Douglas Jay, Durbin has also been credited with adapting the Keynesian revolution into practical policies for the British Labour Party. However, he always remained sceptical about some aspects of *The General Theory*; he did not believe it provided a solution of the cyclical problem, and he was concerned about the inflationary potential of continued expansion.

Beginning in 1931 through the Fabian Society, Durbin and Gaitskell also organized systematic research into the theory and practice of socialist planning and the appropriate criteria for assessing efficiency in a socialist economy. Thus Durbin was in the forefront of the planning controversies of the Thirties, contributing articles to the development of the 'competitive' solution for market socialism and to the marginal cost-pricing debate. He was one of the first to argue that a mixed economy was fundamental to the notion of democratic socialism; the market provided individuals freedom to choose jobs and goods and incentives to innovate, and the government provided the programme and policies to sustain growth, to allocate resources in the public interest and to ensure social justice.

Durbin's main achievements were to present a practical forerunner of the postwar mainstream case for government intervention and to lay the intellectual foundations for the continuing debates about the nature of the socialist vision in Britain.

See Also

- ▶ Fabian Economics
- ▶ Social Democracy

Selected Works

1933a. *Purchasing power and trade depression*. London: Chapman & Hall.
1935a. *The problem of credit policy*. London: Chapman & Hall.
1935b. The social significance of the theory of value. *Economic Journal* 45: 700–710.
1940. *The politics of democratic socialism*. London: Routledge & Kegan Paul.
1949. *Problems of economic planning*. London: Routledge & Kegan Paul.

References

Durbin, E. 1985. *New Jerusalems: The labour party and the economics of democratic socialism*. London: Routledge & Kegan Paul.

Durbin-Watson Statistic

James G. MacKinnon

Keywords

Durbin–Watson statistic; Linear regression models; Monte Carlo test; Ordinary least squares (OLS) estimator; Serial correlation; Testing; DW statistic

JEL Classifications

C1

The well-known Durbin–Watson, or DW, statistic, which was proposed by Durbin and Watson (1950, 1951), is used for testing the null hypothesis that the error terms of a linear regression model are serially independent.

Consider the linear regression model with AR (1) errors,

$$y_t = \boldsymbol{X}_t b + u_t, u_t = \rho u_{t-1} + \varepsilon_t, \varepsilon_t \sim \text{IID}\left(0, \sigma^2\right). \quad (1)$$

Here the scalar y_t is an observation on a dependent variable, $\boldsymbol{X}_t$ is a $1 \times k$ vector of observations on independent variables that may be treated as fixed, and $\boldsymbol{\beta}$ is a k-vector of parameters to be estimated. There are n observations, and we wish to test the null hypothesis that $\rho = 0$, under which the model (1) reduces to

$$y_t = \boldsymbol{X}_t b + u_t, u_t \sim \text{IID}\left(0, \sigma^2\right), \quad (2)$$

for which the ordinary least squares (OLS) estimator is efficient. This estimator is usually written as $\hat{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$, where the $n \times k$ matrix $\boldsymbol{X}$ has t^{th} row $\boldsymbol{X}_t$ and the n-vector $\boldsymbol{y}$ has t^{th} element y_t.

The Durbin–Watson d statistic for testing (2) against (1) is solely a function of the OLS residuals $\hat{u}_t = y_t - \boldsymbol{X}_t\hat{b}$. It is defined as

$$d = \frac{\sum_{t=2}^{n}\left(\hat{u}_t - \hat{u}_{t-1}\right)^2}{\sum_{t=1}^{n}\hat{u}_t^2}. \quad (3)$$

It is easy to see that d is approximately equal to $2 - 2\hat{\rho}$, where $\hat{\rho}$ is the OLS estimate of ρ in a regression of $\hat{u}_t$ on $\hat{u}_{t-1}$. Thus d will be approximately equal to 2 if the residuals do not display any serial correlation, and it will be less (greater) than 2 whenever $\hat{\rho}$ is more than a little bit greater (less) than 0.

The Exact Distribution of the DW Statistic

The DW statistic can be written as a ratio of quadratic forms in the n-vector $\hat{\boldsymbol{u}}$ of OLS residuals, the t^{th} element of which is $\hat{u}_t$. Specifically,

$$d = \frac{\hat{\boldsymbol{u}}'\boldsymbol{A}\hat{\boldsymbol{u}}}{\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}}}, \quad (4)$$

where $\boldsymbol{A}$ is the $n \times n$ matrix

$$\frac{1}{2}\begin{bmatrix} 1 & -1 & 0 & 0 & \dots & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & \dots & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & \dots & 0 & -1 & 1 \end{bmatrix}$$

Durbin and Watson (1950) actually considered a number of statistics that can be written in the form of (4) for different choices of the matrix $\boldsymbol{A}$ and chose to focus on d for reasons of computational and theoretical convenience. Because both the numerator and the denominator are proportional to σ^2, d is invariant to σ.

The exact distribution of d depends on $\boldsymbol{X}$ and the distribution of the u_t. When the error terms are i.i.d. normal, Durbin and Watson (1951) tabulated bounds on the critical values for tests based on d against the one-sided alternative that $\rho > 0$. These bounds, denoted d_L and d_U, depend on the sample size and the number of regressors. We can reject the null hypothesis when $d < d_L$, cannot reject it when $d < d_U$, and can draw no firm conclusion when $d_L < d < d_U$. To test against the alternative that $\rho < 0$, we would replace d by $4 - d$ and use the same procedure.

The original Durbin–Watson tables have been extended by various authors, notably Savin and White (1977). However, since $d_U - d_L$ can be quite large, tests based on the bounds often have indeterminate outcomes. It is much better to perform exact tests conditional on $\boldsymbol{X}$, and this is easy to do with modern computing technology. There are two approaches.

The first approach is to calculate an exact P value for d using one of several methods for calculating the distribution of a ratio of quadratic forms in normal random variables. The method of Imhof (1961) is probably the best known of these, but the more recent method of Ansley et al. (1992) is faster. If a suitable computer program is readily available, this approach is the best one.

An alternative approach is to perform a Monte Carlo test. As can be seen from (4), the statistic d depends only on the vector $\boldsymbol{u}$ and the matrix $\boldsymbol{X}$,

since $\hat{\boldsymbol{u}} = \boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{u}$, where $\mathbf{M}_{\boldsymbol{X}} = \mathbf{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. Because of its invariance to σ, d does not depend on any unknown parameters. This implies that a Monte Carlo test will be exact.

To perform a Monte Carlo test at level α, we first choose B such that $\alpha(B + 1)$ is an integer (999 is often a reasonable choice) and generate B vectors $\boldsymbol{u}_j^*$, each of which is multivariate standard normal. Each of the $\boldsymbol{u}_j^*$ is regressed on $\boldsymbol{X}$ to calculate a vector of residuals $\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{u}_j^*$, which is then used to compute a simulated test statistic d_j^* according to (3). We can then calculate simulated P values for a one-tailed test against either $\rho > 0$ or $\rho < 0$ or for a two-tailed test. For example, the simulated P value for a one-tailed test against $\rho > 0$ is

$$P^*(d) = \frac{1}{\boldsymbol{B}} \sum_{j=1}^{\boldsymbol{B}} I\left(d_j^* < d\right),$$

where $I(\cdot)$ is the indicator function that is equal to 1 when its argument is true and equal to 0 otherwise. We reject the null hypothesis whenever $P^*(d) < \alpha$. For more on the calculation of P values for bootstrap and Monte Carlo tests, see Davidson and MacKinnon (2006).

Limitations of the DW Statistic

The Durbin–Watson statistic is valid only when all the regressors can be treated as fixed. It is not valid, even asymptotically, when $\boldsymbol{X}_t$ includes a lagged dependent variable or any variable that depends on lagged values of y_t. Because $\hat{\rho}$ is biased towards 0 when $\boldsymbol{X}_t$ includes a lagged dependent variable, d is biased towards 2 in this case. Thus, a test based on the DW statistic will tend to under-reject when the null hypothesis is false.

Numerous procedures have been proposed for testing for serial correlation in models that include lagged dependent variables. The simplest is to rerun regression (2), with the addition of the lagged residuals from that regression. The test statistic is then the t statistic on the lagged residuals. This procedure, which is due to Durbin (1970) and Godfrey (1978), does not yield an exact test and should be bootstrapped when the sample size is small.

Of course, since the finite-sample distribution of the DW statistic depends on the distribution of the u_t, we cannot expect to obtain an exact test even when the $\boldsymbol{X}_t$ are exogenous if the normality assumption is not a good one. In principle, we could bootstrap d by using re-sampled residuals instead of multivariate standard normal vectors for the $\boldsymbol{u}_j^*$. This would probably work very well in most cases, but it would not actually yield an exact test.

See Also

- Artificial Regressions
- Serial Correlation and Serial Dependence

Bibliography

Ansley, C., R. Kohn, and T. Shively. 1992. Computing *p*-values for the generalized Durbin–Watson statistic and other invariant test statistics. *Journal of Econometrics* 54: 277–300.

Davidson, R., and J. MacKinnon. 2006. Bootstrap methods in econometrics. In *Palgrave handbooks of econometrics: volume 1: Econometric theory*, ed. T. Mills and K. Patterson. Basingstoke: Palgrave Macmillan.

Durbin, J. 1970. Testing for serial correlation in least-squares regression when some of the regressors are lagged dependent variables. *Econometrica* 38: 410–421.

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

Durbin, J., and G. Watson. 1951. Testing for serial correlation in least squares regression II. *Biometrika* 38: 159–177.

Godfrey, L. 1978. Testing against general autoregressive and moving average error models when the regressors include lagged dependent variables. *Econometrica* 46: 1293–1301.

Imhof, J. 1961. Computing the distribution of quadratic forms in normal variables. *Biometrika* 48: 419–426.

Savin, N., and K. White. 1977. The Durbin–Watson test for serial correlation with extreme sample sizes or many regressors. *Econometrica* 45: 1989–1996.

Durkheim, Emile (1858–1917)

Peter Bearman

Born in Epinal near Strasbourg, Durkheim attended the Ecole Normale Supérieure in Paris, taking his *agregation* in 1882. His first important academic appointment was as Professor of Sociology and Education at Bordeaux in 1887. The Bordeaux appointment marked the first sociology professorship in France. In 1902 Durkheim was appointed as Professor of Sociology (and Education), at the Sorbonne where he remained until his death in 1917. Of Durkheim's four major works, *The Division of Labour* (1893), *The Rules of Sociological Method* (1895), *Suicide*: *A Study in Sociology* (1897) and *The Elementary Forms of Religious Life* (1912), the first three were written while he was at Bordeaux. At the Sorbonne, Durkheim devoted considerable effort towards the establishment of sociology as a professional discipline. He founded a journal, *L'année sociologique*, and was active in supervision of younger scholars, most notably Granet, Mauss, and Halbwachs.

The theoretical agenda of Durkheim's major works centred around understanding the collective bases for social order in modern societies characterized by increased individuation and autonomy. For Durkheim, a stable social order was possible only if the members of a group shared a common set of beliefs (*conscience collective*) governing individual behaviour. With the progression of the division of labour in society – which by definition leads to greater individuation and specialization of persons and roles – the values that increasingly heterogeneous individuals hold are seen to become more abstract, and are thus less able to shape and constrain individual social action. The social order, in this context, may become weak. Durkheim recognized that countervailing the weakness of the collective conscience in modern society was increased functional interdependence of persons. Yet, in *The Division of Labour*, Durkheim attacked the notion (attributed to Spencer as a representative utilitarian) that such interdependence (the need for individuals to exchange the products of their labour) was by itself robust enough to guarantee social stability. Rather, he asserted that exchange is possible only because of the existence of shared sentiments which govern the determination of 'individual interest' and behaviour. Contracts presume prior sentiments constraining self-interested social action.

Durkheim's fundamental methodological contribution to sociology is the recognition that macro-level outcomes cannot be accounted for from the analysis (or empirical observation) of micro-level (individual) action. Rather, he argued that social scientists must recognize that society is a 'reality sui generis', a 'thing' greater than, and not reducible to, its constituent parts. In this framework, sociology is the positive science of social facts, phenomena whose own structure can be used as an indicator of the social solidarity of a group which one cannot directly apprehend from observation alone.

Durkheim is considered a founder of modern sociology and anthropology. On both substantive and methodological grounds, his work can be considered a sustained attack on economic theory which typically elides the problem of social order and assumes that aggregate social outcomes are the products of individual social action and individual self-interest.

Selected Works

1893. In *The division of labor in society.* Trans. and ed. G. Simpson. New York: Free Press, 1947.

1895. In *The rules of sociological method* (trans: Solovay, S.A. and Mueller, J.H.), ed. G. Catlin. New York: Free Press, 1958.

1897. In *Suicide: A study in sociology* (trans: Spaulding, J.A. and Simpson, G.), ed. G. Simpson. New York: Free Press, 1951.

1912. In *The elementary forms of the religious life* (trans: Swain, J.W.). New York: Free Press, 1954.

References

Lukes, S. 1973. *Emile Durkheim: His life and work: A historical and critical study.* Harmondsworth: Penguin.

Parsons, T. 1968. Emile Durkheim. In *International encyclopedia of the social sciences*, ed. D.L. Shils. New York: Macmillan.

Dutch Disease

Thorvaldur Gylfason

Abstract

This article outlines the 'Dutch disease', the fear of de-industrialization first seen in the Netherlands in the wake of the appreciation of the Dutch guilder following the discovery of natural gas deposits in the North Sea around 1960. It considers its symptoms, and asks whether it is indeed a 'disease' with negative economic implications. It also briefly reviews some cases of Dutch disease, and the case of Norway, which appears to have successfully avoided it.

Keywords

De-industrialization; Dutch disease; Oil

JEL Classifications

F43

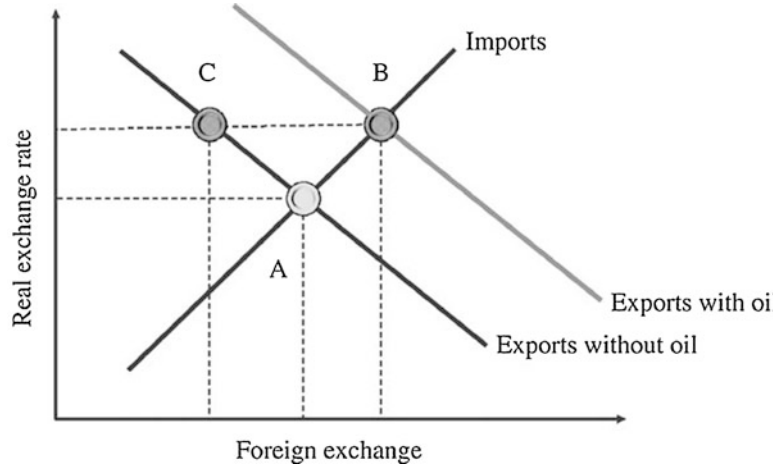
Dutch disease, in the original sense of the term, refers to the fears of de-industrialization that gripped the Netherlands in the wake of the appreciation of the Dutch guilder following the discovery of natural gas deposits in the North Sea around 1960. The appreciation of the guilder following the gas export boom hurt the profitability of manufacturing and service exports. Total exports from the Netherlands decreased markedly relative to Gross Domestic Product (GDP) during the 1960s. The growth of petroleum exports in the 1960s hurt other exports disproportionately. Many feared dire consequences for Dutch manufacturing. The problem proved short-lived, however. From the late 1960s onward, exports of goods and services have increased from less than 40 per cent of GDP to more than 70 per cent, a high export ratio by world standards. The expected de-industrialization did not materialize, but the name stuck. It can be said that, being neither Dutch nor a disease, the Dutch disease is a double misnomer. But when a disease bears the name of the first patient diagnosed with it, it would seem a bit harsh to require the patient to remain sick for the name to stick.

Is it a disease? Some view it as matter of one sector benefiting at the expense of others, without seeing any macroeconomic or social damage done. Others view the Dutch disease as an ailment, pointing to the potentially harmful consequences of the resulting reallocation of resources – from high-tech, high-skill intensive service industries to low-tech, low-skill intensive primary production, for example – for economic growth and diversification.

Symptoms

An overvalued currency was the first symptom associated with the Dutch disease, but later several other symptoms came to light. Figure 1 illustrates how an oil export boom lifts the equilibrium real exchange rate at which total exports of goods, services, and capital match total imports. In the figure, non-oil exports decline from A to C and hence by less than oil exports increase, so that total exports rise from A to B. For total exports to decline the import schedule would have to shift to the left (for instance through capital inflow) by an amount that exceeds the increase in oil exports, measured by the distance between B and C.

Natural resource discoveries and dependence tend to go hand in hand with booms and busts: the prices and supplies of raw materials and related commodities fluctuate a great deal in world markets. Fish stocks, for example, are notoriously volatile. Oil wells are drilled, and then go dry, and mines are depleted. The



Dutch Disease,
Fig. 1 How an oil export boom crowds out nonoil exports

resulting fluctuations in export earnings trigger exchange rate volatility, perhaps no less so under fixed exchange rates than under floating rates. Unstable currencies create uncertainty, which tends to hurt exports and imports as well as foreign investment. Further, the Dutch disease can strike even in countries that do not have a national currency of their own (as, for instance, in Greenland, which uses the Danish krone and depends on fish). In this case, the natural-resource-based industry is able to pay higher wages and also higher interest rates than other industries, thus making it difficult for the latter to stay competitive. This problem can become particularly acute in countries with centralized wage bargaining (or with oligopolistic banking systems, for that matter) where the natural-resource-intensive industries set the tone in nation-wide wage negotiations and dictate wage settlements which other industries can ill afford. In one or all of these ways, the Dutch disease tends to reduce the level of total exports or skew the composition of exports away from manufacturing and service exports which could be particularly conducive to economic growth over time. Exports of capital, including inward foreign direct investment, may also suffer.

The Netherlands recovered quickly from the Dutch disease, and has seen a persistent upward trend in its total exports relative to GDP since the mid-1960s. On the other hand, in Norway, the world's third largest oil exporter after Saudi Arabia and Russia, total exports have risen slowly relative to GDP, to a level well below that of the Netherlands (45 per cent in Norway in 2005 compared with 71 per cent in the Netherlands), even if the Dutch economy is almost three times as large as that of Norway. Also, the share of manufactured exports in merchandise exports was 68 per cent in the Netherlands in 2005 compared with 17 per cent in Norway. Exports and manufacturing are good for growth. Openness to trade invigorates imports of goods and services, capital, technology, ideas and know-how. The Dutch disease matters mainly because of its potentially harmful consequences for economic growth.

Channels

Experience seems to suggest six main channels of transmission from heavy natural resource dependence to sluggish economic growth. At the top of the list is the Dutch disease. In second place, huge natural resource rents, especially in conjunction with ill-defined property rights, imperfect or missing markets, and lax legal structures, may lead to rent-seeking behaviour that diverts resources away from more socially fruitful economic activity. The struggle for resource rents may lead to a concentration of economic and political power in the hands of elites which, once in power, use the rent to placate their political supporters and thus

secure their hold on power, with stunted or weakened democracy and slow growth as a result. Extensive rent seeking – in other words, seeking to make money from market distortions – can breed corruption, thus reducing both economic efficiency and social equality.

Third, natural resource abundance may imbue people with a false sense of security and lead governments to lose sight of the need for growth-friendly economic management, including free trade, foreign investment, bureaucratic efficiency and good institutions, including democracy. Incentives to create wealth through good policies and institutions may wane because of the relatively effortless ability to extract wealth from the soil or the sea. Fourth, abundant natural resources may likewise weaken incentives to accumulate human capital, even if the rent stream from the resources may enable nations to give a high priority to education. Fifth, natural resource abundance may blunt private and public incentives to save and invest in real capital no less than in human capital, and thereby weaken financial institutions and reduce economic growth. Sixth, natural resource wealth is a fixed factor of production that hampers economic growth potential by causing a growing labour force and a growing stock of capital to run into diminishing returns.

In sum, an abundance of natural capital, if not well managed, may erode or reduce the quality of human, physical, social, financial and foreign capital, and thus stand in the way of rapid economic growth. Manna from heaven can be a mixed blessing. Consider the attitudes of individuals to their own and to other people's money. A person's respect for money tends to vary inversely with his or her distance from the effort expended to make the money. For example, loot tends to be invested with less forethought than honest wages. The same argument applies to unrequited foreign aid. An influx of aid tends to increase the real exchange rate, thereby hurting exports as in Fig. 1. Import restrictions exacerbate the appreciation of the currency, hurting exports further. The figure suggests that aid needs to be accompanied by trade liberalization to avoid currency appreciation and its consequences.

Cases

The list of natural-resource-abundant countries beset by economic and political difficulties is a long one. Take Libya. Without its oil export revenues, Libya (population 6 million) would hardly have had the means to purchase 700 military aircraft, submarines and helicopters to pursue the foreign ambitions of Colonel Gaddafi, in power since 1969. In Equatorial Guinea, following oil discoveries, the purchasing power of per capita GDP increased by a factor of six or seven from 1990 to 2005, while life expectancy plunged from 46 years to 42. One child in five dies before reaching its fifth birthday. More than a half of the population of 500.000 lives on less than a dollar a day. President Mbasogo has ruled the country with an iron fist since 1979, usurping the country's oil wealth for himself and his family and cronies. The readiness of the rest of the world to import oil from Equatorial Guinea, and thus to buy stolen goods, is an integral part of the problem because a people's right to its natural resources is a human right proclaimed in primary documents of international law and enshrined in many national constitutions. Article 1 of the International Covenant on Civil and Political Rights states that 'All people may, for their own ends, freely dispose of their natural wealth and resources.' Neither Libya nor Equatorial Guinea exports any manufactures to speak of.

The list of countries afflicted by various symptoms of the Dutch disease could be extended to include Iran, Iraq, Mexico, Nigeria, Russia, Saudi Arabia, Sudan and Venezuela, among several others. Some other countries have managed to avoid such afflictions. A prime example is Norway, where, before the first drop of oil emerged, the oil and gas reserves within Norwegian jurisdiction were defined by law as common property resources, thereby clearly establishing the legal rights of the Norwegian people to the resource rents. On this legal basis, the government has absorbed about 80 per cent of the resource rent over the years, having learnt the hard way in the 1970s to use a relatively small portion of the total to meet current fiscal needs. Most oil revenue is set aside in the state petroleum fund, recently renamed

the pension fund to reflect its intended use. The government laid down economic as well as ethical principles (commandments) to guide the use and exploitation of the oil and gas for the benefit of current and future generations of Norwegians. The main political parties share an understanding that the national economy needs to be shielded from an excessive influx of oil money to avoid overheating and waste. The Central Bank (Norges Bank), which was granted increased independence from the government in 2001, manages the fund (currently around US$400 billion or $85,000 per Norwegian) on behalf of the Ministry of Finance. This arrangement maintains a distance between politicians and the fund. Almost 40 years after discovering their oil, the Norwegians have a smaller central government than Denmark, Finland and Sweden next door.

Norway's tradition of democracy since long before the advent of oil has probably helped immunize the country from the ailments that afflict most other oil- rich nations. Large-scale rent seeking has been averted in Norway, investment performance has been adequate, and the country's education record is excellent. Even so, some (weak) signs of the Dutch disease can be detected, notably sluggish exports and foreign direct investment and the absence of a large, vibrant high-tech manufacturing sector as in Sweden and Finland. Norway's lack of interest in joining the European Unions can also be viewed in this light.

Then there is Botswana. Having managed its diamonds quite well and used the rents to support rapid growth, Botswana has become the richest country in Africa, measured by the purchasing power of per capita GDP. Its rapid growth since 1965 has been accompanied by political stability and a steady advance of democracy. Unlike Sierra Leone's alluvial diamonds, which are easy to mine by shovel and pan, and easy to loot, Botswana's kimberlite diamonds lie deep in the ground and can only be mined with large hydraulic shovels and other sophisticated equipment. They are therefore not very lootable. This difference probably helped Botswana succeed while Sierra Leone failed, and so, most likely, did South African involvement in Botswana's diamond industry.

See Also

▶ Oil and the Macroeconomy

Bibliography

Corden, W.M. 1984. Booming sector and Dutch disease economics: Survey and consolidation. *Oxford Economic Papers* 36: 359–380.

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

Ross, M. 2001. Does oil hinder democracy? *World Politics* 53: 325–361.

Van Wijnbergen, S. 1984. The 'Dutch disease': A disease after all? *Economic Journal* 94: 41–45.

Wenar, L. 2008. Property rights and the resource curse. *Philosophy and Public Affairs* 36: 1–32.

Dutch Disease and Foreign Aid

Christopher Adam

Abstract

Academic and policy debates on aid effectiveness frequently emphasise the vulnerability of recipients to the Dutch Disease, through which aid inflows appreciate the real exchange rate, thereby taxing the tradable export sector with potentially deleterious effects on growth. Fear of the Dutch Disease is remarkably pervasive, even though there is little decisive evidence that aid-induced Dutch Disease effects are either large or widespread amongst poor countries, at least against most plausible counterfactuals. The lack of strong evidence reflects a variety of factors, including problems of measurement, but is primarily due to the fact that aid flows are often purposive – designed to address pre-existing distortions in the recipient economy – and are accompanied by policy measures specifically designed to mitigate latent Dutch Disease effects. Although the conventional macroeconomic transmission

channels may therefore be weak, the language of the Dutch Disease continues to be used as a metaphor for the wide range of political economy concerns associated with aid surges.

