U

Uncertainty

Peter P. Wakker

Abstract

This article deals with individual decision making under uncertainty (unknown probabilities). Risk (known probabilities) is not treated as a separate case, but as a sub-case of uncertainty. Many results from risk naturally extend to uncertainty. The Allais paradox, commonly applied to risk, also reveals empirical deficiencies of expected utility for uncertainty. The Ellsberg paradox reveals deviations from expected utility in a relative, not an absolute, sense, giving *within-person* comparisons: for some events (ambiguous or otherwise) subjects deviate more from expected utility than for other events. Besides aversion, many other attitudes towards ambiguity are empirically relevant.

Keywords

Additive probabilities; Allais paradox; Allais, M.; Ambiguity and ambiguity aversion; Ambiguity attitudes; Bayesian statistics; Bernoulli, D.; Betweenness models; Concave utility; Convex analysis; Convex probability weighting; Convex utility; Cumulative prospect theory; De Finetti, B.; Decision theory; Decision under risk; Disappointment aversion theory; Ellsberg paradox; Existence of equilibria; Expected utility; Gambling; Greenspan, A.; Insurance; Kahneman, D.; Keynes, J. M.; Knight, F.; Known probabilities; Likelihood; Loss aversion; Moral hazard; Multiple priors model; Neuroeconomics; Non-expected utility; Objective probability; Pessimism; Probabilistic risk attitudes; Probabilistic sophistication; Probability; Prospect theory; Ramsay, F.; Rank dependence; Rank-dependent models; Rank-dependent utility; Reference dependence; Revealed preference; Risk; Risk aversion; Savage, L.; Sensitivity; Separability; Source preference; State spaces; Stochastic dominance; Subjective probability; Sure-thing principle; Tversky, A.; Uncertainty; Unknown probabilities; Wald, A.; Weighted utility; Within-person comparison

JEL Classifications

D8

In most economic decisions where agents face uncertainties, no probabilities are available. This point was first emphasized by Keynes (1921) and Knight (1921). It was recently reiterated by Greenspan (2004, p. 38):

> . . . how . . . the economy might respond to a monetary policy initiative may need to be drawn from evidence about past behavior during a period only

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

> roughly comparable to the current situation. . . . In pursuing a risk-management approach to policy, we must confront the fact that only a limited number of risks can be quantified with any confidence.

Indeed, we often have no clear statistics available. Knight went so far as to call probabilities unmeasurable in such cases. Soon after Knight's suggestion, Ramsey (1931), de Finetti (1931), and Savage (1954) showed that probabilities can be defined in the absence of statistics after all, by relating them to observable choice. For example, $P(E) = 0.5$ can be derived from an observed indifference between receiving a prize under event E and receiving it under not-E (the complement to E). Although widely understood today, the idea that something as intangible as a subjective degree of belief can be made observable through choice behaviour, and can even be quantified precisely, was a major intellectual advance.

Ramsey, de Finetti and Savage assumed that the agent, after having determined the probabilities subjectively (as required by some imposed rationality axioms), proceeds as under expected utility for given objective probabilities. The Allais (1953) paradox (explained later) revealed a descriptive difficulty: for known probabilities, people often do not satisfy expected utility. Hence, we need to generalize expected utility. Another, more fundamental, difficulty was revealed by the Ellsberg (1961) paradox (also explained later): for unknown probabilities, people behave in ways that cannot be reconciled with any assignment of subjective probabilities at all, so that further generalizations are needed. (The term 'subjective probability' always refers to additive probabilities in this article.)

Despite the importance and prevalence of unknown probabilities, understood since 1921, and the impossibility of modelling these through subjective probabilities, understood since Ellsberg (1961), decision theorists continued to confine their attention to decision under risk with given probabilities until the late 1980s. Wald's (1950) multiple priors model did account for unknown probabilities, but attracted little attention outside statistics.

As a result of an idea of David Schmeidler (1989, first version 1982), the situation changed in the 1980s. Schmeidler introduced the first theoretically sound decision model for unknown probabilities without subjective probabilities, called rank-dependent utility or Choquet expected utility. At the same time, Wald's multiple priors model was revived when Gilboa and Schmeidler (1989) established its theoretical soundness; a similar result was obtained independently by Chateauneuf (1991, Theorem 2). These discoveries provided the basis for non-expected utility with unknown probabilities that had been sorely missing since 1921. Since the late 1980s, the table has turned in decision theory. Nowadays, most studies concern unknown probabilities. Gilboa (2004) contains recent papers and applications. This article concentrates on conceptual issues of individual decisions in the possible absence of known probabilities.