Keywords

Aid absorption; Dutch disease; Export growth; Foreign aid; Political economy; Real exchange rate appreciation; Transfer paradox

JEL Classifications

F21; F23; F43

Introduction

An aid-induced 'Dutch Disease' occurs when foreign aid inflows result in a sharp appreciation of the recipient country's real exchange rate, undercutting the international competiveness of its export sector to such an extent that it degrades an important driver of economic growth. In principle, the disease may be so strong as to completely overwhelm the beneficial effects of the aid transfer. Concerns that the Dutch Disease phenomenon, traditionally associated with natural resource windfalls (Gylfason 2008), might also plague foreign aid flows began to emerge in the 1980s and early 1990s when donor support to low-income countries was shifting away from project finance – where the bulk of aid was used to purchase non-competitive project-related imports – towards resource transfers that provided general balance of payments and direct support to government budgets (van Wijnbergen 1984; Younger 1992). Since the late 1990s, when large-scale debt relief initiatives to low-income countries were well under way, the fear of the Dutch Disease effects of aid has become a deeply entrenched theme of the macroeconomic policy discourse between international financial institutions, dofgylnor agencies and aid-receiving countries (for example, Heller 2005).

It would be natural to conclude that such fears are grounded in robust empirical evidence. But this is not the case: although the existence of a Dutch Disease channel is frequently advanced as an important explanation for the weak measured growth impacts of foreign aid flows to poor countries (most notably in Rajan and Subramanian 2011), the empirical evidence tends to suggest that measured aid-induced Dutch Disease effects are neither widespread amongst poor aid-dependent economies nor particularly large (Tarp 2008).

That the Dutch Disease remains so prominent in policy debates is therefore intriguing. At least three possible explanations suggest themselves. The first concerns measurement and argues that while the effect exists and may be serious in a range of settings, the fact that aid surges are rare and relatively modest in scale (certainly compared to many natural resource booms) means that it is difficult to identify the effects in the aggregate data. The second explanation is similar and centres on the counterfactual: in this case the fear of the Dutch Disease is also warranted, but in anticipation donors and recipients put in place mitigation measures that allow aid to be absorbed without triggering the disease. Had they not, the damage would have been large and clearly identifiable in the data. A third explanation is that while the conventional channel may not be particularly strong, the language of the Dutch Disease has been co-opted to describe a wide range of political economy and other pathologies associated with large and rapidly increasing inflows of external assistance to small economies, the sorts of problems emphasised by Peter Bauer and other critics of aid (Dorn 2008).

The remainder of this entry consists of four sections. Section "Aid Flows and the Dutch Disease: The Basic Mechanism" lays out in more detail the basic mechanics of the Dutch Disease effects of aid and section "Evidence" briefly discusses some of the main themes in the empirical evidence. Section "Aid and the Dutch Disease: Why the Dog Doesn't (Often) Bark" considers how the particularities of aid flows modify the standard Dutch Disease story and Section "Conclusions" concludes.

Aid Flows and the Dutch Disease: The Basic Mechanism

The essential nature of the aid-induced Dutch Disease derives from the observation that while the growth and diversification of exports play a central role in the transformation of poor countries, the mechanics of aid absorption may generate a negative feedback from aid to export competitiveness and growth, thereby acting as a brake on development (Bevan 2006). The mechanism works as follows. Foreign aid – which is composed entirely of traded goods or claims on traded goods – augments domestic resources, leaving the economy as a whole better off, at least initially; how much better off depends on how these resources are absorbed and with what consequence.[1]

To focus on the basic macroeconomics of the aid inflow, we abstract from the details of to whom the aid accrues (whether it is the public or private sector) and the form of expenditure (consumption or investment, for example), returning to these distinctions later. The *ex post* current account of the balance of payments must be financed by some combination of aid inflows and the drawdown of private and/or official net foreign assets:

$$CA_t = A_t - KA_t - \Delta Z_t, \tag{1}$$

in which A_t denotes the net aid inflow, KA_t the flow position on the capital account, and ΔZ_t official reserve accumulation. CA_t denotes the current account deficit before aid and can be expressed as the excess of domestic expenditure over national income (inclusive of net factor income), $CA_t = E_t - Y_t = (M_t - X_t)$, where Y_t and E_t denote Gross National Income and aggregate expenditure (both public and private) respectively, M_t. denotes imports and X_t exports. Substituting into (1) we can re-write the external balance as

$$(M_t - X_t) = A_t - KA_t - \Delta Z_t. \tag{2}$$

Equation 2 offers a precise definition of aid absorption as the extent to which the current account deficit (net of factor payments) increases in response to an increase in aid; absorption is thus a measure of the transfer of resources from (foreign) donor to (domestic) recipient used to augment domestic expenditure. Three examples help to isolate the Dutch Disease. In the first case, the aid inflow is not absorbed at all but instead is saved in the form of either public or private foreign asset accumulation. Thus, using d to denote a change, $d(M_t - X_t) = 0$ and $dA_t = dKA_t + d\Delta Z_t$; some fraction of the aid increase is saved on the official side through net international reserve accumulation and the remainder on the private side by capital outflows, either legitimate or not.[2] In this case there is no pressure on the real exchange rate and no Dutch Disease, but only because there is no absorption. This situation is unlikely to occur, however, at least beyond the very short run. Donors are generally averse to this sort of outcome: their motive for providing development assistance is precisely for it to be absorbed and spent in the recipient economy, not to be piled up in offshore bank vaults, even if this might in fact be a prudent macroeconomic response.[3] The second polar case is where aid is fully absorbed – so that net imports, $(M_t - X_t)$, rise dollar-for-dollar with the increase in aid – but where the associated increased expenditure is entirely in terms of non-competitive final consumption imports such as military hardware, Ferraris or foreign travel. In this case $dM_t = dA_t$ and $dX_t = 0$: the aid is fully absorbed, but the

[1] Traded goods, consisting of importables and exportables are those goods produced and consumed in world markets. Domestic demand and supply conditions therefore have no impact on the (world) price of tradables. Non-tradable or domestic goods, on the other hand are only produced locally; their price is determined by domestic market conditions. By definition, donor aid can only consist of traded goods (e.g. food aid) or a claim over traded goods (i.e. a dollar flow of aid).

[2] On the link between aid and (legal and illegal) capital flight see Ndikumana and Boyce (2003).

[3] If aid flows are temporary a high rate of (official) saving may be consistent with efficient expenditure smoothing. But see Buffie et al. (2010) on how 'use it or lose it' constraints on aid flows affect recipients' monetary and fiscal policy choices.

entire demand impulse leaks offshore through 'self-sterilising' expenditures and hence there is no transmission of demand pressures to domestic production.

The more interesting and certainly more common case is the intermediate one in which the higher expenditure facilitated by the aid inflow leads to an increase in the demand for both imports and domestic goods and services: in this case $dA_t = dM_t - dX_t$. Hence the aid is fully absorbed so that net imports still rise dollar-for-dollar with the aid, but absorption is represented by some increase in imports and a *fall* in exports. This is simply the 'spending effect' from the conventional Dutch Disease theory. Viewed from this perspective, the absorption of aid inflows becomes a classical small-country 'transfer problem'. As a price-taker, the increased demand for importable goods can be met via imports at the prevailing world prices. The increased demand for domestic non-tradable goods, on the other hand, can only be satisfied if the domestic supply of non-tradables increases which requires resources, in particular labour, to be bid away from the production of tradables (i.e. exports and import-substitutes). The price of non-tradables relative to tradable goods must therefore rise in order to shift demand in favour of tradable goods and supply in favour of non-tradables. This relative price movement – i.e. the real exchange rate appreciation – is what restores both internal balance (i.e. equilibrates the demand and supply for non-tradable goods) and external balance (i.e. the adjustment of net exports, conditional on any changes in public and private net foreign asset positions) following the aid inflow. How much the real exchange rate needs to appreciate and how large the resource movement will be is case-specific, determined by consumer preferences and firms' production behaviour. The more elastic are demand and supply in response to movements in the real exchange rate the easier it is to shift resources between sectors and the milder the required real exchange rate appreciation (see Adam and Bevan (2006) for a formal derivation).

The distributional consequences of this adjustment are as follows. Producers of tradable goods stand to lose as the purchasing power of export revenues declines relative to the cost of consumption, while profit margins are squeezed by the rising cost of labour and non-tradable inputs. At the margin, firms in this sector go out of business. Producers of non-tradables, on the other hand, gain: and if the production of non-tradables is labour-intensive – as is likely to be the case in many low-income countries – wage earners in aggregate will gain (as suggested by the Stolper–Samuelson Theorem).

Up to this point, the real exchange rate appreciation is an efficient macroeconomic response to the aid inflow, providing the signal for a welfare-maximising reallocation of resources. For this adjustment to be harmful to the recipient *in aggregate* – in other words for there to be a 'disease' – requires something else to be going on. There are at least two standard ways of motivating this. The dominant conventional explanation for the disease is that the tradable sector (typically the manufacturing sector or commercial agriculture) is the source of some positive externality – for example if the export sector is an incubator of economy-wide productivity growth. If so, the aid-induced contraction of the sector, relative to the counterfactual in which aid does not increase, is a socially inefficient tax on growth. It is commonly assumed that the relevant externality resides in a learning-by-doing mechanism that generates dynamic economies of scale in the production of non-traditional exportable goods, whereby productivity growth is increasing in the cumulative output (or exports) of the sector (see Beaudry (2008), Clerides et al. (1998), and Martins and Yang (2009), amongst others).

A second perspective emphasises the short-run volatility of the real exchange rate as opposed to its long-run level in generating Dutch Disease effects (for example, Bulíř and Harmann (2008) and Eifert and Gelb (2005)). Here the driving force is hysteresis, whereby the short-run temporary appreciation of the real exchange rate may have long-run adverse effects on export growth. If credit market imperfections mean affected firms in the export sector are unable to borrow against future expected profits when the real exchange rate appreciation passes, they will be unable to

sustain the losses caused by cheaper imports and rising domestic wage costs. At the margin, firms will exit this sector, and if they face fixed costs of re-entry (in terms of specific marketing relationships or long-term supply-chain relationships, for example) or fall behind global technology frontiers, they will re-engage the export market at a lower level than before, even when the real exchange rate returns to its previous level.

For completeness, it is worth briefly mentioning a third alternative representation of the aid-induced Dutch Disease phenomenon, one that does not appeal to conventional growth externalities. Drawing on the 'transfer paradox' literature originating in the post-First World War debates between Keynes and Ohlin (Brock 2008), Yano and Nugent (1999) argue that for the small open economy, aid flows can reduce aggregate welfare if the tradable goods sector is tariff-distorted and aid transfer takes the form of an increase in the installed capital stock. They provide some (weak) evidence in support of this claim, but as Tokarick (2008) demonstrates, aid flows will only be welfare-reducing in this framework if the non-tradable goods are strongly complementary to the (tariff-protected) imported good. When, as is the conventional case, non-traded and traded goods are substitutes in consumption, aid cannot reduce welfare.

Two important qualifications are relevant at this point. First, the conventional Dutch Disease result does not depend on aid flowing to government but rather on its impact on the aggregate demand for non-tradables (and the presence of growth externalities). Aid typically does accrue to government in the first instance, but it can be spent in a variety of ways: directly by government on current or capital goods (an aid-funded deficit); transferred to the private sector through income transfers; used to retire debt; or to finance tax cuts. As we discuss below, the *form* of public expenditure will play a crucial role in determining whether latent Dutch Disease effects are exacerbated or mitigated.

Second, the foregoing analysis is independent of whether the country adopts a fixed nominal exchange rate regime or a float or, indeed, any hybrid arrangement. What matters, at least over the medium term, is how relative prices change and not how this change is brought about. In a flexible exchange rate regime, the real exchange rate appreciation is typically achieved by an appreciation of the nominal exchange rate; in a fixed exchange rate regime, adjustment occurs through rising domestic non-tradable prices relative to tradable good prices. Over the short run, however, particularly when domestic prices are sticky, the dynamics of the real exchange rate do depend on the nominal exchange rate regime (Ghosh et al. 2003; Fielding and Gibson 2013).

Evidence

Although the empirical question is relatively easy to define, the evidence on whether aid-induced Dutch Disease effects exist is mixed and often highly contested. If the Dutch Disease channel is important, we expect rising aid inflows to be associated with an appreciating real exchange rate in recipient countries and for the non-tradable sector to expand at the expense of the exportable sector. Critically, the contraction of exports should be associated with lower overall growth. This pattern should be present in both time-series and the cross-section or panel evidence.

The identification of this channel is not straightforward, for a number of reasons. First, although there are exceptions – the small islands of the Pacific, some post-conflict countries and a small number of Sub-Saharan African countries – aid flows rarely exceed even 10% of GDP, and large aid surges (the events that allow the econometrician to statistically identify Dutch Disease effects) are comparatively uncommon. Second, unlike natural resource windfalls, aid is rarely exogenous with respect to the recipient's economic performance. Rather, aid is purposive, allocated to countries in poor economic straits, afflicted by structural and policy conditions that themselves generate low growth, over-valued real exchange rates and small tradable sectors. In other words, aid tends to flow into countries displaying the symptoms of the Dutch Disease: controlling for the endogeneity of aid in these circumstances is a formidable challenge. Third, policy actions by

aid recipients, in the short term through fiscal and monetary policy and in the longer term by public investment and other policy reforms, will, if successful, mitigate the incipient Dutch Disease pressures from aid. And finally, as with all research that seeks to assess aid effectiveness on the basis of country-level data, the profound changes in recent decades in the political economy and institutional environment shaping aid relationship further complicate matters: not only have underlying structural conditions changed greatly over the decades, but the geopolitics of aid have fundamentally changed the aid allocation behaviour of donors, particularly since the end of the Cold War.

These points notwithstanding, the research literature does tend to find some evidence supporting the Dutch Disease channel. In particular, there is support for the first two links in the process: aid inflows and aid surges are indeed associated with a tendency for the real exchange rate to appreciate, although rarely is this effect particularly strong (Werker et al. 2009; Magud and Sosa 2010; Fielding and Gibson 2013). Similarly, a number of papers identify a link between aid inflows and the relative size of the exportable sector, the most notable contribution in this field being from Rajan and Subramanian (2011) who exploit the within-country cross-industry variation in growth rates to identify the effect of aid on manufacturing growth. They find quite sizeable effects: using alternative measures to identify exportable sectors, they suggest that an additional one percentage point increase in the share of aid to GDP results in exportable industries growing between 0.5% and 1.4% per annum more slowly than non-exportable industries.[4] By contrast, Werker et al. (2009), using oil price shocks experienced by OPEC donors to identify the exogenous variation in their aid disbursement, find that while *net* imports respond strongly to aid inflows – aid is indeed absorbed – this occurs principally through increased imports, with only a very limited reduction in exports.

[4]These are, of course, relative growth rates; without controlling for the relative size of sectors it is not possible to infer the impact on aggregate output growth.

The empirical evidence suggests that the Dutch Disease operates principally through the misalignment, specifically the overvaluation of the short-run real exchange rate, rather than through the appreciation in the equilibrium real exchange rate itself (Arellano et al. 2009; Elbadawi et al. 2012; Rajan and Subramanian 2011; Kang et al. 2012). This reinforces the argument that whether aid flows are likely to trigger Dutch Disease effects depends on how recipient governments set the relevant macroeconomic, public expenditure and structural policy instruments, a point that is reinforced by country case studies examining specific aid surges. In their study of aid surges in the wake of the low-income country debt-relief initiatives in the early 2000s, Berg et al. (2007, 2010) argue that the noticeable absence of Dutch Disease effects in the wake of these aid surges often reflected conscious decisions not to 'absorb' the aid because of an underlying fear of triggering Dutch Disease effects, even to the point that the lack of absorption risked undermining the developmental objective of the resource transfer.

Aid and the Dutch Disease: Why the Dog Doesn't (Often) Bark

The ambiguity of the empirical evidence hints at why, in practice, the characterisation of the Dutch Disease outlined in the simple model described above is rather incomplete. First, the model assumes a pre-aid equilibrium in which all factors are fully employed so that the increased demand for non-tradables necessarily entails a contraction of tradable production. In the aid context, however, recipient countries are typically characterised by unemployed resources, particularly labour. If these can be brought into use in the production of non-tradables as demand increases, aid can then be absorbed without driving up real wages and drawing resources away from the tradable/exportable sector. A failure to recognise idle capacity may thus lead to a systematic 'over-expectation' that aid will induce a Dutch Disease problem (Nkusu 2004). The key question, then, is how much effective excess capacity actually exists: measured unemployment may not be a sufficient statistic if

non-tradable production is intensive in specific factors such as skilled labour that are not easily substituted for the abundant unemployed factors.

This highlights the second key feature of aid inflows, noted above: that aid is intentionally targeted at countries that are not just poor but often heavily constrained by distortions and bottlenecks, so that the economy operates at less than full capacity given its factor endowments. Moreover, aid tends to be conditioned on specific policy actions, directed towards increasing the supply of key inputs to production, either by the provision of imports or through technical assistance, for example, and is bundled with specific conditionality aimed at removing policy distortions or institutional bottlenecks. Although conditionality is not always effective, the blending of resources with policy reform and expertise can allow reform-oriented countries to absorb aid flows and simultaneously avoid the trade-off implied by the Dutch Disease. This was certainly the case across a number of countries in the late 1990s and early 2000s, when aid-supported macroeconomic stabilisation and supply-side reforms combined to generate both rising domestic consumption and rising net exports as aid helped previously heavily distorted economies to decompress and grow rapidly: the cases of Ghana and Uganda in these years stand as good examples of this process.

Clearly, there are limits to decompression. How long countries can dodge the bullet of the Dutch Disease depends both on the depth of reform but equally on how aid supports the expansion of the supply side of the economy. In the short run, government expenditure patterns that are biased towards imports can mitigate pressures (albeit at the cost of potentially distorting otherwise efficient public spending programmes) while monetary policy can be geared to matching the path of public spending to the rate of aid absorption so as to minimise excess real exchange rate appreciation (Adam et al. 2009).

But the obvious mechanism to mitigate or reverse Dutch Disease effects of aid over the medium term is to expand aggregate supply through higher public and/or private investment. One channel for this is if aid is used to reduce domestic taxation and borrowing so that private investment is crowded-in through higher net of tax returns and lower domestic interest rates. The alternative channel is via public investment. Adam and Bevan (2006), amongst others,[5] explore the interplay between two dynamic externalities in the aid-dependent economy. On the one hand is the demand-side Dutch Disease channel, in which a short-run exchange rate appreciation is associated with the contraction of the exportable sector through a learning-by-doing externality, and on the other a growth externality coming from public infrastructure investment which delivers increasing returns to private factors of production. The precise outcome clearly depends on the relative strength and timing of these offsetting effects – for example Dutch Disease effects may dominate if learning-by-doing effects are particularly strong, while the returns to public investment are small and/or take a long time to realise – but it also depends on the sector-intensity of returns to public investment. If the productivity-enhancing effects of public investment are skewed in favour of production of tradables, the real exchange rate may still appreciate, as the spending effect is reinforced by a resource movement *into* the now more productive exportable sector, but the higher productivity of the sector offsets the effects of the appreciation. Thus the aid inflow is still associated with an appreciation of the real exchange rate, but in this instance with an *expansion* rather than a contraction of the exportable sector. Alternatively, if productivity gains are skewed in favour of the non-tradable sector, expansion of the exportable sector is driven by the *falling* relative price of non-tradables, so that we observe aid inflows associated, over the medium term, with a depreciating real exchange rate and an expanding exportable sector. Adam and Bevan (2006) show that these effects are magnified when the non-tradable good is the principal wage good in the economy. A specific example of such a process might be public

[5]Very similar models appear in Torvik (2001), Agénor et al. (2008), and Berg et al. (2010).

investment in the road network, which lowers the cost of transporting (non-tradable) food to urban areas, thus helping to drive down the cost of manufactured goods (Gollin and Rogerson 2010).

Conclusions

On the basis of the empirical evidence, the conventionally defined Dutch Disease effects of aid clearly does exist. It is not merely a 'theoretical quirk', but at the same time even the most robust evidence cannot point to particularly large effects. Indeed, well-designed aid programmes, combined with appropriate macroeconomic and supply-side policy responses, may be associated with exactly the opposite effect, with aid supporting increased investment and export-led growth – what Berg et al. (2010) refer to as 'Dutch Vigour'.

But this leaves open the question as to why concerns about the Dutch Disease effects of aid remain so firmly on the table. The most persuasive explanation is that the language of the Dutch Disease – the idea that an unrequited transfer may be welfare-reducing – continues to be commonly used as a metaphor for the wide range of political-economy concerns associated with aid surges. These may include dysfunctional rent-seeking behaviour, in which productive resources and entrepreneurial talent are devoted to the capture and distribution of rents from aid contracts (see Klein and Harford (2005), amongst others). They may also involve skilled personnel within the public sector being increasingly deployed on the management of often highly bureaucratic aid programmes (Brautigam and Knack 2004). And on the political side, the 'Dutch Disease' label is frequently used to describe the adverse effects that visible and vocal aid donors may have on domestic systems of political accountability – where politicians and bureaucrats may feel more accountable to the donors who finance the lion's share of the budget than to the domestic electorate – what Peter Bauer (1972) referred to as the 'politicization of daily life' in aid-receiving countries. It is predominantly these factors, rather than the pure macroeconomics of aid, that keep the Dutch Disease centre stage.

See Also

- ▶ Dutch Disease
- ▶ Foreign Aid

Acknowledgments I thank Radhika Goyal for excellent research assistance and the Oppenheimer Fund of the Department of International Development, University of Oxford for financial support.

Bibliography

Adam, C., and D. Bevan. 2006. Aid and the supply side: Public investment, export performance and the Dutch Disease in low-income countries. *World Bank Economic Review* 20(2): 261–290.

Adam, C., E. Buffie, S. O'Connell, and C. Pattillo. 2009. Monetary policy rules for managing aid surges in Africa. *Review of Development Economics* 13(3): 464–490.

Agénor, P., N. Bayraktar, and K. El Aynaoui. 2008. Roads out of poverty? Assessing the links between aid, public investment, growth and poverty reduction. *Journal of Development Economics* 86(2): 277–295.

Arellano, C., A. Bulíř, T. Lane, and L. Lipschitz. 2009. The dynamic implications of foreign aid and its variability. *Journal of Development Economics* 88(1): 87–102.

Bauer, P.T. 1972. *Dissent on development*. Boston: Harvard University Press.

Beaudry, P. 2008. Growth and learning-by-doing. *The new palgrave dictionary of economics*. 2nd edn. Basingstoke: Palgrave.

Berg, A., Aiyar, S., Hussain, M., Roache, S., Mirzoev, T. and Mahone, A. 2007. The macroeconomics of scaling up aid: Lessons from recent experience. IMF Occasional Paper No. 253.

Berg, A., Gottschalk, J., Portillo, R. and Zanna, L.-F. 2010. The macroeconomics of medium-term aid scaling up scenarios. IMF Working Paper 10/160.

Bevan, D. 2006. An analytical overview of aid absorption. In *The macroeconomic management of foreign aid: Opportunities and pitfalls*, ed. P. Isard, L. Lipschitz, A. Mourmouras, and B. Yontcheva. Washington, DC: International Monetary Fund.

Brautigam, D., and S. Knack. 2004. Foreign aid, institutions, and governance in Sub-Saharan Africa. *Economic Development and Cultural Change* 52(2): 255–286.

Brock, P. 2008. Transfer problem. *The new palgrave dictionary of economics*. 2nd edn. Basingstoke: Palgrave.

Buffie, E., S. O'Connell, and C. Adam. 2010. Fiscal inertia, donor credibility and the monetary management of aid

surges. *Journal of Development Economics* 93(2): 287–298.
Bulíř, A., and J. Hamann. 2008. Volatility of development aid: From the frying pan into the fire. *World Development* 36(10): 2048–2066.
Clerides, S., S. Lach, and J. Tybout. 1998. Is learning by exporting important? Micro-dynamic evidence from Colombia, Mexico and Morocco. *Quarterly Journal of Economics* 113(3): 903–947.
Dorn, J. A. 2008. Peter Bauer. *The new palgrave dictionary of economics*, 2nd edn. Basingstoke: Palgrave.
Eifert, B. and Gelb, A. 2005. Improving the dynamics of aid: Toward more predictable budget support. World Bank Policy Research Working Paper 3732.
Elbadawi, I., L. Kaltani, and R. Soto. 2012. Aid, real exchange rate misalignment and economic performance in Sub Saharan Africa. *World Development* 40(4): 681–700.
Fielding, D., and F. Gibson. 2013. Aid and Dutch disease in Sub-Saharan Africa. *Journal of African Economies* 22(1): 1–21.
Ghosh, A., A.-M. Gulde, and H. Wolf. 2003. *Exchange rate regimes: Choices and consequences*. Cambridge, MA: MIT Press.
Gollin, D. and Rogerson, R. 2010. Agriculture, roads and economic development in Uganda. NBER Working Paper 15863.
Gylfason, T. 2008. Dutch disease. *The new palgrave dictionary of economics*, 2nd edn. Basingstoke: Palgrave.
Heller, P. S. 2005. Pity the finance minister: Issues in managing a substantial scaling up of aid flows. IMF Working Paper 05/180.
Kang, J., A. Prati, and A. Rebucci. 2012. Aid, exports and growth: A time series perspective on the Dutch Disease hypothesis. *Review of Economics and Institutions* 3(2).
Klein, M., and T. Harford. 2005. *The market for aid*. Washington, DC: World Bank Publications.
Magud, N. and Sosa, S. 2010. When and why worry about real exchange rate appreciation: The missing link between Dutch Disease and growth. IMF Working Paper 10/271.
Martins, P., and Y. Yang. 2009. The impact of exporting on firm productivity: A metaanalysis of the learning-by-doing hypothesis. *Review of World Economics* 145(3): 431–445.
Ndikumana, L., and J.K. Boyce. 2003. Public debts and private assets: Explaining capital flight from Sub-Saharan African countries. *World Development* 31(1): 107–130.
Nkusu, M. 2004. Aid and the Dutch Disease in low income countries: Informed diagnosis for prudent prognosis. IMF Working Paper 04/49.
Rajan, R., and S. Subramanian. 2011. Aid, Dutch Disease and manufacturing growth. *Journal of Development Economics* 94(1): 106–118.
Tarp, F. 2008. Foreign aid. *The new palgrave dictionary of economics*, 2nd edn. Basingstoke: Palgrave.
Tokarick, S. 2008. Welfare-worsening aid flows to small countries: The role of non-traded goods. *Review of Development Economics* 12(4): 818–827.
Torvik, R. 2001. Learning-by-doing and the Dutch Disease. *European Economic Review* 5: 285–306.
Van Wijnbergen, S. 1984. The 'Dutch Disease': A disease after all? *Economic Journal* 84(373): 41–55.
Werker, E., F. Ahmed, and C. Cohen. 2009. How is foreign aid spent? Evidence from a natural experiment. *American Economic Journal: Macroeconomics* 1(2): 225–244.
Yano, M., and J. Nugent. 1999. Aid, nontraded goods and the transfer paradox in small countries. *American Economic Review* 89(3): 431–449.
Younger, S. 1992. Aid and the Dutch Disease: Macroeconomic management when everybody loves you. *World Development* 20(11): 1587–1597.

Dynamic Models with Non-clearing Markets

Jean-Pascal Bénassy

Abstract

This article studies a new class of models which synthesize the two traditions of general equilibrium with non-clearing markets and imperfect competition on the one hand, and dynamic stochastic general equilibrium (DSGE) models on the other hand. This line of models has become a central paradigm of modern macroeconomics for at least three reasons: (*a*) it displays solid microeconomic foundations, (*b*) it is a highly synthetic theory, which combines in a unified framework general equilibrium, non-clearing markets, imperfect competition, growth theory and rational expectations, and (*c*) it is also an empirical success, leading to substantial progress towards matching real world statistics.

Keywords

Budget constraints; Dynamic models with non-clearing markets; Dynamic stochastic general equilibrium (DSGE) models; Efficiency wages; Fixprice macroeconomic model; General equilibrium; Growth theory; Implicit contracts; IS–LM model; Keynesianism; Market clearing; Menu costs; Microfoundations; Monetary shocks; Monopolistic competition; Nominal rigidities; Non-clearing markets in general equilibrium; Objective

demand curve; Phillips curve; Quantity constraint; Quantity signals; Rational expectations; Real business cycles; State dependent price rigidities; Sticky prices; Technological shocks; Time-dependent contracts; Walrasian equilibrium

JEL Classifications
D5

This article studies a new class of models which synthesize the two traditions of general equilibrium with non-clearing markets and imperfect competition on the one hand, and dynamic stochastic general equilibrium (DSGE) models on the other hand. Although this line of models is still recent, it has clearly become in a short time a central paradigm of modern macroeconomics. The reasons are at least threefold.

The first is that it displays solid microeconomic foundations. This is quite natural since from the two constituent fields above this one inherited a strong general equilibrium framework where all agents (households or firms) maximize their respective objectives subject to well defined constraints.

The second is that it is a highly synthetic theory, which combines in a unified framework general equilibrium, non-clearing markets, imperfect competition, growth theory and rational expectations, so that it can appeal to macroeconomists with very different backgrounds.

The third reason is empirical. A key motivation for DSGE models is to compare the 'statistics' generated by these models with the real-world ones. In that respect the addition of non-clearing markets and imperfect competition has led to substantial progress in matching these statistics, and this has certainly been an important factor in the success of these models.

Now such a wide synthesis did not come all at once. So we begin by recalling briefly a little bit of history and some of the antecedents of the field.

We then present a series of models with explicit solutions. These will demonstrate analytically how the introduction of non-clearing markets allows us to substantially improve the ability of DSGE to reproduce a number of macroeconomic facts.

History

Early Times

At the time when many of the developments leading to these models were initiated, there was a profound split between microeconomics and macroeconomics. On the one hand microeconomics, in its general equilibrium version, was dominated by Walras's (1874) model, as developed by Arrow and Debreu (1954), Arrow (1963), and Debreu (1959). In these models all adjustments are carried out via fully flexible prices, and agents never experience any quantity constraint. On the other hand in the standard macroeconomic model in the Keynes (1936) and Hicks (1937) tradition, as exemplified by the IS–LM model, there are price and wage rigidities, unemployment is present and most adjustments are carried out through variations in real income, a quantity, not a price.

Confronted with this inconsistency, the strategies of macroeconomists turned out to be quite diverse and they took two different routes.

General Equilibrium with Non-clearing Markets

On the one hand, a first set of authors aimed at achieving a synthesis between the then existing microeconomics and macroeconomics. This was achieved by generalizing the traditional general equilibrium model, by introducing non-clearing markets, introducing quantity signals into demand and supply functions, and endogenizing prices in a framework of imperfect competition.

Patinkin (1956) and Clower (1965) showed that the presence of quantity constraints in non-clearing markets would drastically modify the demands for labour and goods, an insight further emphasized by Leijonhufvud (1968). Barro and Grossman (1971, 1976) combined these insights into a fixprice macromodel. Drèze (1975) and Bénassy (1975, 1982) constructed full general equilibrium concepts with price rigidities, where price movements are partially replaced by endogenous quantity constraints. Bénassy (1976)

linked these concepts with general equilibrium under imperfect competition à la Negishi (1961). This link was furthered with the construction of a full general equilibrium concept of objective demand curve based on quantity constraints (Bénassy 1988; see also Gabszewicz and Vial 1972, for a Cournotian view). All these developments are reviewed in the dictionary entry 'non clearing markets in general equilibrium'.

Dynamic Market Clearing Macroeconomics

A second set of authors achieved consistency between microeconomics and macroeconomics by importing into macroeconomics the basic assumption of the then dominant general equilibrium microeconomic models, market clearing. At the same time they paid strong attention to the issues of dynamics and expectations. A central part of these developments was the use of 'rational expectations' in the sense of Muth (1961). This was an important addition, as in the Keynesian system it was sometimes difficult to disentangle the results due to price or wage rigidity from those due to incorrect expectations. Rational expectations allowed the suppression of the second type of results. It appeared also that, even with rational expectations and market clearing, it was possible to build rigorous models displaying fluctuations (Lucas 1972; Kydland and Prescott 1982; Long and Plosser 1983).

Non-Walrasian Cycles

Starting in the mid-1980s authors began combining elements of the two paradigms described above, achieving the synthesis that is the subject of this article. Svensson (1986) studies a dynamic stochastic general equilibrium monetary economy subject to supply and demand shocks. Prices are preset one period in advance by monopolistically competitive firms, so we have both imperfect competition and sticky prices. Because of price presetting the model has multiple regimes.

Various types of rigidities have been then introduced in dynamic models, leading to different patterns of cycles. Andersen (1994) reviews various causes and consequences of price and wage rigidities.

A first type of rigidities is 'real' rigidities, which create an endogenous noncompetitive wedge between various prices. As an example, monopolistic competition à la Dixit and Stiglitz (1977) introduces a markup between marginal cost and price. In this class Danthine and Donaldson (1990) introduce efficiency wages, Danthine and Donaldson (1991, 1992) introduce implicit contracts in the vein of Azariadis (1975), Baily (1974) and Gordon (1974). Rotemberg and Woodford (1992, 1995) study imperfect competition.

Models with nominal rigidities study situations where the nominal prices themselves (and not relative prices) are sluggish. Several devices have been used. The first, following the early works on wage and price contracts by Gray (1976), Fischer (1977), Phelps and Taylor (1977), Taylor (1979, 1980) and Calvo (1983), assumes that there is a system of contracts expiring at deterministic or stochastic dates. For that reason they are called 'time dependent'. Such contracts have been integrated in DSGE models by Cho (1993), Cho and Cooley (1995), Bénassy (1995, 2002, 2003a, b), Yun (1996), Cho et al. (1997), Andersen (1998), Jeanne (1998), Ascari (2000), Chari et al. (2000), Collard and Ertz (2000), Ascari and Rankin (2002), Huang and Liu (2002), Smets and Wouters (2003) and Christiano et al. (2005), to name only a few.

Another type of price rigidity, called 'state dependent', is based on costs of changing prices. Two specifications are favourite in the literature: quadratic costs of changing prices (Rotemberg 1982a, b), which have been implemented, for example, in Hairault and Portier (1993), and fixed costs of changing prices (Barro 1972), often renamed 'menu costs'. Clearly these costs should be interpreted as surrogates for other unspecified causes, and identifying these causes is a challenge that faces this line of research.

Now most of the contributions of this field are based on numerical evaluations of various models. So we present next a number of models with explicit solutions which will make clear why this line of models has been successful in solving problems that were difficult to solve in market-clearing models.

An Analytical Illustration

We shall now show in this section in a series of explicitly solved models how the introduction of nominal rigidities in DSGE models allows to considerably improve the capacities of these models to reproduce the dynamic evolutions of actual economies.

We first present a basic model and compute as a reference its Walrasian equilibrium and dynamics. Then we introduce a first nominal rigidity, one-period wage contracts. This improves some correlations, but cannot create strong persistence as in reality. We next introduce multi-periodic wage contracts, and show that this allows us to obtain a persistent response of output to demand shocks. Finally, simultaneous rigidities of wages and prices are considered, and we show that one can obtain in this way with fairly realistic values of the parameters a persistent and hump-shaped response of both output and inflation.

The Basic Model

We study a dynamic monetary economy à la Sidrauski (1967) and Brock (1975), where goods are exchanged against money at the (average) price P_t and work against money at the (average) wage W_t. There are two types of agents: households and firms. Firms have a simple technology:

$$Y_t = Z_t N_t^{\alpha} \tag{1}$$

where N_t is the quantity of labour used by firms and Z_t a technological shock common to all firms. Note that we do not introduce capital in this model. Because its rate of depreciation is low, it would not add much to our argument, and would substantially complicate the results and exposition.

The representative household works N_t, consumes C_t, and ends period t with a quantity of money M_t. It maximizes the expectation of its discounted utility:

$$U = E_0 \sum_{t=0}^{\infty} \beta^t \left[\log C_t + \omega \log \frac{M_t}{P_t} - \xi \frac{N_t^{\nu}}{\nu} \right]. \tag{2}$$

At the beginning of period t the household faces a monetary shock à la Lucas (1972), whereby the quantity of money M_{t-1} coming from $t-1$ is multiplied by μ_t, so that its budget constraint for period t is:

$$C_t + \frac{M_t}{P_t} = \frac{W_t}{P_t} N_t + \frac{\mu_t M_{t-1}}{P_t}. \tag{3}$$

There are thus two shocks in this economy, the technology shock Z_t and the monetary shock $\mu_t = M_t/M_{t-1}$. As an illustration we shall use below the following traditional processes (in all that follows lower-case letters represent the logarithm of the variable represented by the corresponding uppercase letter):

$$m_t - m_{t-1} = \frac{\varepsilon_{mt}}{1-\rho L} z_t = \frac{\varepsilon_{zt}}{1-\phi L} \tag{4}$$

where ε_{zt} and ε_{m_t}, the innovations in z_t and m_t, are uncorrelated white noises with:

$$\mathrm{var}(\varepsilon_{zt}) = \sigma_z^2 \mathrm{var}(\varepsilon_{mt}) = \sigma_m^2 \tag{5}$$

Walrasian Dynamics

As a benchmark we shall study here the case where both labour and goods markets are in Walrasian equilibrium in each period, as in the first traditional real business cycle (RBC) models, and we shall see how this economy reacts to technological and monetary shocks. Solving the model we find that money holdings are a multiple of consumption:

$$\frac{M_t}{P_t C_t} = \frac{\omega}{1-\beta} \tag{6}$$

and that employment N_t is constant:

$$N_t = N = (\alpha/\xi)^{1/\nu}. \tag{7}$$

Using (1) and (7) we find (we eliminate some irrelevant constant terms):

$$n_t = n y_t = z_t + \alpha n w_t - p_t = y_t - n. \tag{8}$$

Although we will not do any real calibration in this article, we can note at this stage a few issues

that posed a problem to researchers in the RBC domain.

First, real wages are much too pro-cyclical in this Walrasian model. From (8) we see that the real wage–output correlation is equal to 1. Even though this correlation is lower than 1 in calibrated models where N_t varies, it is always quite above what is observed in real economies.

A second problem concerns the inflation–output correlation, a problem related to the literature on the Phillips curve. Whereas it is generally considered that this correlation is positive, the above Walrasian model yields a negative correlation:

$$\mathrm{cov}(\Delta p_t, y_t) = -\frac{\sigma_z^2}{1+\phi} < 0. \tag{9}$$

Finally, an important and recurrent critique of RBC-type models has been that they do not generate any internal propagation mechanism, and that the only persistence in output movements is that already present in the exogenous process of technological shocks z_t (see, for example, Cogley and Nason 1993, 1995). This appears here in Eq. (8), where the dynamics of output y_t is exactly the same as that of the technological shock z_t.

We shall now introduce wage contracts, first lasting one period, and then multiperiod overlapping contracts, and we shall see that the above problems find a natural solution in this framework.

Single-Period Wage Contracts

Let us thus assume (Bénassy 1995, and Bénassy 2002, for microfoundations), that the wages are predetermined at the beginning of each period at the expected value of the Walrasian wage (in logarithms), and that at this contractual wage the households supply the quantity of work demanded by firms (this type of contract was introduced by Gray 1976).

Combining (6) and $C_t = Y_t$ we find that the Walrasian wage w_t^* is, up to an unimportant constant, equal to m_t, so that the preset wage w_t is given by:

$$w_t = E_{t-1} w_t^* = E_{t-1} m_t \tag{10}$$

where $E_{t\ -1}m_t$ is the expectation of m_t formed at the beginning of period t, before shocks are known.

The difference with the Walrasian case is that employment N_t is now variable and demand determined. Equations (8) become:

$$y_t = z_t + \alpha n_t w_t - p_t = y_t - n_t \tag{11}$$

while $n_t = n$ is replaced by (10). So we first obtain the level of employment in period t:

$$n_t = n + m_t - E_{t-1} m_t = n + \varepsilon_{mt} \tag{12}$$

since $m_t - E_{t-1}m_t = \varepsilon_{mt}$. Contrarily to what happened in the Walrasian version of the model, unanticipated monetary shocks now have an impact on the level of employment, and therefore output. We shall now use the preceding formulas to show that the hypothesis of preset wages allows to substantially improve some correlations relative to the Walrasian model.

Let us start with the real wage which, in the Walrasian model, has a much too high positive correlation with output. Let us combine (11) and (12), to obtain the values of output and real wage:

$$y_t = z_t + \alpha \varepsilon_{mt} w_t - p_t = z_t - (1-\alpha)\varepsilon_{mt}. \tag{13}$$

We see that supply shocks create a positive correlation between the real wage and output. However, monetary shocks create a negative correlation. Our model thus allows us to combine this last characteristic, typical of traditional Keynesians models, with the usual results of RBC models. If one considers the technological and monetary shocks (4), one obtains the following correlation:

$$\mathrm{corr}(w_t - p_t, y_t) = \frac{\sigma_z^2 - (1-\phi^2)\alpha(1-\alpha)\sigma_m^2}{\left[(\sigma_z^2 + (1-\phi^2)\alpha^2\sigma_m^2)\right]^{1/2}\left[\sigma_z^2 + (1-\phi^2)(1-\alpha)^2\sigma_m^2\right]^{1/2}}. \tag{14}$$

We see that the real-wage–output correlation is equal to 1 if there are *only* technological shocks. But this correlation diminishes as soon as there are

monetary shocks, and it can even become negative. One can thus reproduce the correlations observed in reality by adequate combinations of technological and monetary shocks.

Let us now consider the relation between inflation and output, which are generally considered to be positively correlated, at least in Keynesian tradition. If we assume again the monetary and technological shocks (4), we find:

$$\text{Covariance}(\Delta p_t, y_t) = \alpha(1-\alpha)\sigma_m^2 - \frac{\sigma_z^2}{1+\phi}. \tag{15}$$

Formula (15) shows us that the positive covariance (and thus correlation) between inflation and output is linked to the presence of demand shocks, and that the sign of this correlation may change if there are sufficiently strong technological shocks.

So we just saw that one-period contracts allow us to improve some important correlations. We now naturally ask a question already posed for the standard RBC model: is the response to shocks, and in particular to demand shocks, sufficiently persistent? Let us recall Eq. (13):

$$y_t = z_t + \alpha\varepsilon_{mt}. \tag{16}$$

We see that monetary shocks now have an immediate effect on output (and employment), but that, starting with the second period, the effect of these shocks is completely dampened. One-period contracts allow us to solve the puzzle raised by some correlations, but certainly not the persistence problem. We shall see in the next two sections that multi-periodic contracts allow us to solve that problem.

Multi-periodic Wage Contracts

The models that we have examined so far share with traditional RBC models the defect of having an extremely weak internal propagation mechanism. In particular, the response of output to monetary demand shocks is almost entirely transitory. But several empirical studies (see, for example, Christiano et al. 1999, 2005) have pointed out that in reality the response to monetary shocks not only was persistent but also had a hump-shaped response function. We shall now introduce multi-periodic wage contracts in rigorous stochastic dynamic models, and show that they allow us to reproduce these features. Models with such multi-periodic wage or price contracts have been studied notably by Yun (1996), Andersen (1998), Jeanne (1998), Ascari (2000), Chari et al. (2000), Collard and Ertz (2000), and Bénassy (2002, 2003a, b).

In order to make our demonstration analytically, we use a contract, inspired by Calvo (1983) and developed in Bénassy (2002, 2003a), which has three advantages: (*a*) the average duration of contracts can take any value from zero to infinity, (*b*) an analytical solution can be found with both wage and price contracts, and (*c*) it has explicit microfoundations.

In this framework in each period s a contract is made for wages at period $t \geq s$. As in the Gray contract, the contract wage is the expectation of the market-clearing wage in period t. So if we denote as x_{st} the contract wage made in s for period t:

$$x_{st} = E_s\left(w_t^*\right). \tag{17}$$

Now, as in Calvo (1983), each wage contract has a probability γ to stay unchanged, and a probability $1-\gamma$ to be broken. If the contract is broken, a new contract is immediately renegotiated on the basis of current period information. So for $\gamma = 0$, wages are totally flexible, for $\gamma = 1$ they are totally rigid.

It is easy to compute the average duration of these contracts. The probability for a contract to be still valid j periods after the date it was concluded is equal to $(1-\gamma)\gamma^j$. The expected duration of the contract is thus:

$$\sum_{j=0}^{\infty}(1-\gamma)j\gamma^j = \frac{\gamma}{1-\gamma}. \tag{18}$$

We thus see that varying γ from 0 to 1 the average duration of the contract varies from zero to infinity.

The average wage w_t is the mean of past x_{st}'s weighted by the probability for the corresponding contract to be still in effect. Because of the law of large numbers, and since the probability of survival of wage contracts is γ, the proportion of contracts coming from period $s \leq t$ is $(1-\gamma)\gamma^{t-s}$. Therefore, the average wage in the economy is given by:

$$w_t = (1-\gamma)\sum_{s=-\infty}^{t} \gamma^{t-s} x_{st}. \tag{19}$$

If we now solve the model with the shocks (4) we find that the dynamics of employment is characterized by (Bénassy 2002, 2003a):

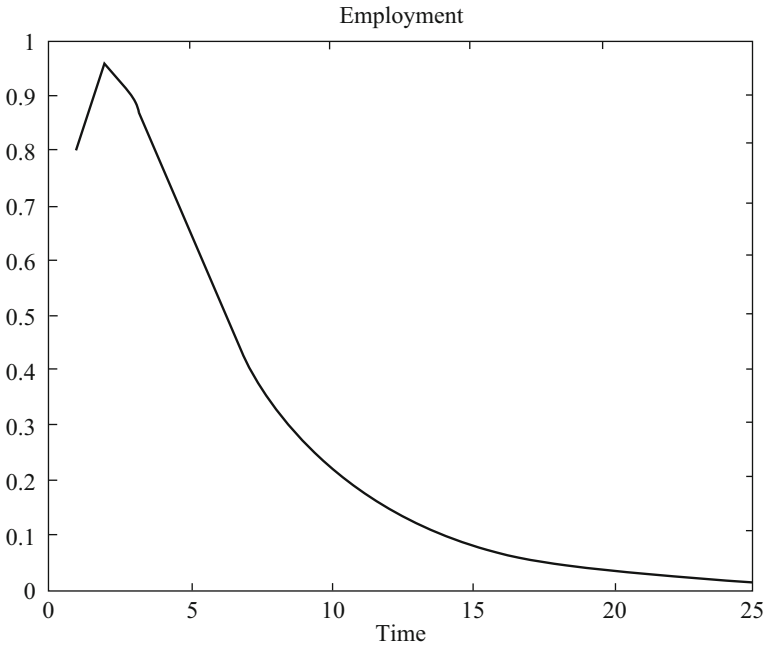
$$n_t = n + \frac{\gamma \varepsilon_{mt}}{(1-\gamma L)(1-\gamma\rho L)} \tag{20}$$

where L is the lag operator: $L^j X_t = X_{t-j}$. The response of output is deduced from that of employment through:

$$y_t = \alpha n_t + z_t. \tag{21}$$

Formula (20) shows clearly that, contrarily to the case of one-period contracts, the response to a monetary shock can be quite persistent. We can have an idea of the temporal profile of this response by computing the response function of output and employment to a monetary shock. The value of ρ most often found in the literature is $\rho = 0{:}5$. As for γ, we saw above (formula 18) that the average duration of wage contracts is equal to $\gamma/(1-\gamma)$. One considers generally that the average duration of wage contracts is about one year (see, for example, Taylor 1999), which corresponds to $\gamma = 4/5$. Figure 1 shows the response of employment (output is derived via 21) to a monetary shock for $\gamma = 4/5$.

We see that the response function displays persistence in the effects of monetary shocks, and has even a hump-shaped response. If we plot, however, the response function of inflation, we find that it is steadily decreasing after the initial jump, whereas it seems to have a delayed hump-shaped response in reality.



Dynamic Models with Non-clearing Markets, Fig. 1

Wage and Price Multi-periodic Contracts

We shall now enlarge our model by considering simultaneously wage and price multi-periodic contracts (see Bénassy 2003b, for such a model with explicit microfoundations). Numerically solved models with both wage and price multi-periodic contracts are found in Christiano et al. (2005), Huang and Liu (2002), Smets and Wouters (2003).

Wage contracts are exactly the same as in the preceding section: each contract is maintained with probability γ, or renegotiated with probability $1 - \gamma$. Symmetrically, price contracts are maintained with probability φ, or break down and are renegotiated with probability $1 - \varphi$. The average price p_t is given by:

$$p_t = (1 - \varphi) \sum_{s=-\infty}^{t} \varphi^{t-s} q_{st} \tag{22}$$

where q_{st} is the price contract negotiated in period s for period t. Using again the shock processes (4), and taking $v = 1$, we find the following dynamics for output and inflation:

$$\begin{aligned} y_t = z_t &- \frac{\varphi \varepsilon_{zt}}{1 - \varphi \phi L} + \frac{\alpha \gamma \varepsilon_{mt}}{(1 - \gamma L)(1 - \gamma \rho L)} \\ &+ \frac{\varphi \varepsilon_{mt}}{(1 - \varphi L)(1 - \varphi \rho L)} \\ &- \frac{\alpha \gamma \varphi \varepsilon_{mt}}{(1 - \gamma \varphi L)(1 - \gamma \varphi \rho L)} \end{aligned} \tag{23}$$

$$\pi_t = (1 - L) p_t = (1 - L)(m_t - y_t). \tag{24}$$

As in the preceding section we take as an illustration $\alpha = 2/3$, $\rho = 1/2$ and $\gamma = 4/5$ (one-year wage contracts). As for prices, we want to take a rather low duration of contracts, so we shall take $\varphi = 1/2$ (one quarter). Simulations show that in that case we obtain a persistent and hump-shaped response for both output and inflation.

So we see that with only reasonable nominal rigidities we obtain some realistic response functions. Clearly the adjunction of 'real' rigidities would allow to reproduce even better the actual dynamic macroeconomic patterns.

See Also

- ▶ Non-clearing Markets in General Equilibrium
- ▶ Real Business Cycles

Bibliography

Andersen, T. 1994. *Price rigidity: Causes and macroeconomic implications*. Oxford: Oxford University Press.

Andersen, T. 1998. Persistency in sticky price models. *European Economic Review* 42: 593–603.

Arrow, K. 1963. The role of securities in the optimal allocation of risk–bearing. *Review of Economic Studies* 31: 91–96.

Arrow, K., and G. Debreu. 1954. Existence of an equilibrium for a competitive economy. *Econometrica* 22: 265–290.

Ascari, G. 2000. Optimising agents, staggered wages and persistence in the real effects of money shocks. *Economic Journal* 110: 664–686.

Ascari, G., and N. Rankin. 2002. Staggered wages and output dynamics under disinflation. *Journal of Economic Dynamics and Control* 26: 653–680.

Azariadis, C. 1975. Implicit contracts and underemployment equilibria. *Journal of Political Economy* 83: 1183–1202.

Baily, M. 1974. Wages and employment under uncertain demand. *Review of Economic Studies* 41: 37–50.

Barro, R. 1972. A theory of monopolistic price adjustment. *Review of Economic Studies* 39: 17–26.

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

Barro, R., and H. Grossman. 1976. *Money, employment and inflation*. Cambridge: Cambridge University Press.

Bénassy, J.-P. 1975. Neo-Keynesian disequilibrium theory in a monetary economy. *Review of Economic Studies* 42: 503–523.

Bénassy, J.-P. 1976. The disequilibrium approach to monopolistic price setting and general monopolistic equilibrium. *Review of Economic Studies* 43: 69–81.

Bénassy, J.-P. 1977. A neoKeynesian model of price and quantity determination in disequilibrium. In *Equilibrium and disequilibrium in economic theory*, ed. G. Schwödiauer. Boston: Reidel Publishing Company.

Bénassy, J.-P. 1982. *The economics of market disequilibrium*. New York: Academic.

Bénassy, J.-P. 1988. The objective demand curve in general equilibrium with price makers. *Economic Journal* 98(Suppl): 37–49.

Bénassy, J.-P. 1990. Non-Walrasian equilibria, money and macroeconomics. In *Handbook of monetary economics*, ed. B. Friedman and F. Hahn. Amsterdam: North-Holland.

Bénassy, J.-P. 1995. Money and wage contracts in an optimizing model of the business cycle. *Journal of Monetary Economics* 35: 303–315.

D

Bénassy, J.-P. 2002. *The macroeconomics of imperfect competition and nonclearing markets: A dynamic general equilibrium approach*. Cambridge, MA: M.I.T. Press.

Bénassy, J.-P. 2003a. Staggered contracts and persistence: microeconomic foundations and macroeconomic dynamics. *Recherches Economiques de Louvain* 69: 125–144.

Bénassy, J.-P. 2003b. Output and inflation persistence under price and wage staggering: Analytical results. *Annales d'Economie et de Statistique, n* 69: 1–30.

Brock, W. 1975. A simple perfect foresight monetary model. *Journal of Monetary Economics* 1: 133–150.

Calvo, G. 1983. Staggered prices in a utility-maximizing framework. *Journal of Monetary Economics* 12: 383–398.

Chari, V., P. Kehoe, and El McGrattan. 2000. Sticky price models of the business cycle: Can the contract multiplier solve the persistence problem? *Econometrica* 68: 1151–1179.

Cho, J.-O. 1993. Money and the business cycle with one-period nominal contracts. *Canadian Journal of Economics* 26: 638–659.

Cho, J.-O., and T. Cooley. 1995. Business cycles with nominal contracts. *Economic Theory* 6: 13–34.

Cho, J.-O., T. Cooley, and L. Phaneuf. 1997. The welfare cost of nominal wage contracting. *Review of Economic Studies* 64: 465–484.

Christiano, L., M. Eichenbaum, and C. Evans. 1999. Monetary policy shocks: what have we learned and to what end? In *Handbook of macroeconomics*, ed. J. Taylor and M. Woodford, Vol. 1A. Amsterdam: North-Holland.

Christiano, L., M. Eichenbaum, and C. Evans. 2005. Nominal rigidities and the dynamic effects of a shock to monetary policy. *Journal of Political Economy* 113: 1–46.

Clower, R. 1965. The Keynesian counterrevolution: A theoretical appraisal. In *The theory of interest rates*, ed. F. Hahn and F. Brechling. London: Macmillan.

Cogley, T., and J. Nason. 1993. Impulse dynamics and propagation mechanisms in a real business cycle model. *Economics Letters* 43: 77–81.

Cogley, T., and J. Nason. 1995. Output dynamics in real-business-cycle models. *American Economic Review* 85: 492–511.

Collard, F., and G. Ertz. 2000. Stochastic nominal wage contracts in a cash-in-advance model. *Recherches Economiques de Louvain* 66: 281–301.

Danthine, J.-P., and J. Donaldson. 1990. Efficiency wages and the business cycle puzzle. *European Economic Review* 34: 1275–1301.

Danthine, J.-P., and J. Donaldson. 1991. Risk sharing, the minimum wage and the business cycle. In *Equilibrium theory and applications: A conference in honor of Jacques Drèze*, ed. W. Barnett et al. Cambridge: Cambridge University Press.

Danthine, J.-P., and J. Donaldson. 1992. Risk sharing in the business cycle. *European Economic Review* 36: 468–475.

Debreu, G. 1959. *Theory of value*. New York: Wiley.

Dixit, A., and J. Stiglitz. 1977. Monopolistic competition and optimum product diversity. *American Economic Review* 67: 297–308.

Drèze, J. 1975. Existence of an exchange equilibrium under price rigidities. *International Economic Review* 16: 301–320.

Fischer, S. 1977. Long-term contracts, rational expectations, and the optimal money supply rule. *Journal of Political Economy* 85: 191–205.

Gabszewicz, J., and J.-P. Vial. 1972. Oligopoly 'A la Cournot' in a general equilibrium analysis. *Journal of Economic Theory* 42: 381–400.

Gordon, D. 1974. A neo-classical theory of Keynesian unemployment. *Economic Inquiry* 12: 431–459.

Gray, J.-A. 1976. Wage indexation: A macroeconomic approach. *Journal of Monetary Economics* 2: 221–235.

Hairault, J.-O., and F. Portier. 1993. Money, new-Keynesian macroeconomics and the business cycle. *European Economic Review* 37: 1533–1568.

Hicks, J. 1937. Mr. Keynes and the 'classics': A suggested interpretation. *Econometrica* 5: 147–159.

Huang, K., and Z. Liu. 2002. Staggered price-setting, staggered wage-setting, and business cycle persistence. *Journal of Monetary Economics* 49: 405–433.

Jeanne, O. 1998. Generating real persistent effects of monetary shocks: How much nominal rigidity do we really need? *European Economic Review* 42: 1009–1032.

Keynes, J.M. 1936. *The general theory of employment, interest and money.* New York: Harcourt Brace.

Kydland, F., and E. Prescott. 1982. Time to build and aggregate fluctuations. *Econometrica* 50: 1345–1370.

Leijonhufvud, A. 1968. *On Keynesian economics and the economics of Keynes*. Oxford: Oxford University Press.

Long, J., and C. Plosser. 1983. Real business cycles. *Journal of Political Economy* 91: 39–69.

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

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

Negishi, T. 1961. Monopolistic competition and general equilibrium. *Review of Economic Studies* 28: 196–201.

Patinkin, D. 1956. *Money, interest and prices*. 2nd ed, 1965. New York: Harper and Row.

Phelps, E., and J. Taylor. 1977. Stabilizing powers of monetary policy under rational expectations. *Journal of Political Economy* 85: 163–190.

Rotemberg, J. 1982a. Monopolistic price adjustment and aggregate output. *Review of Economic Studies* 44: 517–531.

Rotemberg, J. 1982b. Sticky prices in the United States. *Journal of Political Economy* 90: 1187–1211.

Rotemberg, J., and M. Woodford. 1992. Oligopolistic pricing and the effects of aggregate demand on economic activity. *Journal of Political Economy* 100: 1153–1207.

Rotemberg, J., and M. Woodford. 1995. Dynamic general equilibrium models with imperfectly competitive product markets. In *Frontiers of business cycle research*, ed. T. Cooley. Princeton: Princeton University Press.

Sidrauski, M. 1967. Rational choice and patterns of growth in a monetary economy. *American Economic Review* 57(Suppl): 534–544.
Smets, F., and R. Wouters. 2003. An estimated dynamic stochastic general equilibrium model of the euro area. *Journal of the European Economic Association* 1: 1123–1175.
Svensson, L. 1986. Sticky goods prices, flexible asset prices, monopolistic competition and monetary policy. *Review of Economic Studies* 53: 385–405.
Taylor, J. 1979. Staggered wage setting in a macro model. *American Economic Review* 69: 108–113.
Taylor, J. 1980. Aggregate dynamics and staggered contracts. *Journal of Political Economy* 88: 1–23.
Taylor, J. 1999. Staggered price and wage setting in macroeconomics. In *Handbook of macroeconomics*, ed. J. Taylor and M. Woodford, Vol. 1. Amsterdam: North-Holland.
Walras, L. 1874. *Eléments d'économie politique pure*. Lausanne: Corbaz. Definitive edition Trans. W. Jaffe as *Elements of Pure Economics*. London: Allen and Unwin, 1954.
Yun, T. 1996. Nominal price rigidity, money supply endogeneity, and business cycles. *Journal of Monetary Economics* 37: 345–370.

Dynamic Programming

John Rust

Abstract

This article reviews the history and theory of dynamic programming (DP), a recursive method of solving sequential decision problems under uncertainty. It discusses computational algorithms for the numerical solution of DP problems, and an important limitation in our ability to solve realistic large-scale dynamic programming problems, the 'curse of dimensionality'. It also summarizes recent research in complexity theory that delineates situations where the curse can be broken (allowing us to solve DPs using fast polynomial time algorithms), and situations where it is insuperable. The literature on econometric estimation and testing of DP models is reviewed, as is another 'scientific limit to knowledge', namely, the identification problem.

Keywords

Backward induction; Bellman equation; Computational complexity; Computational experiments; Concavity; Continuous and discrete time models; Curse of dimensionality; Decision variables; Discount factor; Dynamic discrete choice models; Dynamic games; Dynamic programming; Econometric estimation; Euler equations; Game tree; Identification; Independence; Indirect inference; Infinite horizons; Kalman filtering; Kuhn–Tucker th; Markov chain Monte Carlo methods; Markovian decision problems; Maximum likelihood; Method of simulated moments; Method of simulated scores; Minimum residual method; Monotonicity; Monte Carlo integration; Neural networks; Nonlinear regression; Nonparametric regression; Optimal decision rules; Policy iteration; Principle of optimality; Rational expectations; Sequential decision problems; Simulated maximum likelihood; Simulated method of moments; Simulated minimum distance; Simulation-based estimation; State variables; Stationarity; Statistical decision theory; Structural estimation; Subgame perfection; Uncertainty; Wald, A

JEL Classification

C51; C61

Introduction

Dynamic programming is a recursive method for solving *sequential decision problems* (hereafter abbreviated as SDP). Also known as *backward induction*, it is used to find *optimal decision rules* in 'games against nature' and *subgame perfect equilibria* of dynamic multi-agent games, and competitive equilibria in dynamic economic models. Dynamic programming has enabled economists to formulate and solve a huge variety of problems involving sequential decision-making under uncertainty, and as a result it is

now widely regarded as the single most important tool in economics. Section "History" provides a brief history of dynamic programming. Section "Theory" discusses some of the main theoretical results underlying dynamic programming, and its relation to game theory and optimal control theory. Section "Numerical dynamic programming and the curse of dimensionality" provides a brief survey of numerical dynamic programming. Section "Empirical dynamic programming and the identification problem" surveys the experimental and econometric literature that uses dynamic programming to construct empirical models economic behaviour.

History

The earliest reference to the use of the method of backward induction to solve decision problems appears to be Arthur Cayley's 1875 solution to the secretary problem (I am grateful to Arthur F. Veinott Jr. for alerting me to this). In the mid-1940s a number of different researchers in economics and statistics appear to have independently discovered backward induction as a way to solve SDPs involving risk or uncertainty. Von Neumann and Morgenstern, in their seminal work on game theory (1944), used backward induction to find what we now call *subgame perfect equilibria of extensive form games*. ('We proceed to discuss the game Γ by starting with the last move $\mathcal{M}_v$ and then going backward from there through the moves $\mathcal{M}_{v-1}, \mathcal{M}_{v-2} \cdots$': 1944, p. 126.) Abraham Wald, who is credited with the invention of *statistical decision theory*, extended this theory to sequential decision-making in his 1947 book *Sequential Analysis*. Wald generalized the problem of gambler's ruin from probability theory and introduced the *sequential probability ratio test* that minimizes the expected number of observations in a sequential generalization of the classical hypothesis test. However, the role of backward induction is less obvious in Wald's work. It was more clearly elucidated in the 1949 paper by Arrow, Blackwell and Girshick. They studied a generalized version of the statistical decision problem and formulated and solved it in a way that is a readily recognizable application of modern dynamic programming. Following Wald, they characterized the optimal rule for making a statistical decision (for example, accept or reject a hypothesis), accounting for the costs of collecting additional observations. In the section 'The Best Truncated Procedure' they show how the optimal rule can be approximated 'Among all sequential procedures not requiring more than N observations ...' and solve for the optimal truncated sampling procedure 'by induction backwards' (1949, p. 217).

Other early applications of backward induction include the work of Pierre Massé (1945, p. 196) on statistical hydrology and the management of reservoirs, and Dvoretzky et al. (1952) analysis of optimal inventory policy. Richard Bellman is widely credited with recognizing the common structure underlying SDPs, and showing how backward induction can be applied to solve a huge class of SDPs under uncertainty. Most of Bellman's work in this area was done at the RAND Corporation, starting in 1949. It was there that he invented the term 'dynamic programming' that is now the generally accepted synonym for backward induction. Bellman (1984, p. 159) explained that he invented the name 'dynamic programming' to hide the fact that he was doing mathematical research at RAND under a Secretary of Defense who 'had a pathological fear and hatred of the term, research'. He settled on 'dynamic programming' because it would be difficult give it a 'pejorative meaning' and because 'It was something not even a Congressman could object to'.

Theory

Dynamic programming can be used to solve for optimal strategies and equilibria of a wide class of SDPs and multiplayer games. The method can be applied both in discrete time and continuous time settings. The value of dynamic programming is that it is a 'practical' (that is, *constructive*) method for finding solutions to extremely complicated problems. However, continuous time problems involve technicalities that I wish to avoid in this survey. If a continuous time problem does not admit a closed-form solution, the most commonly used numerical approach is to solve an

approximate discrete time version of the problem or game, since under very general conditions one can find a sequence of discrete time DP problems whose solutions converge to the continuous time solution the time interval between successive decisions tends to zero (Kushner 1990). I start by describing how dynamic programming is used to solve single agent 'games against nature'. The approach can be extended to solve multiplayer games, dynamic contracts, principal–agent problems, and competitive equilibria of dynamic economic models. See recursive competitive equilibrium.

Sequential Decision Problems

There are two key variables in any dynamic programming problem: a *state variable* s_t, and a *decision variable* d_t (the decision is often called a 'control variable' in the engineering literature). These variables can be vectors in R^n, but in some cases they might be *infinite-dimensional* objects. For example, in Bayesian decision problems, one of the state variables might be a *posterior distribution* for some unknown quantity θ. In general, this posterior distribution lives in an infinite dimensional space of all probability distributions on θ. In heterogeneous agent equilibrium problems state variables can also be distributions. The state variable evolves randomly over time, but the agent's decisions can affect its evolution. The agent has a *utility* or *payoff function* $U(s_1, d_1, \ldots, s_T, d_T)$ that depends on the realized states and decisions from period $t = 1$ to the *horizon* T. In some cases $T = \infty$, and we say the problem is *infinite horizon*. In other cases, such as a life-cycle decision problem, T might be a random variable, representing a consumer's date of death. As we will see, dynamic programming can be adapted to handle either of these possibilities. Most economic applications presume a *discounted*, *time-separable* objective function, that is, U has the form

$$U(s_1, d_1, \ldots, s_T, d_T) = \sum_{t=1}^{T} \beta^t u_t(s_t, d_t) \qquad (1)$$

where β is known as a *discount factor* that is typically presumed to be in the (0, 1) interval, and $u_t(s_t, d_t)$ is the agent's *period t utility* (*payoff*) function. Discounted utility and profits are typical examples of time separable payoff functions studied in economics. However, the method of dynamic programming does not require time separability, and so I will describe it without imposing this restriction.

We model the uncertainty underlying the decision problem via a family of history and decision-dependent conditional probabilities $\{p_t(s_t|H_{t-1})\}$ where $H_{t-1} = (s_1, d_1, \ldots, s_{t-1}, d_{t-1})$ denotes the *history*, that is the realized states and decisions from the initial date $t = 1$ to date $t = T$. Note that this includes all deterministic SDPs as a special case where the transition probabilities p_t are degenerate. In this case we can represent the 'law of motion' for the state variables by deterministic functions $s_{t+1} = f_t(s_t, d_t)$. This implies that in the most general case, $\{s_t, d_t\}$, evolves as a history dependent stochastic process. Continuing the 'game against nature' analogy, it will be helpful to think of $\{p_t(s_t|H_{t-1})\}$ as constituting a 'mixed strategy' played by 'nature' and the agent's optimal strategy as a 'best response' to nature's strategy.

The final item we need to specify is the *timing of decisions*. Assume that the agent selects d_t *after* observing s_t, which is 'drawn' from the distribution $p_t(s_t|H_{t-1})$. The alternative case where d_t is chosen before s_t is realized can also be handled, but requires a small change in the formulation of the problem. The agent's choice of d_t is restricted to a *state-dependent constraint (choice) set* $D_t(H_{t-1}, s_t)$. We can think of D_t as the generalization of a 'budget set' in standard static consumer theory. The choice set could be a finite set, in which case we refer to the problem as *discrete choice*, or D_t could be a subset of R^k with non-empty interior, then we have a *continuous choice* problem. In many cases, there is a mixture of types of choices, which we refer to as *discrete-continuous choice problems*. An example is commodity price speculation; see for example Hall and Rust (2006), where a speculator has a discrete choice of whether or not to order to replenish his inventory and a continuous decision of how much of the commodity to order. Another example is retirement: a person has a discrete decision of

whether to retire and a continuous decision of how much to consume.

Definition A (single agent) *sequential decision problem* (SDP) consists of (1) a *utility function* U, (2) a sequence of *choice sets* $\{D_t\}$, and (3) a sequence of *transition probabilities* $\{pt(s_t|H_{t-1})\}$ where we assume that the process is initialized at some given initial state s_1.

In order to solve this problem, we have to make assumptions about how the decision-maker evaluates alternative risky strategies. The standard assumption is that the decision-maker maximizes *expected utility*. Backward induction does not necessarily result in optimal strategies for *non-expected utility maximizers*, except for certain classes of *recursive preferences*.

As the name implies, an expected utility maximizer makes decisions that maximize their *ex ante* expected utility. However, since information unfolds over time, it is generally not optimal to *pre-commit* to any fixed sequence of actions $(d_1, \ldots, d_T)$. Instead, the decision-maker can generally obtain higher expected utility by adopting a *history-dependent strategy* or *decision rule* $(\delta_1, \ldots, \delta_T)$. This is a sequence of *functions* such that for each time t the realized decision is a function of all available information. In the engineering literature, a decision rule that does not depend on evolving information is referred to as an *open-loop* strategy, whereas one that does is referred to as a *closed-loop* strategy. In deterministic control problems, the closed-loop and open-loop strategies are the same since both are simple functions of time. However in stochastic control problems, open-loop strategies are a strict subset of closed-loop strategies. Under our timing assumptions the information available at time t is (H_{t-1}, s_t), so we can write $d_t = \delta_t(H_{t-1}, s_t)$. By convention we set $H_0 = \varnothing$ so that the available information for making the initial decision is just s_1. A decision rule is *feasible* if it also satisfies $\delta_t(H_{t-1}, s_t) \in D_t(H_{t-1}, s_t)$ for all (s_t, H_{t-1}). Each feasible decision rule can be regarded as a 'lottery' whose payoffs are utilities, the expected value of which corresponds to expected utility associated with the decision rule. An *optimal decision rule* $\delta^* \equiv (\delta_1^*, \ldots, \delta_T^*)$ is simply a feasible decision rule that maximizes the decision-maker's expected utility

$$\delta^* = \underset{\delta \in \mathcal{F}}{\operatorname{argmax}} E\{U(\{\tilde{s}_t, \tilde{d}_t\}_\delta)\}, \qquad (2)$$

where $\mathcal{F}$ denotes the class of feasible history-dependent decision rules, and $\{\tilde{s}_t, \tilde{d}_t\}_\delta$ denotes the stochastic process induced by the decision rule $\delta \equiv (\delta_1, \ldots, \delta_T)$. Problem (2) can be regarded as a static, *ex ante* version of the agent's problem. In game theory, (2) is referred to as the *normal form* or the *strategic form* of a dynamic game, since the dynamics are suppressed and the problem has the superficial appearance of a static optimization problem or game in which an agent's problem is to choose a best response, either to nature (in the case of single agent decision problems) or to other rational opponents (in the case of games). The strategic formulation of the agent's problem is quite difficult to solve since the solution is a *sequence of history-dependent functions* $\delta^* = (\delta_1^*, \ldots, \delta_T^*)$ for which standard finite dimensional constrained optimization techniques (for example, the Kuhn–Tucker th) are inapplicable. (If we consider problems where all states can assume only a finite number of values, it is possible to apply standard finite dimensional Kuhn–Tucker constrained optimization methods, but if the state variables can assume a continuum of possible values, the programming problem becomes an infinite dimensional programming problem for which optimal control and dynamic programming methods are more appropriate. See Luenberger 1969, for a more thorough discussion of how Lagrange multipliers and Kuhn–Tucker methods can be extended to problems where decisions are infinite-dimensional objects. These methods are usually applied in deterministic context, and there is a specialized literature on optimal control for solving such problems.) See Pontryagin's principle of optimality.

Solving Sequential Decision Problems by Backward Induction

To carry out backward induction, we start at the last period, T, and for each possible combination (H_{T-1}, s_T) we calculate the time T *value function*

and *decision rule* (we will discuss how backward induction can be extended to cases where T is random or where $T = \infty$ shortly).

$$V_T(H_{T-1}, s_T) = \max_{d_T \in D_T(H_{T-1}, s_T)} U(H_{T-1}, s_T, d_T)$$
$$\delta_T(H_{T-1}, s_T) = \operatorname*{argmax}_{d_T \in D_T(H_{T-1}, s_T)} U(H_{T-1}, s_T, d_T), \tag{3}$$

where we have written $U(H_{T-1}, s_T, d_T)$ instead of $U(s_1, d_1, \ldots, s_T, d_T)$ since $H_{T-1} = (s_1, d_1, \ldots, s_{T-1}, d_{T-1})$. Next we move backward one time period to time $T-1$ and compute

$$\begin{aligned} V_{T-1}(H_{T-2}, s_{T-1}) &= \max_{d_{T-1} \in D_{T-1}(H_{T-2}, s_{T-1})} E\{V_T(H_{T-2}, s_{T-1}, d_{T-1}, \tilde{s}_T) | H_{T-2}, s_{T-1}, d_{T-1}\} \\ &= \max_{d_{T-1} \in D_{T-1}(H_{T-2}, s_{T-1})} \int V_T(H_{T-2}, s_{T-1}, d_{T-1}, s_T) p_T(s_T | H_{T-2}, s_{T-1}, d_{T-1}) \\ \delta_{T-1}(H_{T-2}, s_{T-1}) &= \operatorname*{argmax}_{d_{T-1} \in D_{T-1}(H_{T-2}, s_{T-1})} E\{V_T(H_{T-2}, s_{T-1}, d_{T-1}, \tilde{s}_T) | H_{T-2}, s_{T-1}, d_{T-1}\} \end{aligned} \tag{4}$$

where the integral in Eq. (4) is the formula for the conditional expectation of V_T, where the expectation is taken with respect to the random variable $\tilde{s}_T$ whose value is not known as of time $T-1$. We continue the backward induction recursively for time periods $T-2, T-3, \ldots$ until we reach time period $t = 1$. The equation for the value function V_t in an arbitrary period t is defined recursively by an equation that is now commonly called the *Bellman equation*

$$\begin{aligned} &V_t(H_{t-t}, s_t) \\ &= \max_{d_t \in D_t(H_{t-1}, s_t)} E\{V_{t+1}(H_{t-1}, s_t, d_t, \tilde{s}_{t+1}) | H_{t-1}, s_t, d_t\} \\ &= \max_{d_t \in D_t(H_{t-1}, s_t)} \int V_{t+1}(H_{t-1}, s_t, d_t, s_{t+1}) p_{t+1}(s_{t+1} | H_{t-1}, s_t, d_t). \end{aligned} \tag{5}$$

The decision rule δ_t is defined by the value of d_t that attains the maximum in the Bellman equation for each possible value of (H_{t-1}, s_t)

$$\begin{aligned} &\delta_t(H_{t-t}, s_t) \\ &= \operatorname*{argmax}_{d_t \in D_t(H_{t-1}, s_t)} E\{V_{t+1}(H_{t-1}, s_t, d_t, \tilde{s}_{t+1}) | H_{t-1}, s_t, d_t\}. \end{aligned} \tag{6}$$

Backward induction ends when we reach the first period, in which case, as we will now show, the function $V_1(s_1)$ provides the expected value of an optimal policy, starting in state s_1 implied by the recursively constructed sequence of decision rules $\delta = (\delta_1, \ldots, \delta_T)$.

The Principle of Optimality

The key idea underlying why backward induction produces an optimal decision rule is called

The Principle of Optimality *An optimal decision rule* $\delta^* = \left(\delta_1^*, \ldots, \delta_T^*\right)$ *has the property that given any* $t \in \{1, \ldots, T\}$ *and any history* H_{t-1} in the support of the controlled process $\{s_t, d_t\}_{\delta^*}$, δ^* *remains optimal for the 'subgame' starting at time* t *and history* H_{t-1}. *That is,* δ^* *maximizes the "continuation payoff" given by the conditional expectation of utility from period* t *to* T, *given history* H_{t-1}:

$$\delta^* = \operatorname*{argmax}_{\delta} E\left\{U\left(\{s_t, d_t\}_{\delta}\right) | H_{t-1}\right\}. \tag{7}$$

In game theory, the principle of optimality is equivalent to the concept of a *subgame perfect equilibrium* in an *extensive form game*. When all actions and states are discrete, the stochastic decision problem can be diagrammed as a *game tree*. The principle of optimality, which in game theory is equivalent to the concept of a *subgame perfect*

equilibrium, guarantees that if δ^* is an optimal strategy (or equilibrium strategy) for the overall game tree, then it must also be an optimal strategy for every subgame, or, more precisely, *all subgames that are reached with positive probability from the initial node*.

It should now be evident why there is a need for the qualification 'for all H_{t-1} in the support of $\{s_t, d_t\}_{\delta^*}$' in the statement of the principle of optimality. There are some subgames that are never reached with positive probability under an optimal strategy. Thus, it is easy to construct alternative optimal decision rules that do not satisfy the principle of optimality because they involve taking suboptimal decisions on 'zero probability subgames'. Since these subgames are never reached, such modifications do not jeopardize *ex ante* optimality. However we cannot be sure *ex ante* which subgames will be irrelevant *ex post* unless we carry out the full backward induction process. Dynamic programming results in strategies that are optimal in *every possible subgame*, even those which will never be reached when the strategy is executed. Since backward induction results in a decision rule δ that is optimal for all possible subgames, it is intuitively clear that δ is optimal for the game as a whole, that is, it is a solution to the *ex ante* strategic form of the optimization problem (2).

For a formal proof of this result for games against nature (with appropriate care taken to ensure measurability and existence of solutions), see Gihman and Skorohod (1979). If in addition to 'nature' we extend the game tree by adding another rational expected utility maximizing player, then backward induction can be applied in the same way to solve this alternating move dynamic game. Assume that player 1 moves first, then player 2, then nature, and so on. Dynamic programming results in a *pair of strategies* for both players. Nature still plays a 'mixed strategy' that could depend on the entire previous history of the game, including all the previous moves of both players. The backward induction process ensures that each player can predict the future choices of their opponent, not only in the succeeding move but in all future stages of the game. The pair of strategies (δ^1, δ^2) produced by dynamic programming are mutual best responses, as well as being best responses to nature's moves. Thus, these strategies constitute a *Nash equilibrium*. They actually satisfy a stronger condition: they are Nash equilibrium strategies in every possible subgame of the original game, and thus are *subgame-perfect* (Selten 1975). Subgame-perfect equilibria exclude 'implausible equilibria' based on *incredible threats*. A standard example is an incumbent's threat to engage in a price war if a potential entrant enters the market. This threat is incredible if the incumbent would not really find it advantageous to engage in a price war (resulting in losses for both firms) if the entrant called its bluff and entered the market. Thus the set of all Nash equilibria to dynamic multiplayer games is strictly larger than the subset of subgame-perfect equilibria, a generalization of the fact that, in single agent decision problems, the set of optimal decision rules includes ones which take suboptimal decisions on subgames that have zero chance of being reached for a given optimal decision rule. Dynamic programming ensures that the decision-maker would never mistakenly reach any such subgame, similar to the way subgame perfection ensures that a rational player would not be fooled by an incredible threat.

Dynamic Programming for Stationary, Markovian, Infinite-Horizon Problems

The *complexity* of dynamic programming arises from the exponential growth in the number of possible histories as the number of possible values for the state variables, decision variables, and/or number of time periods T increases. For example, in a problem with N possible values for s_t and D possible values for d_t in each time period t, there are $[ND]^T$ possible histories, and thus the required number of calculations to solve a general T period, history-dependent dynamic programming problem is $O([ND]^T)$. Bellman and Dreyfus (1962) referred to this exponential growth in the number of calculations as the *curse of dimensionality*. In the next section, I will describe various strategies for dealing with this problem, but an immediate solution is to restrict attention to *time*

separable Markovian decision problems. These are problems where the payoff function U is additively separable as in Eq. (1), and where both the choice sets $\{D_t\}$ and the transition probabilities $\{p_t\}$ depend only on the contemporaneous state variable s_t and not on the entire previous history H_{t-1}. We say a conditional distribution p_t satisfies the *Markov property* if it depends on the previous history only via the most recent values, that is, if $p_t(s_t|H_{t-1}) = p_t(s_t|s_{t-1}, d_{t-1})$. In this case backward induction becomes substantially easier. For example, in this case the dynamic programming optimizations have to be performed only at each of the N possible values of the state variable at each time t, so only $O(NDT)$ calculations are required to solve a time T period time separable Markovian problem instead of $O([ND]^T)$ calculations when histories matter. This is part of the reason why, even though time non-separable utilities and non-Markovian forms of uncertainty may be more general, most dynamic programming problems that are solved in practical applications are both time separable and Markovian.

SDPs with random horizons $\tilde{T}$ can be solved by backward induction provided there is some finite time $\overline{T}$ satisfying $Pr\{\tilde{T} \leq \overline{T} = 1$. In this case, backward induction proceeds from the maximum possible value $\overline{T}$ and the *survival probability* $\rho_t = Pr\{\tilde{T} > t | \tilde{T} \geq t$ is used as to capture the probability that the problem will continue for at least one more period. The Bellman equation for the discounted, time-separable utility with uncertain lifetime is

$$\begin{aligned} V_t(s_t) &= \max_{d \in D_t(s_t)} [u_t(s_t, d) + \rho_t \beta E V_{t+1}(s_t, d)] \\ \delta_t(s_t) &= \operatorname*{argmax}_{d \in D_t(s_t)} [u_t(s_t, d) + \rho_t \beta E V_{t+1}(s_t, d)], \end{aligned} \tag{8}$$

where

$$EV_{t+1}(s, d) = \int_{s'} V_{t+1}(s') p_{t+1}(s'|s, d). \tag{9}$$

In many problems there is no finite upper bound $\overline{T}$ on the horizon. These are called *infinite horizon problems* and they occur frequently in economics. For example, SDPs used to model decisions by firms are typically treated as infinite horizon problems. It is also typical in infinite horizon problems to assume *stationarity*. That is, the utility function $u(s, d)$, the constraint set $D(s)$, the survival probability ρ, and the transition probability $p(s'|s, d)$ do not explicitly depend on time t. In such cases, it is not hard to show the value function and the optimal decision rules are also stationary, and satisfy the following version of Bellman's equation

$$\begin{aligned} V(s) &= \max_{d \in D(s)} [u(s, d) + \rho \beta E V(s, d)] \\ \delta(s) &= \operatorname*{argmax}_{d \in D(s)} [u(s, d) + \rho \beta E V(s, d)], \end{aligned} \tag{10}$$

where

$$EV(s, d) = \int_{s'} V(s') p(s'|s, d). \tag{11}$$

This is a fully recursive definition of V, and as such there is an issue of existence and uniqueness of a solution. In addition, it is not obvious how to carry out backward induction, since there is no 'last' period from which to begin the backward induction process. However, under relatively weak assumptions one can show there is a unique V satisfying the Bellman equation, and the implied decision rule in Eq. (10) is an optimal decision rule for the problem. Further, this decision rule can be approximated by solving an approximate finite horizon version of the problem by backward induction.

For example, suppose that $u(s, d)$ is a continuous function of (s, d), the state space S is compact, the constraint sets $D(s)$ are compact for each $s \in S$, and the transition probability $p(s'|s, d)$ is weakly continuous in (s, d) (that is, $EV(s, d) \equiv \int_{s'} W(s') p(s'|s, d)$ is a continuous function of (s, d) for each continuous function $W: S \to R$). Blackwell (1965a, b), Denardo (1967) and others have proved that, under these sorts of assumptions, V is the unique fixed point to the *Bellman operator* $\Gamma: B \to B$, where B is the Banach space of

continuous functions on S under the supremum norm, and Γ is given by

$$\Gamma(W)(s) = \max_{d \in D(s)} \left[u(s,d) + \rho\beta \int_{s'} W(s') p(s'|s,d) \right]. \tag{12}$$

The existence and uniqueness of V is a consequence of the *contraction mapping th*, since Γ can be shown to satisfy the contraction property,

$$\| \Gamma W - \Gamma V \| \leq \alpha \| W - V \|, \tag{13}$$

where $\alpha \in (0,1)$ and $\|W\| = \sup_{s \in S} |W(s)|$. In this case, $\alpha = \rho\beta$, so the Bellman operator will be a contraction mapping if $\rho\beta \in (0,1)$.

The proof of the optimality of the decision rule δ in Eq. (10) is somewhat more involved. Using the Bellman equation (10), we will show that (see Eq. (34) in section "Numerical dynamic programming and the curse of dimensionality"),

$$\begin{aligned} V(s) &= u(s,\delta(s)) \\ &\quad + \rho\beta \int_{s'} V(s') p(s'|s,\delta(s)) \\ &= E\left\{ \sum_{t=0}^{\infty} [\rho\beta]^t u?(s_t,\delta(s_t)) \middle| s_0 = s \right\}, \end{aligned} \tag{14}$$

that is, V is the value function implied by the decision rule δ. Intuitively, the boundedness of the utility function, combined with discounting of future utilities, $\rho\beta \in (0,1)$, implies that if we truncate the infinite horizon problem to a T period problem, the error in doing so would be arbitrarily small when T is sufficiently large. Indeed, this is the key to understanding how to find approximately optimal decision rules to infinite horizon SDPs: *we approximate the infinite horizon decision rule δ by solving an approximate finite horizon version of the problem by dynamic programming*. The validity of this approach can be formalized using a well-known property of contraction mappings, namely, that the *method of successive approximations* starting from any initial guess W converges to the fixed point of Γ, that is

$$\lim_{t \to \infty} V_t = \Gamma^t(W) = V \ \ \forall W \in B, \tag{15}$$

where $\Gamma^t W$ denotes t successive iterations of the Bellman operator Γ,

$$\begin{aligned} V_0 &= \Gamma^0(W) = W \\ V_1 &= \Gamma^1(W) \\ &\cdots \\ V_t &= \Gamma^t(W) = \Gamma\left(\Gamma^{t-1} W\right) = \Gamma(V_{t-1}). \end{aligned} \tag{16}$$

If $W = 0$ (that is the zero function in B), then $V_T = \Gamma^T(0)$ is simply the period $t = 1$ value function resulting from the solution of a T period dynamic programming problem. Thus, this result implies that the optimal value function V_T for a T-period approximation to the infinite horizon problem converges to V as $T \to \infty$. Moreover, the difference in the two functions satisfies the bound

$$\|V_T - V\| \leq \frac{[\rho\beta]^T \| u \|}{1 - \rho\beta}. \tag{17}$$

Let $\delta_T = \delta_{1,T}, \delta_{2,T}, \ldots, \delta_{T,T}$ be the optimal decision rule to the T period problem. It can be shown that, if we follow this decision rule up to period T and then use $\delta_{1,T}$ in every period after T, the resulting decision rule is approximately optimal in the sense that the value function for this infinite horizon problem also satisfies inequality (17), and thus can be made arbitrarily small as T increases.

In many cases in economics the state space S has no natural upper bound. An example might be where s_t denotes an individual's wealth at time t, or the capital stock of the firm. If the unboundedness of the state space results in unbounded payoffs, the contraction mapping argument must be modified since the Banach space structure under the *supremum* norm no longer applies to unbounded functions. Various alternative approaches have been used to prove existence of optimal decision rules for unbounded problems. One is to use an alternative norm (for example, a *weighted norm*) and demonstrate that the Banach space/contraction mapping argument still applies. However, there are cases where there are no natural weighted norms, and the contraction mapping property cannot hold

since the Bellman equation can be shown to have multiple solutions. The most general conditions under which the existence and uniqueness of the solution V to the Bellman equation and the optimality of the implied stationary decision rule δ has been established is in Bhattacharya and Majumdar (1989). However, as I discuss in the next section, considerable care must be taken in solving unbounded problems numerically.

Numerical Dynamic Programming and the Curse of Dimensionality

The previous section showed that dynamic programming is a powerful tool that has enabled us to formulate and solve a wide range of economic models involving sequential decision-making under uncertainty – at least 'in theory'. Unfortunately, the cases where dynamic programming results in *analytical, closed-form solutions* are rare and often rather fragile in the sense that small changes in the formulation of a problem can destroy the ability to obtain an analytical solution. However even though most problems do not have analytical solutions, the theorems in the previous section guarantee the *existence* of solutions, and these solutions can be calculated (or approximated) by numerical methods. Since the 1980s, faster computers and better numerical methods have made dynamic programming a tool of substantial practical value by significantly expanding the range of problems that can be solved. In particular, it has led to the development of a large and rapidly growing literature on econometric estimation and testing of 'dynamic structural models' that I will discuss in the next section.

However, there are still many difficult challenges that prevent us from formulating and solving models that are as detailed and realistic as we might like, a problem that is especially acute in empirical applications. The principal challenge is what Bellman and Dreyfus (1962) called *the curse of dimensionality*. We have already illustrated this problem in section "Dynamic programming for stationary, Markovian, infinite-horizon problems": for history-dependent SDPs with a finite horizon T and a finite number of states N and actions D, dynamic programming requires $O([ND]^T)$ operations to find a solution. Thus it appears that the time required to compute a solution via dynamic programming increases *exponentially fast* with the number of possible decisions or states in a dynamic programming problem.

Fortunately, computer power (for example, operations per second) has also been growing exponentially fast, a consequence of *Moore's Law* and other developments in information technology, such as improved communications and massive parallel processing. Bellman and Dreyfus (1962) carried out calculations on RAND's 'Johnniac' computer (named in honour of Jon von Neumann, whose work contributed to the development of the first electronic computers) and reported that this machine could do 12,500 additions per second. Nowadays, in 2007, a typical laptop computer can do over a billion operations per second and we now have supercomputers that are approaching a *thousand trillion operations per second* – a level known as a 'petaflop'. In addition to faster 'hardware', research on numerical methods has resulted in significantly better 'software' that has had a huge impact on the spread of numerical dynamic programming and on the range of problems we can solve. In particular, algorithms have been developed that succeed in 'breaking' the curse of dimensionality, enabling us to solve in polynomial time classes of problems that were previously believed to be solvable only in exponential time. The key to breaking the curse of dimensionality is the ability to recognize and exploit *special structure* in an SDP problem. We have already illustrated an example of this in section "Dynamic programming for stationary, Markovian, infinite-horizon problems": if the SDP is Markovian and utility is time separable, a finite horizon, finite state SDP can be solved by dynamic programming in only $O(NDT)$ operations, compared to the $O([ND]^T)$ operations that are required in the general history-dependent case. There is only enough space here to discuss several of the most commonly used and most effective numerical methods for solving different types of SDPs by dynamic programming. I refer the reader to

Puterman (1994), Rust (1996) and Judd (1998) for more in-depth surveys on the literature on numerical dynamic programming. See computational methods in econometrics.

Naturally, the numerical method that is appropriate or 'best' depends on the type of problem being solved. Different methods are applicable depending on whether the problem has (*a*) finite versus infinite horizon, (*b*) finite versus continuous-valued state and decision variables, and (*c*) single versus multiple players. In finite horizon problems, backward induction is the essentially the only approach, although as we will see there are many different choices about how to most implement it most efficiently – especially in discrete problems where the number of possible values for the state variables is huge (for example, chess) or in problems with continuous state variables. In the latter case, it is clearly not possible to carry out backward induction for every possible history (or value of the state variable at stage t if the problem is Markovian and time separable), since there are infinitely many (indeed a continuum) of them. In these cases, it is necessary to *interpolate* the value function, whose values are only explicitly computed at a finite number of points in the state space. I use the term 'grid' to refer to the finite number of points in the state space where the backward induction calculations are actually performed. Grids might be *lattices* (that is, regularly spaced sets of points formed as Cartesian products of unidimensional grids for each of the continuous state variables), or they may be *quasi-random grids* formed by randomly sampling the state space from some probability distribution, or by generating deterministic sequences of points such as *low discrepancy sequences*. The reason why one might choose a random or low-discrepancy grid instead of regularly spaced lattice is to break the curse of dimensionality, as I discuss shortly. Also, in many cases it is advantageous to refine the grid over the course of the backward induction process, starting out with an initial 'coarse' grid with relatively few points and subsequently increasing the number of points in the grid as the backward induction progresses. I will have more to say about such *multigrid* and *adaptive grid* methods when I discuss solution of infinite horizon problems below.

Once a particular grid is chosen, the backward induction process is carried out in the way it would be normally be done in a finite state problem. On the assumption that the problem is Markovian and the utility is time separable and there are n grid points $\{s_1, \ldots, s_n\}$, this involves the following calculation at each grid point s_i, $i = 1,\ldots,n$

$$V_t(s_i) = \max_{d \in D_t(s_i)} \left[u_t(s_i, d) + \rho\beta\widehat{E}V_{t+1}(s_i, d)\right], \tag{18}$$

where $\widehat{E}V_{t+1}(s_i, d)$ is a numerical estimate of the conditional expectation of next period's value function. I will be more specific below about which numerical integration methods are appropriate, but at this point it suffices to note that they are all simple weighted sums of values of the value function at $t + 1$, $V_{t+1}(s)$. We can now see that, even if the actual backward induction calculations are carried out only at the n grid points $\{s_1,\ldots, s_n\}$, we will still have to do numerical integration to compute $\widehat{E}V_{t+1}(s_i, d)$ and the latter calculation may require values of $V_{t+1}(s)$ at points s off the grid, that is at points $s \notin \{s_1, \ldots, s_n\}$. This is why some form or interpolation (or in some cases *extrapolation*) is typically required. Almost all methods of interpolation can be represented as weighted sums of the value function at its known values $\{V_{t+1}(s_1), \ldots, V_{t+1}(s_n)\}$ at the n grid points, which were calculated by backward induction at the previous stage. Thus, we have

$$\widehat{V}_{t+1}(s) = \sum_{j=1}^{n} w_i(s)V_{t+1}(s_i), \tag{19}$$

where $w_i(s)$ is a weight assigned to the i^{th} grid point that depends on the point s in qst. These weights are typically positive and sum to 1. For example in *multilinear interpolation* or *simplicial interpolation* the $w_i(s)$ weights are those that allow s to be represented as a convex combination of the vertices of the smallest lattice hypercube containing s. Thus, the weights $w_i(s)$ will be zero for all i except the immediate neighbours of the

point s. In other cases, such as *kernel density* and *local linear regression*, the weights $w_i(s)$ are generally non-zero for all i, but the weights will be highest for the grid points $\{s_1, \ldots, s_n\}$ which are the *nearest neighbours* of s. An alternative approach can be described as *curve fitting*. Instead of attempting to interpolate the calculated values of the value function at the grid points, this approach treats these values as a *data-set* and estimates parameters θ of a flexible functional form approximation to $V_{t+1}(s)$ by *nonlinear regression*. Using the estimated $\widehat{\theta}_{t+1}$ from this nonlinear regression, we can 'predict' the value of $V_{t+1}(s)$ at any $s \in S$

$$\widehat{V}_{t+1}(s) = f\left(s, \widehat{\theta}_{t+1}\right). \tag{20}$$

A frequently used example of this approach is to approximate $V_{t+1}(s)$ as a linear combination of K 'basis functions' $\{b_1(s),\ldots, b_K(s)\}$. This implies that $f(s, \theta)$ takes the form of a *linear regression* function

$$f(s,\theta) = \sum_{k=1}^{K} \theta_k b_k(s), \tag{21}$$

and $\widehat{\theta}_{t+1}$ can be estimated by *ordinary least squares*. *Neural networks* are an example where f depends on θ in a nonlinear fashion. Partition θ into subvectors $\theta = (\gamma, \lambda, \alpha)$, where γ and λ are vectors in R^J, and $\alpha = (\alpha_1, \ldots, \alpha_J)$, where each α_j has the same dimension as the state vector s. Then the neural network f is given by

$$\begin{aligned} f(s,\theta) &= f(s,\gamma,\lambda,\alpha) \\ &= \sum_{j=1}^{J} \gamma_j \varphi\left(\lambda_j + \langle s, \alpha_i\rangle\right) \end{aligned} \tag{22}$$

where $\langle s, \alpha_j\rangle$ is the inner product between s and the conformable vector α_j, and φ is a 'squashing function' such as the logistic function $\varphi(x) = \exp\{x\}/(1 + \exp\{x\})$. Neural networks are known to be 'universal approximators' and require relatively few parameters to provide good approximations to nonlinear functions of many variables. For further details on how neural networks are applied, see the book by Bertsekas and Tsitsiklis (1996) on *Neuro-Dynamic Programming*.

All these methods require extreme care for problems with *unbounded state spaces*. By definition, any finite grid can cover only a small subset of the state space in this case, and thus any of the methods discussed above would require *extrapolation* of the value function to predict its values in regions where there are no grid points, and thus 'data' on what its proper values should be. Not only may mistakes that lead to incorrect extrapolations in these regions lead to errors in the regions where there are no grid points, but the errors can 'unravel' and also lead to considerable errors in approximating the value function in regions where we do have grid points. Attempts to 'compactify' an unbounded problem by arbitrarily truncating the state space may also lead to inaccurate solutions, since the truncation is itself an implicit form of extrapolation (for example, some assumption needs to be made what to do when state variables approach the 'boundary' of the state space: do we assume a 'reflecting boundary', an 'absorbing boundary', and so on?). For example in life-cycle optimization problems, there is no natural upper bound on wealth, even if it is true that there is only a finite amount of wealth in the entire economy. We can always ask the qst, if a person had wealth near the 'upper bound', what would happen to next period wealth if he invested some of it? Here we can see that, if we extrapolate the value function by assuming that the value function is bounded in wealth, this means that by definition there is no incremental return to saving as we approach the upper bound. This leads to lower saving, and this generally leads to errors in the calculated value function and decision rule far below the assumed upper bound. There is no good general solution to this problem except to solve the problem on a much bigger (bounded) state space than one would expect to encounter in practice, in the hope that extrapolation-induced errors in approximating the value function die out the further one is from the boundary. This property should hold for problems where the probability that the next period state will hit or exceed the 'truncation

boundary' gets small the farther the current state is from this boundary.

When a method for interpolating/extrapolating the value function has been determined, a second choice must be made about the appropriate method for *numerical integration* in order to approximate the conditional expectation of the value function $EV_{t+1}(s, d)$ given by

$$EV_{t+1}(s,d) = \int_{s'} V_{t+1}(s')p_{t+1}(s'|s,d). \qquad (23)$$

There are two main choices here: (1) deterministic quadrature rules or (2) (quasi-) Monte Carlo methods. Both methods can be written as weighted averages of form

$$\widehat{EV}_{t+1}(s,d) = \sum_{i=1}^{N} w_i(s,d)V_{t+1}(a_i), \qquad (24)$$

where $\{w_i(s, d)\}$ are *weights*, and $\{a_i\}$ are *quadrature abscissae*. Deterministic quadrature methods are highly accurate (for example, an N-point Gaussian quadrature rule is constructed to exactly integrate all polynomials of degree $2N - 1$ or less), but become unwieldy in multivariate integration problems when *product rules* (tensor products of unidimensional quadrature) are used. *Any* sort of deterministic quadrature method can be shown to be subject to the curse of dimensionality in terms of worst-case computational complexity (see Traub and Werschulz 1998). For example, if $N = O(1/\varepsilon)$ quadrature points are necessary to approximate a univariate integral within ε, then in a d-dimensional integration problem $N^d = O(1/\varepsilon^d)$ quadrature points would be necessary to approximate the integral with an error of ε, which implies that computational effort to find an ε-approximation increases exponentially fast in the problem dimension d. Using the theory of computational complexity, one can prove that *any* deterministic integration procedure is subject to the curse of dimensionality, at least in terms of a 'worst case' measure of complexity. The curse of dimensionality can disappear if one is willing to adopt a Bayesian perspective and place a 'prior distribution' over the space of possible integrands and consider an 'average case' instead of a 'worst case' notion of computational complexity.

Since multivariate integration is a 'subproblem' that must be solved in order to carry out dynamic programming when there are continuous state variables (indeed, dynamic programming in principle involves infinitely many integrals in order to calculate $EV_{t+1}(s, d)$, one for each possible value of (s, d)), if there is a curse of dimensionality associated with numerical integration of a single multivariate integral, then it should also not be surprising that dynamic programming is also subject to the same curse. There is also a curse of dimensionality associated with global optimization of nonconvex objective functions of continuous variables. Since optimization is also a sub-problem of the overall dynamic programming problem, this constitutes another reason why dynamic programming is subject to a curse of dimensionality. Under the standard worst case definition of computational complexity, Chow and Tsitsiklis (1989) proved that *no* deterministic algorithm can succeed in breaking the curse of dimensionality associated with a sufficiently broad class of dynamic programming problems with continuous state and decision variables. This negative result dashes the hopes of researchers dating back to Bellman and Dreyfus (1962), who conjectured that there might be sufficiently clever deterministic algorithms that can overcome the curse of dimensionality.

However, there are examples of *random algorithms* that can circumvent the curse of dimensionality. Monte Carlo integration is a classic example. Consider approximating the (multidimensional) integral in Eq. (23) by using *random* quadrature abscissae $\{\tilde{a}_i\}$ that are N independent and identically distributed (*IID*) draws from the distribution $p_{t+1}(s'|s, d)$ and uniform quadrature weights equal to $w_i(s, d) = 1/N$. Then the law of large numbers and the central limit theorem imply that the Monte Carlo integral $\tilde{EV}_{t+1}(s,d)$ converges to the true conditional expectation $EV_{t+1}(s, d)$ at rate $1/\sqrt{N}$ *regardless of the dimension of the state space d*. Thus a random algorithm, Monte Carlo integration, succeeds in breaking the curse of dimensionality of multivariate integration. Unfortunately,

randomization does *not* succeed in breaking the curse of dimensionality associated with general nonconvex optimization problems with continuous multidimensional decision variables d (see Nemirovsky and Yudin 1983).

However, naive application of Monte Carlo integration will not necessarily break the curse of dimensionality of the dynamic programming problem. The reason is that a form of *uniform convergence* (as opposed to pointwise) convergence of the conditional expectations $\widehat{EV}_{t+1}(s,d)$ to $EV_{t+1}(s, d)$ is required in order to guarantee that the overall backward induction process converges to the true solution as the number of Monte Carlo draws, N, gets large. To get an intuition why, note that if separate *IID* sets of quadrature abscissae $\{\tilde{a}_i\}$ where drawn for each (s, d) point that we wish to evaluate the Monte Carlo integral $\widehat{EV}_{t+1}(s,d)$ at, the resulting function would be an extremely 'choppy' and irregular function of (s, d) as a result of all the random variation in the various sets of quadrature abscissae. Extending an idea introduced by Tauchen and Hussey (1991) to solve rational expectations models, Rust (1997) proved that it is possible to break the curse of dimensionality in a class of SDPs where the choice sets $D_t(s)$ are finite, a class he calls *discrete decision processes*. The restriction to finite choice sets is necessary, since, as noted above, randomization does not succeed in breaking the curse of dimensionality of nonconvex optimization problems with continuous decision variables. The key idea is to choose, as a *random grid*, the same set of random points that are used quadrature abscissae for Monte Carlo integration. That is, suppose $p_{t+1}(s'|s, d)$ is a transition *density* and the state space (perhaps after translation and normalization) is identified with the d-dimensional *hypercube* $S = [0,1]^d$. Apply Monte Carlo integration by drawing N *IID* points $\{\tilde{s}_1, \ldots, \tilde{s}_N\}$ from the this hypercube (this can be accomplished by drawing each component of s_i from the uniform distribution on the [0,1] interval). We have

$$\widehat{EV}_{t+1}(s,d) = \frac{1}{N}\sum_{i=1}^{N} V_{t+1}(\tilde{s}_i)p_{t+1}(\tilde{s}_i|s,d). \quad (25)$$

Applying results from the theory of *empirical processes* (Pollard 1989), Rust showed that this form of the Monte Carlo integral does result in uniform convergence (that is, $\mathrm{P}\widehat{EV}_{t+1}(s,d) - EV_{t+1}(s,d)\mathrm{P} = O_p\left(1/\sqrt{N}\right)$), and, using this, he showed that this randomized version of backward induction succeeds in breaking the curse of dimensionality of the dynamic programming problem. The intuition of why this works is, instead of trying to approximate the conditional expectation in (23) by computing *many independent Monte Carlo integrals* (that is, drawing separate sets of random abscissae $\{\tilde{a}_i\}$ from $p_{t+1}(s'|s, d)$ for each possible value of (s, d)), the approach in Eq. (25) is to compute a *single Monte Carlo integral* where the random quadrature points $\{\tilde{s}_i\}$ are drawn from the uniform distribution on $[0,1]^d$, and the integrand is treated as the *function* $V_{t+1}(s')$ $p_{t+1}(s'|s, d)$ instead of $V_{t+1}(s')$. The second important feature is that Eq. (25) has a *self-approximating* property: that is, since the quadrature abscissae are the same as the grid points at which we compute the value function, no auxiliary interpolation or function approximation is necessary in order to evaluate $\widehat{EV}_{t+1}(s,d)$. In particular, if $p_{t+1}(s'|s, d)$ is a smooth function of s, then $\widehat{EV}_{t+1}(s,d)$ will also be a smooth function of s. Thus, backward induction using this algorithm is extremely simple. Before starting backward induction we choose a value for N and draw N *IID* random vectors $\{\tilde{s}_1, \ldots, \tilde{s}_N\}$ from the uniform distribution on the d-dimensional hypercube. This constitutes a random grid that remains fixed for the duration of the backward induction. Then we begin ordinary backward induction calculations, at each stage t computing $V_t(\tilde{s}_i)$ at each of the N random grid points, and using the self-approximating formula (25) to calculate the conditional expectation of the period $t + 1$ value function using only the N stored values $(V_{t+1}(\tilde{s}_1), \ldots, V_{t+1}(\tilde{s}_N))$ from the previous stage of the backward induction. See Keane and Wolpin (1994) for an alternative approach, which combines Monte Carlo integration with the curve-fitting approaches discussed above. Note that the Keane and Wolpin approach will not generally succeed in breaking the curse of dimensionality since it requires approximation of functions of

D

d variables which is also subject to a curse of dimensionality, as is well known from the literature on *nonparametric regression*.

There are other subclasses of SDPs for which it is possible to break the curse of dimensionality. For example, the family of *linear quadratic/Gaussian* (LQG) can be solved in polynomial time using highly efficient matrix methods, including efficient methods for solving the *matrix Ricatti equation* which is used to compute the *Kalman filter* for Bayesian LQG problems (for example, problems where agents only receive a noisy signal of a state variable of interest, and they update their beliefs about the unknown underlying state variable via Bayes rule).

Now consider stationary, infinite horizon Markovian decision problems. As noted in section "Dynamic programming for stationary, Markovian, infinite-horizon problems", there is no 'last' period from which to begin the backward induction process. However, if the utility function is time separable and discounted, then, under fairly general conditions, it will be possible to approximate the solution arbitrarily closely by solving a finite horizon version of the problem, where the horizon T is chosen sufficiently large. As we noted in section "Dynamic programming for stationary, Markovian, infinite-horizon problems", this is equivalent to solving for V, the fixed point to the contraction mapping $V = \Gamma(V)$ by the method of *successive approximations*, where Γ is the *Bellman operator* defined in Eq. (12) of section "Dynamic programming for stationary, Markovian, infinite-horizon problems".

$$V_{t+1} = \Gamma(V_t). \tag{26}$$

Since successive approximations converges at a geometric rate, with errors satisfying the upper bound in Eq. (17), this method can converge at an unacceptably slow rate when the discount factor is close to 1. A more effective algorithm in such cases is *Newton's method* whose iterates are given by

$$V_{t+1} = V_t - [I - \Gamma'(V_t)]^{-1}[V_t - \Gamma(V_t)], \tag{27}$$

where Γ' is the *Gateaux* or *directional derivative* of Γ, that is, it is the linear operator given by

$$\Gamma'(V)(W) = \lim_{t \to 0} \frac{\Gamma(V + tW) - \Gamma(V)}{t}. \tag{28}$$

Newton's method converges *quadratically* independent of the value of the discount factor, as long as it is less than 1 (to guarantee the contraction property and the existence of a fixed point). In fact, Newton's method turns out to be equivalent to the method of *policy iteration* introduced by Howard (1960). Let δ be any stationary decision rule, that is, a candidate *policy*. Define the policy-specific conditional expectation operator E_δ by

$$E_\delta V(s) = \int_{s'} V(s')p(s'|s, \delta(s)). \tag{29}$$

Given a value function V_t, let δ_{t+1} be the decision rule implied by V_t, that is

$$\delta_{t+1}(s) = \underset{d \in D(s)}{\operatorname{argmax}} \left[u(s,d) + \rho\beta \int_{s'} V_t(s')p(s'|s,d) \right]. \tag{30}$$

It is not hard to see that the value of policy δ_{t+1} must be at least as high as V_t, and for this reason, Eq. (30) is called the *policy improvement step* of the policy iteration algorithm. It is also not hard to show that

$$\Gamma'(V_t)(W)(s) = \rho\beta E_{\delta_{t+1}} W(s), \tag{31}$$

and this implies that the Newton iteration, Eq. (27), is numerically identical to *policy iteration*

$$V_{t+1}(s) = \left[I - \rho\beta E_{\delta_{t+1}} \right]^{-1} u(s, \delta_{t+1}(s)), \tag{32}$$

where δ_{t+1} is given in Eq. (30). Equation (32) is called the *policy valuation step* of the policy iteration algorithm since it calculates the value function implied by the policy δ_{t+1}. Note that, since E_δ is an expectation operator, it is linear and satisfies $\|E_\delta\| \leq 1$, and this implies that the operator

$[I - \rho\beta E_\delta]$ is invertible and has the following geometric series expansion

$$[I - \rho\beta E_\delta]^{-1} = \sum_{j=0}^{\infty} [\rho\beta]^j E_\delta^j, \tag{33}$$

where E_δ^j is the j step ahead expectations operator. Thus, we see that

$$[I - \rho\beta E_\delta]^{-1} u(s, \delta(s)) = \sum_{j=0}^{\infty} [\rho\beta]^j E_\delta^j u(s, \delta(s))$$
$$= E\left\{ \sum_{t=0}^{\infty} [\rho\beta]^t u(s_t, \delta(s_t)) | s_0 = s \right\}, \tag{34}$$

so that value function V_t from the policy iteration (32) corresponds to the expected value implied by policy (decision rule) δ_t.

If there are an infinite number of states, the expectations operator E_δ is an infinite-dimensional linear operator, so it is not feasible to compute an exact solution to the policy iteration Eq. (32). However if there are a finite number of states (or an infinite state space is discretized to a finite set of points, as per the discussion above), then E_δ is an $N \times N$ transition probability matrix, and policy iteration is feasible using ordinary matrix algebra, requiring at most $O(N^3)$ operations to solve a system of linear equations for V_t at each policy valuation step. Further, when there are a finite number of possible actions as well as states, there are only a finite number of possible policies $|D|^{|S|}$, where $|D|$ is the number of possible actions and $|S|$ is the number of states, and policy iteration can be shown to converge in a finite number of steps, since the method produces an improving sequences of decision rules, that is $V_t \leq V_{t+1}$. Thus, since there is an upper bound on the number of possible policies and policy iteration cannot cycle, it must converge in a finite number of steps. The number of steps is typically quite small, far fewer than the total number of possible policies. Santos and Rust (2004) show that the number of iterations can be bounded independent of the number of elements in the state space $|S|$. Thus, policy iteration is the method of choice for infinite horizon problems for which the discount factor is sufficiently close to 1. However, if the discount factor is far enough below 1, then successive approximations can be faster since policy iteration requires $O(N^3)$ operations per iteration whereas successive approximations requires $O(N^2)$ operations per iteration. At most $T(\varepsilon, \beta)$ successive approximation iterations are required to compute an ε-approximation to an infinite horizon Markovian decision problem with discount factor β, where $T(\varepsilon, \beta) = \log((1 - \beta)\varepsilon)/\log(\beta)$. Roughly speaking, if $T(\varepsilon, \beta) < N$, then successive approximations are faster than policy iteration.

Successive approximations can be accelerated by a number of means discussed in Puterman (1994) and Rust (1996). *Multigrid algorithms* are also effective: these methods begin backward induction with a coarse grid with relatively few grid points N, and then as iterations proceed, the number of grid points is successively increased, leading to finer and finer grids as the backward induction starts to converge. Thus, computational time is not wasted early on in the backward induction iterations when the value function is far from the true solution. *Adaptive grid* methods are also highly effective in many problems: these methods can automatically detect regions in the state space where there is higher curvature in the value function, and in these regions more grid points are added in order to ensure that the value function is accurately approximated, whereas in regions where the value function is 'flatter' grid points can be removed, so as to direct computational resources to the regions of the state space where there is the highest payoff in terms of accurately approximating the value function. See Grüne and Semmler (2004) for more details and an interesting application of adaptive grid algorithms.

I conclude this section with a discussion of several other alternative approaches to solving stationary infinite horizon problems that can be extremely effective relative to 'discretization' methods when the number of grid points N required to obtain a good approximation becomes very large. Recall the curve-fitting approach discussed above in finite horizon SDPs: we approximate the value function V by a parametric function as $V_\theta(s) \equiv f(s, \theta)$ for some

D

flexible functional form f, where θ are treated as unknown parameters to be 'estimated'. For infinite horizon SDPs, our goal is to find parameter values θ so that the implied value function satisfies the Bellman equation as well as possible. One approach to doing this, known as the *minimum residual method*, is a direct analogue of nonlinear least squares: if θ is a vector with K components, we select $N \geq K$ points in the state space (potentially at random) and find $\widehat{\theta}$ that minimizes the squared deviations or *residuals* in the Bellman equation

$$\widehat{\theta} = \underset{\theta \in R^K}{\operatorname{argmin}} \sum_{i=1}^{N} \left[\widehat{\Gamma}(V_\theta)(s_i) - V_\theta(s_i) \right]^2, \quad (35)$$

where $\widehat{\Gamma}$ denotes an approximation to the Bellman operator, where some numerical integration and optimization algorithm are used to approximate the true expectation operator and maximization in the Bellman equation (12). Another approach, called the *collocation method*, finds $\widehat{\theta}$ by choosing K grid points in the state space and setting the residuals at those K points to zero:

$$\begin{aligned} V_{\widehat{\theta}}(s_1) &= \widehat{\Gamma}\left(V_{\widehat{\theta}}\right)(s_1) V_{\widehat{\theta}}(s_2) \\ &= \widehat{\Gamma}\left(V_{\widehat{\theta}}\right)(s_2) \cdots V_{\widehat{\theta}}(s_K) \\ &= \widehat{\Gamma}\left(V_{\widehat{\theta}}\right)(s_K). \end{aligned} \quad (36)$$

Another approach, called *parametric policy iteration*, carries out the policy iteration algorithm in Eq. (32) above, but, instead of solving the linear system (32) for the value function V_t at each policy valuation step, they approximately solve this system by finding $\widehat{\theta}_t$ that solves the regression problem

$$\theta_t = \underset{\theta \in R^K}{\operatorname{argmin}} \sum_{i=1}^{N} \left[V_{\widehat{\theta}_t}(s_i) - u(s_i, \delta_t(s_i)) - \rho\beta E_{\delta_t} V_{\theta_t}(s_i) \right]^2. \quad (37)$$

Other than this, policy iteration proceeds exactly as discussed above. Note that, due to the linearity of the expectations operator, the regression problem above reduces to an ordinary linear regression problem when V_θ is approximated as a linear combination of basis functions as in (21) above.

There are variants of the minimum residual and collocation methods that involve parameterizing the *decision rule* rather than the value function. These methods are frequently used in problems where the control variable is continuous, and construct residuals from the *Euler equation* – a functional equation for the decision rule that can in certain classes of problems be derived from the first-order necessary condition for the optimal decision rule. These approaches then try to find $\widehat{\theta}$ so that the Euler equation (as opposed to the Bellman equation) is approximately satisfied, in the sense of minimizing the squared residuals (minimum residual approach) or setting the residuals to zero at K specified points in the state space (collocation method). See Judd (1998) for further discussion of these methods and a discussion of strategies for choosing the grid points necessary to implement the collocation or minimum residual method.

There is a variety of other *iterative stochastic algorithms* for approximating solutions to dynamic programming problems that have been developed in the computer science and 'artificial intelligence' literatures on *reinforcement learning*. These methods include *Q-learning, temporal difference learning*, and *real time dynamic programming*. The general approach in all these methods is to iteratively update an estimate of the value function, and recursive versions of Monte Carlo integration methods are employed in order to avoid doing numerical integrations to calculate conditional expectations. Using methods adapted from the literature on *stochastic approximation*, it is possible to prove that these methods converge to the true value function in the limit as the number of iterations tends to infinity. A key assumption underlying the convergence proofs is that there is sufficient stochastic noise to ensure that all possible decisions and decision nodes are visited 'infinitely often'. The intuition of why such an assumption is necessary follows from the discussion in section "Theory": suppose

that at some state *s* an initial estimate of the value function for decision that is actually optimal happens to be so low that the action is deemed to be 'nonoptimal' relative to the initial estimate. If the agent does not 'experiment' sufficiently, and thus fails to choose suboptimal decisions infinitely often, the agent may fail to learn that the initial estimated value was an underestimate of the true value, and therefore the agent might never learn that the corresponding action really is optimal. There is a trade-off between learning and experimentation, of course. The literature on 'multi-armed bandits' (Gittins 1979) shows that a fully rational Bayesian decision-maker will generally not find it optimal to experiment infinitely often. As a result such an agent can fail to discover actions that are optimal in an *ex post* sense. However, this does not contradict the fact that their behaviour is optimal in an *ex ante* sense: rather, it is a reflection that learning and experimentation is a costly activity, and thus it can be optimal to be incompletely informed, a result that has been known as early as Wald (1947a). A nice feature of many of these methods, particularly the real time dynamic programming developed in Barto et al. (1995), is that these methods can be used in 'real time', that is, we do not have to 'pre-calculate' the optimal decision rule in 'offline' mode. All these algorithms result in steady improvement in performance with experience. Methods similar to these have been used to produce highly effective strategies in extremely complicated problems. An example is IBM's 'Deep Blue' computer chess strategy, which has succeeded in beating the world's top human chess player, Garry Kasparov. However, the level of computation and repetition necessary to 'train' effective strategies is hugely time consuming, and it is not clear that any of these methods succeed in breaking the curse of dimensionality. For further details on this literature, see Bertsekas and Tsitsiklis (1996). Pakes (2001) applies these methods to approximate Markov perfect equilibria in games with many players. All types of stochastic algorithms have the disadvantage that the approximate solutions can be 'jagged' and there is always at least a small probability that the converged solution can be far from the true solution. However, they may be the only feasible option in many complex, high-dimensional problems where deterministic algorithms (for example, the Pakes and McGuire 1994, algorithm for Markov perfect equilibrium) quickly become intractable due to the curse of dimensionality.

D

Empirical Dynamic Programming and the Identification Problem

The developments in numerical dynamic programming described in the previous section paved the way for a new, rapidly growing literature on empirical estimation and testing of SDPs and dynamic games. This literature began to take shape in the late 1970s, with contributions by Sargent (1978) on estimation of dynamic labour demand schedules in a linear quadratic framework, and Hansen and Singleton (1982), who developed a generalized method of moment estimation strategy for a class of continuous choice SDPs using the *Euler equation* as an *orthogonality condition*. About the same time, a number of papers appeared that provided different strategies for estimation and inference in *dynamic discrete choice models* including Gotz and McCall's (1980) model of retirements of air force pilots, Wolpin's (1984) model of a family's decision whether or not to have a child, Pakes's (1986) model of whether or not to renew a patent, and Rust's (1987) model of whether or not to replace a bus engine. Since 1987, hundreds of different empirical applications of dynamic programming models have been published. For surveys of this literature see Eckstein and Wolpin (1989), Rust (1994), and the very readable book by Adda and Cooper (2003) – which also provides accessible introductions to the theory and numerical methods for dynamic programming. The remainder of this section will provide a brief overview of estimation methods and a discussion of the identification problem.

In econometrics, the term *structural estimation* refers to a class of methods that tries to go beyond simply summarizing the behaviour of

economic agents by attempting to infer their underlying *preferences* and *beliefs*. This is closely related to the distinction between the *reduced-form* of an economic model and the underlying *structure* that 'generates' it. (Structural estimation methods were first developed at the Cowles Commission at Yale University, starting with attempts to structurally estimate the linear simultaneous equations model, and models of investment by firms. Frisch, Haavelmo, Koopmans, Marschak, and Tinbergen were among the earliest contributors to this literature.) The reason why one would want to do structural estimation, which is typically far more difficult (for example, computationally intensive) than reduced-form estimation, is having knowledge of underlying structure enables us to conduct *hypothetical/counterfactual policy experiments*. Reduced-form estimation methods can be quite useful and yield significant insights into behaviour, but they are limited to summarizing behaviour under the status quo. However, they are inherently limited in their ability to forecast how individuals change their behaviour in response to various changes in the environment, or in *policies* (for example, tax rates, government benefits, regulations, laws, and so on) that *change the underlying structure* of agents' decision problems. As long as it is possible to predict how different policies change the underlying structure, we can use dynamic programming to re-solve agents' SDPs under the alternative structure, resulting in corresponding decision rules that represent predictions of how their behaviour (and welfare) will change in response to the policy change.

The rationale for structural estimation was recognized as early as Marschak (1953); however, his message appears to have been forgotten until the issue was revived in Lucas's (1976) critique of the limitations of reduced-form methods for policy evaluation. An alternative way to do policy evaluation is via *randomized experiments* in which subjects are randomly assigned to the *treatment group* (where the 'treatment' is some alternative policy of interest) and the *control group* (who continue with the policy under the status quo). By comparing the outcomes in the treatment and control groups, we can assess the behavioural and welfare impacts of the policy change. However, human experiments can be very time consuming and expensive to carry out, whereas 'computational experiments' using a structural model are very cheap and can be conducted extremely rapidly. The drawback of the structural approach, though, is the issue of *credibility* of the structural model. If the structural model is *misspecified*, it can generate incorrect forecasts of the impact of a policy change. There are numerous examples of how structural models can be used to make policy predictions: see Todd and Wolpin (2005) for an example that compares the prediction of a structural model with the results of a randomized experiment, where the structural model is estimated using subjects from the control group, and *out-of-sample predictions* are made to predict the behavioural response by subjects in the treatment group. They show that the structural model results in accurate predictions of how the treatment group subjects responded to the policy change.

I illustrate the main econometric methods for structural estimation of SDPs in the case of a stationary infinite horizon Markovian decision problem, although all the concepts extend in a straightforward fashion to finite horizon, non-stationary and non-Markovian problems. Estimation requires a specification of the *data generating process*. Assume we observe N agents, and we observe agent i from time period $-\overline{T}_i$ to $\overline{T}_i$ (or via appropriate re-indexing, from $t = 1,\ldots, T_i$). Assume observations of each individual are independently distributed realizations from the controlled process $\{s_t,d_t\}$. However, while we assume that we can observe the decisions made by each agent, it is more realistic to assume that we only observe a subset of the agent's state s_t. If we partition $s_t = (x_t, \varepsilon_t)$, assume that the econometrician observes x_t but not ε_t, so this latter component of the state vector constitutes an *unobserved state variable*. Then the reduced-form of the SDP is the decision rule δ

$$d = \delta(x, \varepsilon), \tag{38}$$

since the decision rule embodies all the behavioural content of the SDP model. The

structure Λ consists of the objects $\Lambda = \{\beta, \rho, u(s, d), p(s'|s, d)\}$. Equation (10) specifies the mapping from the structure Λ into the reduced form, δ. The data-set consists of $\{(x_{i,t}, d_{i,t}), t = 1, \ldots, T_i, i = 1, \ldots, N$. The econometric problem is to infer the underlying structure Λ from our data on the observed states and decisions by a set of individuals. Although the decision rule is potentially a complicated nonlinear function of unobserved state variables in the reduced-form Eq. (38), it is often possible to consistently estimate the decision rule under weak assumptions as $N \to \infty$, or as $T_i \to \infty$ if the data consists only of a single agent or a small number of agents i who are observed over long intervals. Thus, the decision rule δ can be treated as a *known function* for purposes of a theoretical analysis of identification. The *identification problem* is the qst, *under what conditions is the mapping from the underlying structure* Λ *to the reduced form* 1 *to* 1 (*that is invertible*)? If this mapping is 1 to 1, we say that the structure is *identified* since in principle it can be inverted to uniquely determine the underlying structure Λ. In practice, we construct an *estimator* $\widehat{\Lambda}$ based on the available data and show that $\widehat{\Lambda}$ converges to the true underlying structure Λ as $N \to \infty$ and/or $T_i \to \infty$ for each i.

Unfortunately, rather strong a priori assumptions on the form of agents' preferences and beliefs are required in order to guarantee identification of the structural model. Rust (1994) and Magnac and Thesmar (2002) have shown that an important subclass of SDPs, *discrete decision processes* (DDPs), are *nonparametrically unidentified*. That is, if we are unwilling to make any *parametric* functional form assumptions about preferences or beliefs, then in general there are infinitely many different structures Λ consistent with any reduced form δ. In more direct terms, there are many different ways to *rationalize* any observed pattern of behaviour as being 'optimal' for different configurations of preferences and beliefs. It is likely that these results extend to continuous choice problems, since it is possible to approximate a continuous decision process (CDP) by a sequence of DDPs with expanding numbers of elements in their choice sets. Further, for dynamic games, Ledyard (1986) has shown that *any* undominated strategy profile can be a Bayesian equilibrium for some set of preferences and beliefs. Thus, the hypothesis of optimality or equilibrium per se does not have testable empirical content: further a priori assumptions must be imposed in order for SDPs models to be identified and result in empirically testable restrictions on behaviour.

There are two main types of identifying assumptions that have been made in the literature to date: (*a*) parametric functional form assumptions on preferences $u(s,d)$ and components of agents' beliefs $p(s'|s, d)$ that involve unobserved state variables ε and (*b*) *rational expectations*. Rational expectations states that an agent's *subjective beliefs* $p(s'|s, d)$ coincide with *objective probabilities* that can be estimated from data. Of course, this restriction is useful only for those components of s, x, that the econometrician can actually observe. In addition, there are other more general *functional restrictions* that can be imposed to help identify the model. One example is monotonicity and shape restrictions on preferences (for example, concavity and monotonicity of the utility function), and another example is independence or *conditional independence* assumptions about variables entering agents' beliefs. I will provide specific examples below; however, it should be immediately clear why these additional assumptions are necessary.

For example, consider the two parameters ρ (the agent's subjective survival probability) and β (the agent's subjective discount factor). We have seen in section "Theory" that *only the product of* ρ *and* β *enter the SDP model, and not* ρ *and* β *separately*. Thus, at most the product $\rho\beta$ can be identified, but without further assumptions it is impossible to separately identify the subjective survival probability ρ from the subjective discount factor β since both affect an agent's behaviour in a symmetrical fashion. However, we can separately identify ρ and β if we assume that an individual has *rational survival expectations*, that is, that their subjective survival probability ρ coincides with the 'objective' survival probability. Then we can estimate ρ 'outside' the SDP model, using data on the lifetime distributions of similar types of agents, and then β can be

identified if other restrictions are imposed to guarantee that the product $\rho\beta$ is identified. However, it can be very difficult to make precise inferences about agents' discount factors in many problems, and it is easy to think of models where there is heterogeneity in survival probabilities and discount factors, and unobserved variables affecting one's beliefs about them (for example, family characteristics such as a predisposition for cancer, and so on, that are observed by an agent but not by the econometrician) where identification is problematic.

There are two main approaches for conducting inference in SDPs: (*a*) maximum likelihood and (*b*) 'simulation estimation'. The latter category includes a variety of similar methods such as *indirect inference* (Gourieroux and Monfort 1997), *simulated method of moments* (McFadden 1989; Gallant and Tauchen 1996), *simulated maximum likelihood and method of simulated scores* (see simulation-based estimation), and *simulated minimum distance* (Hall and Rust 2006). To simplify the discussion I will define these initially for single agent SDPs and at the end discuss how these concepts naturally extend to dynamic games. I will illustrate maximum likelihood and show how a likelihood can be derived for a class of DDPs; however, for CDPs, it is typically much more difficult to derive a likelihood function, especially when there are issues of *censoring*, or problems involving mixed discrete and continuous choice. In such cases simulation estimation is often the only feasible way to do inference.

For discrete decision processes, assume that the utility function has the following parametric, *additively separable* representation

$$u(x,\varepsilon,d) = u(x,d,\theta_1) + \varepsilon(d)\,(\text{AS}). \tag{39}$$

where $\varepsilon = \{\varepsilon(d)|d \in D(x)\}$, and $\varepsilon(d)$ is interpreted as an unobserved component of utility associated with choice of alternative $d \in D(x)$. Further, suppose that the transition density $p(x', \varepsilon'|x, \varepsilon, d)$ satisfies the following *conditional independence assumption*

$$p(x',\varepsilon'|x,\varepsilon,d) = p(x'|x,d,\theta_2)q(\varepsilon',\theta_3)(\text{CI}). \tag{40}$$

The CI assumption implies that $\{\varepsilon_t\}$ is an *IID* 'noise' process that is independent of $\{x_t, d_t\}$. Thus all of the serially correlated dynamics in the state variables are captured by the observed component of the state vector x_t. If, in addition, $q(\varepsilon_t, \theta_3)$ is a distribution with unbounded support with finite absolute first moments, one can show that the following *conditional choice probabilities* exist

$$P(d|x,\theta) = \int I_\varepsilon\{d = \delta(x,\varepsilon,\theta)\}q(\varepsilon)d\varepsilon, \tag{41}$$

where $\theta = (\rho, \beta, \theta_1, \theta_2, \theta_3)$ constitute the vector of unknown parameters to be estimated. (Identification of fully parametric models is a 'generic' property, that is, if there are two different parameters θ that produce the same conditional choice probability $P(d|x, \theta)$ for all x and $d \in D(x)$ – and thus led to the same limiting expected log-likelihood – small perturbations in the parameterization will 'almost always' result in a nearby model for which θ is uniquely identified.) In general, the parametric functional form assumptions, combined with the assumption of rational expectations and the AS and CI assumptions, are sufficient to identify the unknown parameter vector θ^*. θ^* can be estimated by maximum likelihood, using the *full information* likelihood function L_f given by

$$\begin{aligned}\mathcal{L}_f\left(0|\{x_{i,t},d_{i,t}\},t=\ 1,\ldots,T_i,\ \ i=1,\ldots,N\right)\\ =\prod_{i=1}^{N}\prod_{t=2}^{T_i}P\left(d_{i,t}|x_{i,t},\theta\right)\times p\left(x_{i,t}|x_{i,t-1},d_{i,t-1},\theta_2\right).\end{aligned} \tag{42}$$

A particularly tractable special case is where $q(\varepsilon, \theta_3)$ has a *multivariate extreme value distribution* where θ_3 is a common scale parameter (linearly related to the standard deviation) for each variable in this distribution (see McFadden, Daniel; logit models of individual choice for the exact formula for this density). This specification leads to a dynamic generalization of the *multinomial logit model*

$$P(d|x,\theta) = \frac{\exp\{v(x,d,\theta)/\theta_3\}}{\sum_{d'\ \in\ D(x)}\exp\{v(x,d',\theta)/\theta_3\}}, \tag{43}$$

where $v(x, d, \theta)$ is the expected, discounted utility from taking action d in observed state x given by the unique fixed point to the following *smoothed Bellman equation*

$$v(x,d,\theta) = + \rho\beta \int_{x'} \theta_3 \log\left(\sum_{d' \in D(x')} \exp\{v(x',d',\theta)/\theta_3\} \right) \times p(x'|x,d,\theta_2)dx'. \tag{44}$$

Define $\Gamma_\theta$by

$$\Gamma_\theta(W)(x,d) = u(x,d,\theta_1) + \rho\beta \int_{x'} \theta_3 \log\left(\sum_{d' \in D(x')} \exp\{\exp\{W(x',d',\theta)/\theta_3\} \right) \times p(x'|x,d,\theta_2)dx'. \tag{45}$$

It is not hard to show that under weak assumptions Γ_θ is a contraction mapping, so that $v(x, d, \theta)$ exists and is unique. Maximum likelihood estimation can be carried out using a nested *fixed point* maximum likelihood algorithm consisting of an 'outer' optimization algorithm to search for a value of θ that maximizes $\mathcal{L}_f(\theta)$, and an 'inner' fixed point algorithm that computes $v_\theta = \Gamma_\theta(v_\theta)$ each time the outer optimization algorithm generates a new trial guess for θ. The implicit function theorem guarantees that v_θ is a smooth function of θ. See Aguirregabiria and Mira (2004) for an ingenious alternative that 'swaps' the order of the inner and outer algorithms of the nested fixed-point algorithm resulting in significant computational speedups. See also Rust (1988) for further details on the nested fixed-point algorithm and the properties of the maximum likelihood estimator, and Rust (1994) for a survey of alternative less efficient but computationally simpler estimation strategies.

As noted above, econometric methods for CDPs, that is, problems where the decision variable is continuous (such as firm investment decisions, price settings, or consumption/savings decisions) are harder, since there is no tractable, general specification for the way unobservable state variables to enter the decision rule that result in a *nondegenerate* likelihood function (that is, where the likelihood $\mathcal{L}(\theta)$ is non-zero for any data-set and any value of θ). For this reason, maximum likelihood estimation of CDPs is rare, outside certain special subclasses, such as linear quadratic CDPs (Hansen and Sargent 1980; Sargent 1981). However, simulation-based methods of inference can be used in a huge variety of situations where a likelihood is difficult or impossible to derive. These methods have a great deal of flexibility, a high degree of generality, and often permit substantial computational savings. In particular, generalizations of McFadden's (1989) *method of simulated moments* (MSM) have enabled estimation of a wide range of CDPs. The MSM estimator minimizes a quadratic form between a set of moments constructed from the data, h_N and a vector of *simulated moments* $h_{N,S}(\theta)$, that is

$$h_N = \frac{1}{N}\sum_{i=1}^{N} h(\{x_{it}, d_{it}\})$$
$$h_{N,S}(\theta) = \frac{1}{S}\sum_{j=1}^{S}\frac{1}{N}\sum_{i=1}^{N} h\left(\left\{\tilde{x}_{it}^j(\theta), \tilde{d}_{it}^j(\theta)\right\}\right) \tag{46}$$

where h is a vector of $J \geq K$ 'moments' (that is, functionals of the data that the econometrician is trying to 'match'), where K is the dimension of θ, $\{x_{it}, d_{it}\}$ are the data, and $\left\{\tilde{x}_{it}^j(\theta), \tilde{d}_{it}^j(\theta)\right\}, j = 1, \ldots, S$ are S *IID* realizations of the controlled process.

The estimate $\widehat{\theta}$ is given by

$$\widehat{\theta} = \underset{\theta \in R^K}{\operatorname{argmin}} \left[h_N - h_{N,S}(\theta)\right]' W_N \left[h_N - h_{N,S}(\theta)\right], \tag{47}$$

where W_N is a $J \times J$ positive–definite weighting matrix. The most efficient choice for W_N is $W_N = \left[\widehat{\Omega}_N\right]^{-1}$ where $\widehat{\Omega}_N$ is the variance-covariance matrix formed from the vector of sample moments h_N. Simulation estimators require a nested fixed-

point algorithm since each time the outer minimization algorithm tries a new trial value for θ, the inner fixed point problem must be called to solve the CDP problem, using the optimal decision rule $d_{it}^j(\theta) = \delta\left(x_{it}^j, \varepsilon_{it}^j, \theta\right)$ to generate the simulated decisions, and the transition density $p\left(x_{i,t+1}^j, \varepsilon_{i,t+1}^j | x_{i,t}^j, \varepsilon_{i,t}^j, d_{i,t}^j, \theta_2\right)$ to generate $j = 1,\ldots, S$ *IID* realizations for a simulated panel each potential value of θ. (It is important to simulate using 'common random numbers' that remain fixed as θ varies over the course of the estimation, in order to satisfy the *stochastic equicontinuity conditions* necessary to establish consistency and asymptotic normality of the simulation estimator.)

Simulation methods are extremely flexible for dealing with a number of data issues such as attrition, missing data, censoring and so forth. The idea is that, if we are willing to build a stochastic model of the data 'problem', we can account for it in the process of simulating the behavioural model. For example, Hall and Rust (2006) develop a dynamic model of commodity price speculation in the steel market. An object of interest is to estimate the stochastic process governing wholesale steel prices; however, there is no public commodity market where steel is traded and prices are recorded on a daily basis. Instead, Hall and Rust observe only the actual wholesale prices of a particular steel trader, who records wholesale prices only on the days he actually buys steel in the wholesale market. Since the speculator makes money by 'buying low and selling high', the set of observed wholesale prices are *endogenously sampled*, and failure to account for this can lead to incorrect inferences about wholesale prices – a dynamic analogue of *sample selection bias*. However, in a simulation model it is easy to censor the simulated data in the same way it is censored in the actual data, that is, by discarding simulated wholesale prices on days where no simulated purchases are made. Hall and Rust show that even though moments based on the observed (censored) data are 'biased' estimates, the simulated moments are biased in exactly the same fashion, so minimizing the distance between actual and simulated biased moments nevertheless results in consistent and asymptotically normal estimates of the parameters of the wholesale price process and other parameters entering the speculator's objective function.

Simulation methods have also enabled the use of Bayesian methods, resulting in methods of inference that do not require asymptotic approximations, although they generally use Markov chain Monte Carlo methods to generate simulated draws from a distribution that approximates the exact finite sample posterior distribution for the parameters of interest (see for example, Lancaster 1997; Imai et al. 2005; Nourets 2006).

The most recent literature has extended the methods for estimation of single-agent SDPs to multi-agent dynamic games. For example, Rust (1994) described applications of dynamic discrete choice models to multiple-agent *discrete dynamic games*. The unobserved state variables ε_t entering any particular agent's payoff function are assumed to be unobserved both by the econometrician and by the other players in the game. The *Bayesian–Nash equilibria* of this game can be represented as a vector of conditional choice probabilities $(P_1(d_1|x), \ldots, P_n(d_n|x))$, one for each player, where $P_i(d_i|x)$ represents the econometrician's and the other players' beliefs about the probability player i will take action d_i, 'integrating out' over the unobservable states variable $\varepsilon_{i,t}$ affecting player i's decision at time t similar to Eq. (41) for single-agent problems. If one adapts the numerical methods for Markov-perfect equilibrium described in section "Numerical dynamic programming and the curse of dimensionality", it is possible to compute Bayesian–Nash equilibria of discrete dynamic games using nested fixed-point algorithms. While it is relatively straightforward to write down the likelihood function for the game, actual estimation via a straightforward application of full information maximum likelihood is extremely computationally demanding since it requires a *doubly nested fixed point algorithm* (that is, an 'outer' algorithm to search over θ to maximize the likelihood, and then an inner algorithm to solve the dynamic game for each value of θ, but this inner algorithm is itself a nested fixed-point algorithm). Alternative, less computationally demanding estimation methods

have been proposed by Aguirregabiria and Mira (2007), Bajari and Hong (2006), Bajari et al. (2007), and Pesendorfer and Schmidt-Dengler (2003). This research is at the current frontier of development in numerical and empirical applications of dynamic programming.

Besides econometric methods, which are applied for structural estimation for actual agents in their 'natural' settings, an alternative approach is to try to make inferences about agents' preferences and beliefs (and even their 'mode of reasoning') for artificial SDPs in a laboratory setting. The advantage of a laboratory experiment is *experimental control over preferences and beliefs*. The ability to control these aspects of decision-making can enable much tighter tests of theories of decision-making. For example, Binmore et al. (2002) structured a laboratory experiment to determine whether individuals do backward induction in one- and two-stage alternating offer games, and 'find systematic violations of backward induction that cannot be explained by payoff-interdependent preferences' (2002, p. 49).

Comments

There has been tremendous growth in research related to dynamic programming since the 1940s. The method has evolved into the main tool for solving sequential decision problems, and research related to dynamic programming has led to fundamental advances in theory, numerical methods and econometrics. As we have seen, while dynamic programming embodies the notion of rational decision-making under uncertainty, there is mixed evidence as to whether it provides a good literal description of how human beings actually behave in comparable situations. Although human reasoning and decision-making is undoubtedly both more complex and more 'frail' and subject to foibles and limitations than the idealized notion of 'full rationality' that dynamic programming embodies, the discussion of the identification problem shows that, if we are given sufficient flexibility about how to model individual preferences and beliefs, there exist SDPs whose decision rules provide arbitrarily good approximations to individual behaviour.

Thus, dynamic programming can be seen as a useful 'first approximation' to human decision-making, but it will undoubtedly be superseded by more descriptively accurate psychological models. Indeed, in the future one can imagine behavioural models that are not derived from some a priori axiomatization of preferences, but will result from empirical research that will ultimately deduce human behaviour from yet even deeper 'structure', that is the very underlying neuroanatomy of the human brain.

Even if dynamic programming is unlikely to be a descriptively accurate model of *human* decision-making, it will probably still remain highly relevant for the foreseeable future as the embodiment of *rational* decision-making. There are well-defined problems, for example, profit maximization or cost minimization, where there is agreement on the objective function to be maximized or minimized, and where there will be a demand for dynamic programming methods to find the optimal profit- or cost-minimizing strategies. There are many examples of this in the operations research literature. Practical applications include optimal inventory management (Hall and Rust 2006) and optimal harvesting of timber (Paarsch and Rust 2007).

Some observers such as Kurzweil (2005) predict that in the not too distant future (for example, approximately 2050) a *singularity* will occur, 'during which the pace of technological change will be so rapid, its impact so deep, that human life will be irreversibly transformed' (2005, p. 7). The singularity is a complex of accelerating improvements in computer hardware and software, and a merger of machine- and biological-based intelligence that will blur the distinction between 'artificial intelligence' and human intelligence, that will overcome many of current limitations of the human brain and human reasoning: 'By the end of this century, the nonbiological portion of our intelligence will be trillions and trillions of times more powerful than unaided human intelligence' (2005, p. 9). Dynamic programming will undoubtedly continue to be a critical tool in this brave new world.

Whether this prognosis will ever come to pass, or come to pass as soon as Kurzweil forecast, is

debatable; but it does suggest that there will be continued interest in and research on dynamic programming. However, the fact that reasonably broad classes of dynamic programming problems are subject to a curse of dimensionality suggests that it may be too optimistic to think that human rationality will soon be superseded by 'artificial rationality'. While there are many complicated problems that we would like to solve by dynamic programming in order to understand what 'fully rational' behaviour actually looks like in specific situations, the curse of dimensionality still limits us to very simple 'toy models' that only very partially and simplistically capture the myriad of details and complexities we face in the real world. Although we now have a number of examples where artificial intelligence based on principles from dynamic programming outstrips human intelligence, for example computerized chess, all these cases are for very specific problems in very narrow domains. I believe that it will be a long time before technological progress in computation and algorithms produce truly general-purpose 'intelligent behaviour' that can compete successfully with human intelligence in widely varying domains and in the immensely complicated situations that we operate in every day. Despite all our psychological frailties and limitations, there is an important unanswered question of 'how do we do it?', and more research is required to determine if human behaviour is simply suboptimal, or whether the human brain uses some powerful implicit 'algorithm' to circumvent the curse of dimensionality that digital computers appear to be subject to for solving problems such as SDPs by dynamic programming. For a provocative theory that deep principles of quantum mechanics can enable human intelligence to transcend computational limitations of digital computers, see Penrose (1989).

See Also

- ▶ Bellman Equation
- ▶ Game Theory
- ▶ Logit Models of Individual Choice
- ▶ McFadden, Daniel (Born 1937)
- ▶ Recursive Competitive Equilibrium
- ▶ Recursive Preferences
- ▶ Sequential Analysis
- ▶ Simulation-Based Estimation

Bibliography

This article has benefited from helpful feedback from Kenneth Arrow, Daniel Benjamin, Larry Blume, Moshe Buchinsky, Larry Epstein, Chris Phelan and Arthur F. Veinott, Jr.

Adda, J., and R. Cooper. 2003. *Dynamic economics quantitative methods and applications*. Cambridge, MA: MIT Press.

Aguirregabiria, V., and P. Mira. 2004. Swapping the nested fixed point algorithm: A class of estimators for discrete Markov decision models. *Econometrica* 70: 1519–1543.

Aguirregabiria, V., and P. Mira. 2007. Sequential estimation of dynamic discrete games. *Econometrica* 75: 1–53.

Arrow, K.J., D. Blackwell, and M.A. Girshik. 1949. Bayes and minimax solutions of sequential decision problems. *Econometrica* 17: 213–244.

Bajari, P., and H. Hong. 2006. *Semiparametric estimation of a dynamic game of incomplete information*, Technical Working Paper No. 320. Cambridge, MA: NBER.

Bajari, P., L. Benkard, and J. Levin. 2007. Estimating dynamic models of imperfect competition. *Econometrica* 75: 1331–1370.

Barto, A.G., S.J. Bradtke, and S.P. Singh. 1995. Learning to act using real-time dynamic programming. *Artificial Intelligence* 72: 81–138.

Bellman, R. 1957. *Dynamic programming*. Princeton: Princeton University Press.

Bellman, R. 1984. *Eye of the hurricane*. Singapore: World Scientific.

Bellman, R., and S. Dreyfus. 1962. *Applied dynamic programming*. Princeton: Princeton University Press.

Bertsekas, D.P. 1995. *Dynamic programming and optimal control*, vols 1 and 2. Belmont: Athena Scientific.

Bertsekas, D.P., and J. Tsitsiklis. 1996. *Neuro-dynamic programming*. Belmont: Athena Scientific.

Bhattacharya, R.N., and M. Majumdar. 1989. Controlled semi-Markov models – The discounted case. *Journal of Statistical Planning and Inference* 21: 365–381.

Binmore, K., J. McCarthy, G. Ponti, L. Samuelson, and A. Shaked. 2002. A backward induction experiment. *Journal of Economic Theory* 104: 48–88.

Blackwell, D. 1962. Discrete dynamic programming. *Annals of Mathematical Statistics* 33: 719–726.

Blackwell, D. 1965a. Positive dynamic programming. *Proceedings of the 5th Berkeley Symposium* 3: 415–428.
Blackwell, D. 1965b. Discounted dynamic programming. *Annals of Mathematical Statistics* 36: 226–235.
Cayley, A. 1875. Mathematical qsts and their solutions. Problem No. 4528. *Educational Times* 27, 237.
Chow, C.S., and J.N. Tsitsiklis. 1989. The complexity of dynamic programming. *Journal of Complexity* 5: 466–488.
Denardo, E. 1967. Contraction mappings underlying the theory of dynamic programming. *SIAM Review* 9: 165–177.
Dvoretzky, A., J. Kiefer, and J. Wolfowitz. 1952. The inventory problem: I. Case of known distributions of demand. *Econometrica* 20: 187–222.
Eckstein, Z., and K.I. Wolpin. 1989. The specification and estimation of dynamic stochastic discrete choice models: A survey. *Journal of Human Resources* 24: 562–598.
Gallant, A.R., and G.E. Tauchen. 1996. Which moments to match? *Econometric Theory* 12: 657–681.
Gihman, I.I., and A.V. Skorohod. 1979. *Controlled stochastic processes*. New York: Springer.
Gittins, J.C. 1979. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society B* 41: 148–164.
Gotz, G.A., and J.J. McCall. 1980. Estimation in sequential decision-making models: A methodological note. *Economics Letters* 6: 131–136.
Gourieroux, C., and A. Monfort. 1997. *Simulation-based methods of inference*. Oxford: Oxford University Press.
Grüne, L., and W. Semmler. 2004. Using dynamic programming with adaptive grid scheme for optimal control problems in economics. *Journal of Economic Dynamics and Control* 28: 2427–2456.
Hall, G., and J. Rust. 2006. *Econometric methods for endogenously sampled time series: The case of commodity price speculation in the steel market*. Manuscript: Yale University.
Hansen, L.P., and T.J. Sargent. 1980. Formulating and estimating dynamic linear rational expectations models. *Journal of Economic Dynamics and Control* 2: 7–46.
Hansen, L.P., and K. Singleton. 1982. Generalized instrumental variables estimation of nonlinear rational expectations models. *Econometrica* 50: 1269–1281.
Howard, R.A. 1960. *Dynamic programming and Markov processes*. New York: Wiley.
Imai, S., N. Jain, and A. Ching. 2005. *Bayesian estimation of dynamic discrete choice models*. Manuscript: University of Illinois.
Judd, K. 1998. *Numerical methods in economics*. Cambridge, MA: MIT Press.
Keane, M., and K.I. Wolpin. 1994. The solution and estimation of discrete choice dynamic programming models by simulation: Monte Carlo evidence. *Review of Economics and Statistics* 76: 648–672.
Kurzweil, R. 2005. *The singularity is near when humans transcend biology*. New York: Viking Press.
Kushner, H.J. 1990. Numerical methods for stochastic control problems in continuous time. *SIAM Journal on Control and Optimization* 28: 999–1048.
Lancaster, A. 1997. Exact structural inference in optimal job search models. *Journal of Business Economics and Statistics* 15: 165–179.
Ledyard, J. 1986. The scope of the hypothesis of Bayesian equilibrium. *Journal of Economic Theory* 39: 59–82.
Lucas Jr., R.E. 1976. Econometric policy evaluation: A critique. In *The phillips curve and labour markets*, Carnegie-Rochester Conference on Public Policy, ed. K. Brunner and A.K. Meltzer. Amsterdam: North-Holland.
Lucas Jr., R.E. 1978. Asset prices in an exchange economy. *Econometrica* 46: 1426–1445.
Luenberger, D.G. 1969. *Optimization by vector space methods*. New York: Wiley.
Magnac, T., and D. Thesmar. 2002. Identifying dynamic discrete decision processes. *Econometrica* 70: 801–816.
Marschak, T. 1953. Economic measurements for policy and prediction. In *Studies in econometric method*, ed. W.-C. Hood and T.J. Koopmans. New York: Wiley.
Massé, P. 1945. *Application des probabilités en chaîne á l'hydrologie statistique et au jeu des réservoirs. Report to the Statistical Society of Paris*. Paris: Berger-Levrault.
Massé, P. 1946. *Les réserves et la régulation de l'avenir*. Paris: Hermann.
McFadden, D. 1989. A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica* 57: 995–1026.
Nemirovsky, A.S., and D.B. Yudin. 1983. *Problem complexity and method efficiency in optimization*. New York: Wiley.
Nourets, A. 2006. *Inference in dynamic discrete choice models with serially correlated unobserved state variables*. Manuscript, University of Iowa.
Paarsch, H.J., and J. Rust. 2007. *Stochastic dynamic programming in space: An application to British Columbia forestry*. Working paper.
Pakes, A. 1986. Patents as options: Some estimates of the values of holding European patent stocks. *Econometrica* 54: 755–784.
Pakes, A. 2001. Stochastic algorithms, symmetric Markov perfect equilibria and the 'curse' of dimensionality. *Econometrica* 69: 1261–1281.
Pakes, A., and P. McGuire. 1994. Computing Markov perfect Nash equilibrium: Numerical implications of a dynamic differentiated product model. *RAND Journal of Economics* 25: 555–589.
Penrose, R. 1989. *The emperor's new mind*. New York: Penguin.
Pesendorfer, M., and P. Schmidt-Dengler. 2003. *Identification and estimation of dynamic games*. Manuscript, University College London.
Pollard, D. 1989. Asymptotics via empirical processes. *Statistical Science* 4: 341–386.

Puterman, M.L. 1994. *Markovian decision problems*. New York: Wiley.
Rust, J. 1985. Stationary equilibrium in a market for durable goods. *Econometrica* 53: 783–805.
Rust, J. 1987. Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher. *Econometrica* 55: 999–1033.
Rust, J. 1988. Maximum likelihood estimation of discrete control processes. *SIAM Journal on Control and Optimization* 26: 1006–1024.
Rust, J. 1994. Structural estimation of Markov decision processes. In *Handbook of econometrics*, vol. 4, ed. R.F. Engle and D.L. McFadden. Amsterdam: North-Holland.
Rust, J. 1996. Numerical dynamic programming in economics. In *Handbook of computational economics*, ed. H. Amman, D. Kendrick, and J. Rust. Amsterdam: North-Holland.
Rust, J. 1997. Using randomization to break the curse of dimensionality. *Econometrica* 65: 487–516.
Rust, J., and G.J. Hall. 2007. The (*S*, *s*) rule is an optimal trading strategy in a class of commodity price speculation problems. *Economic Theory* 30: 515–538.
Rust, J., and C. Phelan. 1997. How social security and medicare affect retirement behavior in a world with incomplete markets. *Econometrica* 65: 781–832.
Rust, J., J.F. Traub, and H. Woźniakowski. 2002. Is there a curse of dimensionality for contraction fixed points in the worst case? *Econometrica* 70: 285–329.
Santos, M., and J. Rust. 2004. Convergence properties of policy iteration. *SIAM Journal on Control and Optimization* 42: 2094–2115.
Sargent, T.J. 1978. Estimation of dynamic labor demand schedules under rational expectations. *Journal of Political Economy* 86: 1009–1044.
Sargent, T.J. 1981. Interpreting economic time series. *Journal of Political Economy* 89: 213–248.
Selten, R. 1975. Reexamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory* 4: 25–55.
Tauchen, G., and R. Hussey. 1991. Quadrature-based methods for obtaining approximate solutions to nonlinear asset pricing models. *Econometrica* 59: 371–396.
Todd, P., and K.I. Wolpin. 2005. *Ex ante evaluation of social programs*. Manuscript, University of Pennsylvania.
Traub, J.F., and A.G. Werschulz. 1998. *Complexity and information*. Cambridge: Cambridge University Press.
Von Neumann, J., and O. Morgenstern. 1944. *Theory of games and economic behavior*. Princeton: Princeton University Press. 3rd edn, 1953.
Wald, A. 1947a. Foundations of a general theory of sequential decision functions. *Econometrica* 15: 279–313.
Wald, A. 1947b. *Sequential analysis*. New York: Dover.
Wald, A., and J. Wolfowitz. 1948. Optimum character of the sequential probability ratio test. *Annals of Mathematical Statistics* 19: 326–339.
Wolpin, K. 1984. An estimable dynamic stochastic model of fertility and child mortality. *Journal of Political Economy* 92: 852–874.

Dynamic Programming and Markov Decision Processes

Steven A. Lippman

A great many problems in economics can be reduced to determining the maximum of a given function. Dynamic programming is one of a number of mathematical optimization techniques applicable in such problems. As will be illustrated, the dynamic programming technique or viewpoint is particularly useful in complex optimization problems with many variables in which time plays a crucial role. Unlike calculus-based techniques it does not require the function being optimized to be differentiable in the (decision) variables.

In a nutshell, dynamic programming is a mathematical approach designed for analysing decision processes in which the multi-stage or sequential character of the process is prominent. In particular, dynamic programming is likely to be applicable whenever an economic agent makes a sequence of decisions (such as how much to consume in year i) in a prespecified order (year i's decision is made prior to year $i + 1$'s decision). In contrast to the familiar first-order condition which, for example, balances marginal revenues against marginal costs, the orientation of the typical dynamic programming approach entails balancing current profits against all future profits. In so doing, it transforms a complex n variable optimization problem into n simple one variable optimization problems.

While it is important to understand the conditions under which this approach induces a computationally attractive technique, economists' interest in dynamic programming emanates from its analytic rather than computational power. As empiricists we are interested in the numbers (for example, the optimal amount consumed in period 1), but wearing our theoretical or policy-making hats we are more interested in the intrinsic structure of the solution (for example, the optimal amount consumed in period 1 decreases in

response to an increase in the riskiness associated with the income stream). With this in mind, we emphasize the use of dynamic programming as a conceptual framework enabling us to understand the nature of the solution to the decision-maker's problem.

The basic components describing a multi-stage decision process are states, stages, actions (decisions), rewards, state transitions or law of motion, and constraints. The relevant dynamic programming concepts are those of policy, return function, and functional equation. For pedagogical purposes we define and explain these objects in the context of consumption under uncertainty.

An Example

At the beginning of each of N periods (of equal length such as one year) our economic agent must decide how much of his current wealth to consume and how much to save. After making his consumption decision, his remaining wealth is invested and experiences a (possibly random) return of R per unit of capital. Assume $R \geq 0$ so that losing his entire investment is the worst that can happen. The agent is gainfully employed; accordingly, at the end of each period his wealth is augmented by his (possibly random) labour income L. To keep matters simple, assume that the $2N$ random variables are independent and that the distributions of both return on capital and on labour do not change with time. The agent's goal is to maximize the expected discounted utility of his consumption stream. The literature on this topic usually postulates that the agent's utility function is separable in time and that $u(c)$, the utility of consuming c units of capital in a given period, does not change with time and is strictly concave and strictly increasing. Denoting the one-period discount factor by $\beta > 0$, $\sum_{i=1}^{N} \beta^{i-1} u(c_i)$ is the utility associated with the consumption stream $c_1, c_2, \ldots, c_N$. Setting the agent's initial endowment at w_1, the problem specification is complete.

The *state* of the system or process is the agent's wealth, and the set of possible states, called the *state space,* is the nonnegative numbers. (Although typically the state is a real number or a vector of real numbers, occasionally the state is a more complicated object such as a probability measure.) The points in time when decisions must be made divide the process into *stages.* In this example, each period is a stage. The *action* or decision at each stage is how much to consume. Given a wealth of w, i.e., the agent finds himself in state w, the agent's consumption level c must satisfy the constraint $0 \leq c \leq w$. The agent's *action space* is $[0, w]$ and reflects the fact that his consumption cannot be negative and is not permitted to exceed his current wealth. (Thus, borrowing against future income is prohibited.)

In our consumption example the objective function or overall return (the functional being maximized, here $\sum_{i=1}^{N} \beta^{i-1} u(c_i)$) is additively separable in the consumption levels $c_1, c_2, \ldots, c_N$: the change in the overall return associated with a change in c_i to $\hat{c}_i$ is $\beta^{i-1}[u(\hat{c}_i) - u(c_i)]$ and does not depend upon any of the other consumption levels. (Of course, altering c_i has an impact upon the future wealth levels.) Consequently, we can speak of the *one-period reward function u.* Given the current state (wealth level) w, the system passes (the state is transformed) to a new state $T(w, c)$ in response to the action (consumption decision) c selected. The new state $T(w, c)$ is simply the state of the system at the beginning of the next stage or period. Thus $w_{j+1} = T(w_j, c)$ where w_j is the state of the system in period j. In our example, the *law of motion or transition function* is

$$T(w, c) = (w - c)R + L. \tag{1}$$

It reflects the facts that labour income is unaffected by either consumption or the return on investment and that the investment has constant returns to scale.

It is often more convenient to label time backwards and to speak of the number of stages remaining. Accordingly, define $V_n(w)$ to be the maximum expected discounted utility obtainable when n stages remain and the current wealth is w, $n = 1, 2, \ldots$ and $w \geq 0$; V_n is called the n-period

return function. Our immediate goal is to write a *functional equation* or recursive formula relating V_n to V_{n-1}.

In order to obtain our recursive relation, we shall employ implicitly Bellman's famous Principle of Optimality. Bellman stated it thus:

> An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision. (1, p. 83)

Our use of the Principle of Optimality will become clear as the development of the function equation (3) unfolds. To facilitate the connection between the functional equation and Bellman's Principle of Optimality one last piece of terminology requires introduction. A *policy* is a rule which specifies the decisions to be made as the system passes through the various states. Of course the decisions specified must be *feasible* in that they satisfy the system's constraints: each action selected must lie in the action space associated with the given state. While the action specified by the policy for state s at time n can be a (random) function of the history $(s_1, a_1, s_2, a_2, \ldots, s_{n-1}, a_{n-1}, s_n)$ of the system up to time n where s_i was the state at time i and a_i the action selected at time i, it is usually the case that attention can be restricted to policies under which the action specified for time n depends upon the history of the system only through the state of the system at time n; such a policy is referred to as a *Markov policy.* It may be helpful to think of a (Markov) policy as a contingency plan specifying the action to be selected if a given state is reached at a given stage rather than as a schedule of N actions that will occur. A policy is said to be an *optimal policy* if the return associated with using it equals the maximal return attainable.

Developing the Functional Equation

To begin, recall that the agent's utility function u is nondecreasing, whence he consumes all his wealth in period N when one stage remains:

$$V_1(w) = u(w), \quad \text{all} \quad w \geq 0 \tag{2}$$

To obtain an expression for V_2 in terms of V_1, note that the total return when two stages remain consists of the immediate reward $u(c)$ from the first stage plus the discounted return from the second stage. At the second stage the agent will have an amount $(w - c)R + L$ of wealth to allocate between consumption and saving; clearly, it must be allocated in the best possible manner – in this instance all of it is consumed – in order to obtain an optimal two stage allocation. Thus, given an initial consumption of c, an additional return of $\beta V_1[(w - c)\, R + L]$ is garnered if consumption in the final stage is chosen optimally. Therefore, the total expected return for the two stage process when c is the consumption in the first of the two stages is simply $u(c) + \beta E V_1[(w - c)\ \ R + L]$, where E denotes the expectation with respect to the random variables in the state description. Finally, by selecting the best consumption level in the first of the two stages we obtain the desired relationship between V_1 and V_2:

$$V_2(w) = \max_{0 \leq c \leq w} \{u(c) + \beta E V_1[(w - c)R + L]\}, \quad \text{all} \ \ w \geq 0. \tag{3a}$$

Utilizing the same logic, the return function V_2 can be employed to compute V_3 and, more generally, V_{n-1} can be used to compute V_n as follows:

$$V_n(w) = \max_{0 \leq c \leq w} \{u(w) + \beta E V_{n-1}[(w - c)R + L]\}, \quad n = 2, 3, \ldots, N, \ \ \text{all} \ \ w \geq 0. \tag{3b}$$

For each n and w, define $c_n(w)$ to be the largest value of c for which the maximum in (3) is attained. Then the policy which consumes the amount $c_n(w)$ when in state w with n stages remaining is an optimal policy. When the optimal return exists, it is unique, but it is often the case that there is more than one optimal policy.

Structure of the Optimal Policy

As noted earlier, the problem is not solved until the structure of an optimal policy is exhibited. Our analysis of the consumption problem is typical and illustrative of many analyses of finite planning horizon ($N < \infty$) dynamic programming problems; in particular, mathematical induction is critical to the analysis.

To ensure a positive level of consumption each period, assume $u'(0) = \infty$ and $L \geq \varepsilon > 0$ (so labour income is bounded away from zero whence $V_n(w) > -\infty$ for $w > 0$ as $Eu(L) \geq u(\varepsilon) > -\infty$). The first condition provides the incentive to consume and the second provides the capital. If R and/or β is small, the agent's optimal policy may entail the corner solution of consuming all of his wealth in a given period. Assume R and β are sufficiently large – e.g., $u'(\varepsilon) < \beta E(R)\, Eu'(L)$ – to ensure $c_n(w) < w$ for $n > 1$.

It is our intention to illustrate common analytical approaches as well as lay bare the structure of the optimal policy. In so doing the nature of the return function V_n is also characterized.

One technique often employed is that of computing the return of a suboptimal policy which mimics the actions of another (perhaps optimal) policy. Mimicking is the method of proof of the following minor result, which we facetiously interpret as 'life is worth living'.

Lemma 1 If $Eu(L) > 0$, then $V_{n+1} > V_n$.

Proof Let the consumption $\pi_i(w)$ dictated by policy π when i stages remain and the current wealth is w be specified as follows: $\pi_1(w) = w$ and $\pi_{i+1}(w)$ for $i \geq 1$. Thus, when $i + 1$ stages remain, π acts like an optimal policy if there were but i stages remaining. Consequently, π yields a return of $V_n(w) + \beta^n\, Eu\,(L) > (w)$. The result now follows as the optimal return $V_{n+1}(w)$ is at least as large as the return from using π for $n + 1$ stages. Q.E.D.

The return function often inherits properties of the oneperiod reward function. For example, using induction it can be shown the return function is strictly concave and strictly increasing like u. Doing so in this instance is a bit more difficult than usual.

Lemma 2 The return function V_n is strictly increasing and strictly concave. Consequently, there is a unique optimal policy: $c_n(w)$ is the unique optimal level of consumption.

Proof By (2) V_1 is trivially seen to be strictly increasing. Noting that $c_n\,(w + \delta)$ need not equal $c_n(w) + \delta$ for $\delta > 0$, we have

$$\begin{aligned} V_n(w+\delta) &\geq u[c_n(w)+\delta] \\ &+\beta E V_{n-1}\{[w-c_n(w)]R + L\} \\ &> u[c_n(w)] + \beta E V_{n-1}\{[w-c_n(w)]R + L\} \\ &= V_n(w) \end{aligned}$$

so V_n is strictly increasing.

Strict concavity is proved by induction. Clearly V_1 is strictly concave. Assume V_{n-1} is strictly concave. It is easy to demonstrate (see Heyman and Sobel 1984, p. 535) that the function $J(w,\ c) = V_{n-1}[(w–c)r + l]$ is jointly concave in w and c on the convex set $C = \{(w, c)\colon 0 \leq c \leq w,\ w > 0\}$. (The concavity is strict if $r > 0$.) As the sum of concave functions is itself concave, $EV_{n-1}[(w–c)R + L]$ is strictly concave on C as is $\overline{J}(w,c) = u(c) + \beta E V_{n-1}[(w-c)R + L]$. While the maximum of a set of concave functions need not be concave, a standard result (ibid. p. 525) reveals that V_n, the maximum of the jointly strictly concave function $\overline{J}$, is strictly concave on C. This completes the induction argument. Strict concavity of $\overline{J}(w,c)$ ensures uniqueness. Q.E.D.

The decreasing marginal utility of wealth readily implies that optimal consumption increases with, but not as quickly as, wealth.

Lemma 3 The optimal level of consumption $c_n(w)$ satisfies

$$0 < c_n(w+\Delta) - c_n(w) < \Delta \text{ for } \Delta > 0. \quad (4)$$

Proof For ease in presentation only, assume that u'' exists so that V''_{n-1} and $\mathrm{d}^2 EV_{n-1}[(w-c)R+L]/\mathrm{d}c^2$ both exist. Differentiating the first order condition $u'[c_n(w)] - \beta E\{RV'_{n-1}[(w-c_n(w))R+L]\} = 0$ with respect to w yields

$$\begin{aligned} & c'_n(w)u''[c_n(w)] - \left[1 - c'_n(w)\right] \\ & \quad \times E\left\langle R^2 V''_{n-1}\{[w-c_n(w)]R+L\}\right\rangle \\ & \quad = 0, \end{aligned} \tag{5}$$

As $u'' < 0, R^2 \geq 0$, and $V''_{n-1} < 0, c_n \leq 0$ would violate (5). Similarly, $c'_n \geq 1$ violates (5). Q.E.D.

In the context of a teenager's lament and the associated parental response, the two intuitive inequalities in (4) have the interpretations 'What is money for if not to spend?' and, 'Don't let it burn a hole in your pocket!'

The demonstration that the overall return increases with the time remaining offered in the proof of Lemma 1 was straightforward. Verifying that the marginal utility of wealth declines with the agent's age, whence consumption increases with age, entails the application of a frequently employed induction technique we call bootstrapping induction.

Lemma 4 The marginal utility of wealth decreases and the optimal level of consumption increases with age:

$$c_n(w) > c_{n+1}(w)$$

and

$$V'_n(w) < V'_{n+1}(w).$$

Proof For ease in presentation assume only that u'' exists. Corner solutions have been eliminated: the assumption $u'(\varepsilon) < \beta E(R)Eu'(L)$ ensures $c_n(w) < w$ for $n > 1$ whereas $u'(0) = \infty$ and $L \geq \varepsilon$ ensures $c_n(w) > 0$ for $n \geq 1$. Therefore, the marginal benefit $u'[c_n(w)]$ of immediate consumption equals the marginal benefit $\beta E\langle RV_n'_{-1}\{[w - c_n(w)]R + L\}\rangle$ of savings for $n > 1$. Consequently, regardless of the percentage of marginal increase in wealth the agent allocates to savings, we find

$$V'_n(w) = u'[c_n(w)], \quad n \geq 1. \tag{6}$$

From (2) and the guarantee of an interior solution for $n > 1$ we have $c_1(w) = w > c_2(w)$, whereas (6) and u' strictly decreasing yield $V_1{}'(w) = u'[c_1(w)] < u'[c_2(w)] = V_2{}'(w)$. Assume

$$c_n(w) > c_{n+1}(w), \quad \text{all} \quad w > 0 \tag{7a}$$

$$V'_n(w) < V'_{n+1}(w), \quad \text{all} \quad w > 0\,. \tag{7b}$$

Applying (7b) to the future return results in

$$\begin{aligned} & \frac{\mathrm{d}}{\mathrm{d}_c}\beta EV_n[(w-c)R+L] = \\ & -\beta E\{RV'_n[(w-c)R+L]\} > \\ & -\beta E\left\{RV'_{n+1}\left[(w-c)R+L = \frac{\mathrm{d}}{\mathrm{d}_c}\beta EV_{n+1}[(w-c)R+L]\right.\right. \end{aligned}$$

from which we obtain immediately

$$c_{n+1}(w) > c_{n+2}(w). \tag{8a}$$

Now (6), (8a), and u' strictly decreasing yield

$$\begin{aligned} V'_{n+1}(w) &= u'[c_{n+1}(w)] < u'[c_{n+2}(w)] \\ &= V_{n+2}(w). \end{aligned} \tag{8b}$$

Having established (8), the induction argument is complete. Q.E.D.

The impact on consumption of increased uncertainty (in the sense of second order stochastic dominance) in capital income R or labour income L has been a focal point of the literature which models the agent's allocation between immediate consumption and saving. Will the prospect of either uncertainty vis-à-vis certainty or increased uncertainty induce the agent to increase his immediate consumption as a hedge against the (increasingly) uncertain future in which nature herself may, in effect, consume his wealth, or will the agent decrease his immediate consumption in an attempt to provide against an adverse future? The former strategy adopts a 'get while the getting is good' philosophy while the latter evokes one of 'save for a rainy day'.

The best response to an increase in uncertainty depends on the shape of the utility function as well as the source of the uncertainty. On this account, the family $\{u_\gamma\}$ of utility functions with constant relative risk aversion plays an important role:

$$u_0(c) = \ln c \quad \text{and} \quad u_\gamma(c) = c^\gamma/\gamma, \quad \text{for} \quad \gamma < 1, \quad \gamma \neq 0.$$

When the utility function has a positive third derivative and there is pure income risk (i.e. R is a constant), consumption decreases in the face of an increase in risk (see Miller 1976). For $\gamma < 1$, $u_\gamma''' > 0$. When there is pure capital risk (i.e. L is a constant) and the utility function exhibits constant relative risk aversion, an increase in risk causes consumption to decrease if $\gamma < 0$ and increase if $\gamma > 0$ (see Phelps 1962 or Mirman 1971).

Of course the dynamic programming approach can be gainfully employed to address other interesting questions such as the conditions which imply capital will (on average) accumulate, whether V_n converges, and the impact of an uncertain lifetime. The second question is a recurring one in dynamic programming models; in this instance, V_N converges as $N \to \infty$ provided $\beta ER < 1$. If p_i is the probability the agent lives i or more years and $P_{N+1} = 0$, then the return function V_n satisfies (2) and (3) with $u(c)$ replaced by $P_{N-n+1}\, u(c)$ and an increase in consumption is the agent's response to an increase in the risk associated with his own longevity when there is neither income nor capital risk and $u = u_y$ for $u = u_\gamma$ for $\gamma < 1$ (see Levhari and Mirman 1977).

Markov Decision Processes

The consumption model was considered in unctuous detail for several reasons. It is intrinsically interesting to economists, it is relatively simple to describe, and its risk structure can be ascertained without undue effort. In addition, the analytical approach and the techniques employed as well as the formulation of the functional equation are standard fare in dynamic programming models. The most important reason, however, emanates from the fact that it is an example of the seemingly ubiquitous *Markov Decision Process* (MDP).

A discrete time MDP is a process that is observed at time points 0, 1, 2, ...; the k th observation finds the process to be in some states $s_k \in S$. When in state s at time n, an action $a \in A_s$ is chosen. As a result of this action a reward $r(s, a)$ is received and the next state of the process is determined according to the transition probability of a stationary Markov process. The objective is to maximize the sum of the expected discounted rewards. Thus, if S is indexed by the non-negative integers, the optimal return function V for this infinite stage process can be shown (see Ross 1983 for the standard proof when r is bounded and Lippman 1975 for appropriate conditions on r and P when r is unbounded) to satisfy the functional equation

$$V(i) = \max_{a \in A_i} \left\{ r(i,a) + \beta \sum_{j=0}^{\infty} P_{ij}(a) V(j) \right\}, \quad j = 0, 1, 2, \ldots \tag{9}$$

where $P_{ij}(a)$ is the conditional probability that the process will be in stage j at time $n + 1$ given that it was in state i at time n and action a was selected.

The theory of MDP, including the roles played by successive approximation and policy iteration, is rather extensive, though not by comparison with its host of applications. Excellent modern treatments emphasizing theory and computation, respectively, are given in Heyman and Sobel (1984) and Ross (1983) and in Denardo (1982). Bellman's original book (1957) on dynamic programming remains a very worthwhile read as does Howard's book (1960) on MDP.

See Also

- ▶ Optimal Control and Economic Dynamics
- ▶ Stochastic Optimal Control

Bibliography

Bellman, R. 1957. *Dynamic programming*. Princeton: Princeton University Press.

Blackwell, D. 1962. Discrete dynamic programming. *Annals of Mathematics and Statistics* 33: 719–726.

Blackwell, D. 1965. Discounted dynamic programming. *Annals of Mathematics and Statistics* 36: 226–235.

Denardo, E.V. 1967. Contraction mappings in the theory underlying dynamic programming. *SIAM Review* 9: 165–177.

Denardo, E.V. 1982. *Dynamic programming*. Englewood Cliffs: Prentice-Hall.

Hakansson, N.H. 1970. Optimal investment and consumption strategies under risk for a class of utility functions. *Econometrica* 38: 587–607.

Heyman, D., and M. Sobel. 1984. *Stochastic models in operations research*, vol. II. New York: McGraw-Hill.

Howard, R.A. 1960. *Dynamic programming and Markov processes*. Cambridge, MA: MIT Press.

Levhari, D., and L.J. Mirman. 1977. Savings and uncertainty with an uncertain horizon. *Journal of Political Economy* 85: 265–281.

Lippman, S. 1975. On dynamic programming with unbounded rewards. *Management Science* 21: 1225–1233.

Miller, B.L. 1974. Optimal consumption with a stochastic income stream. *Econometrica* 42: 253–266.

Miller, B.L. 1976. The effect on optimal consumption of increased uncertainty in labor income in the multi period case. *Journal of Economic Theory* 13: 154–167.

Mirman, L.J. 1971. Uncertainty and optimal consumption decisions. *Econometrica* 39: 179–185.

Phelps, E.S. 1962. The accumulation of risky capital: A sequential utility analysis. *Econometrica* 30: 729–743.

Ross, S. 1983. *Introduction to stochastic dynamic programming*. New York: Academic Press.