Theoretical studies of non-expected utility have usually assumed risk aversion for known probabilities (leading to concave utility and convex probability weighting), and ambiguity aversion for unknown probabilities (Camerer and Weber 1992, section 2.3). These phenomena best fit with the existence of equilibria and can be handled using conventional tools of convex analysis (Mukerji and Tallon 2001). Empirically, however, a more complex fourfold pattern has been found. For gains with moderate and high likelihoods, and for losses with low likelihoods, risk aversion is prevalent indeed, but for gains with low likelihoods and for losses with high likelihoods the opposite, risk seeking, is prevalent.

The fourfold pattern resolves the classical paradox of the coexistence of gambling and insurance, and leads, for instance, to new views on insurance. Whereas all classical studies of insurance explain insurance purchasing through concave utility, empirical measurements of utility have suggested that utility is not very concave for losses, often exhibiting more convexity than concavity (surveyed by Köbberling et al. 2006). This finding is diametrically opposite to what has been assumed throughout the insurance literature.

According to modern decision theories, insurance is primarily driven by consumers' overweighting of small probabilities rather than by marginal utility.

The fourfold pattern found for risk has similarly been found for unknown probabilities, and usually to a more pronounced degree. Central qsts in uncertainty today concern how to analyse not only classical marginal utility but also new concepts such as probabilistic risk attitudes (how people process known probabilities), loss aversion and reference dependence (the framing of outcomes as gains and losses), and, further, states of belief and decision attitudes regarding unknown probabilities ('ambiguity attitudes').

We end this introduction with some notation and definitions. *Decision under uncertainty* concerns choices between *prospects* such as $(E_1 : x_1, \ \ldots, E_n : x_n)$, yielding *outcome* x_j if *event* E_j obtains, $j = 1, \ \ldots, n$. Outcomes are monetary. The E_js are events of which an agent does not know for sure whether they will obtain, such as who of n candidates will win an election. The E_js are mutually exclusive and exhaustive. No probabilities of the events need to be given. Because the agent is uncertain about which event obtains, he is uncertain about which outcome will result from the prospect, and has to make decisions under uncertainty.

Decision Under Risk and Non-expected Utility Through Rank Dependence

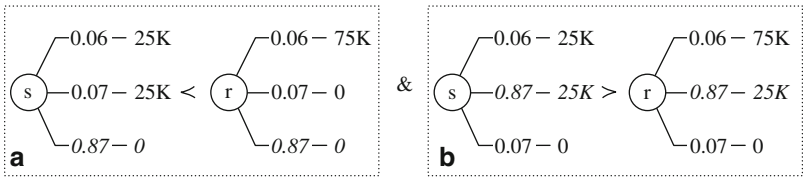
Because risk is a special and simple subcase of uncertainty (as explained later), this chapter on uncertainty begins with a discussion of *decision under risk*, where the probability $p_j = P(E_j)$ is given for each event E_j. We can then write a prospect as $(p_1 : x_1, \ \ldots, p_n : x_n)$, yielding x_j with probability $p_j, j = 1, \ \ldots, n$. Empirical violations of expected-value maximization because of risk aversion (prospects being less preferred than their expected value) led Bernoulli (1738) to propose expected utility, $\sum_{j=1}^{n} p_j U(x_j)$, to evaluate prospects, where U is the utility function. Then risk aversion is equivalent to concavity of U.

Several authors have argued that it is intuitively unsatisfactory that risk attitude be modelled through the utility of money (Lopes 1987, p. 283). It would be more satisfactory if risk attitude were also related to the way people feel about probabilities. Economists often react very negatively to such arguments, based as they are on introspection and having no clear link to revealed preference. Arguments against expected utility that are based on revealed preference were put forward by Allais (1953).

Figure 1 displays preferences commonly found, with K denoting \$1,000:

$$(0.06 : 25\text{K}, 0.07 : 25\text{K}, 0.87 : 0)$$
$$\times \prec (0.06 : 75\text{K}, 0.07 : 0, 0.87 : 0)$$
$$\times \text{and} (0.06 : 25\text{K}, 0.87 : 25K, 0.07 : 25\text{K})$$
$$\times \succ (0.06 : 75\text{K}, 0.87 : 25K, 0.07 : 0).$$

Preference symbols $\succcurlyeq$, $\succ$, $\prec$, $\preccurlyeq$ and are as usual. We denote the outcomes in a rank-ordered manner, from best to worst. In Fig. 1a, people usually prefer the 'risky' (r) prospect because the high payment of 75 K is attractive. In Fig. 1b, people usually prefer the 'safe' (s) certainty of 25 K for sure. These preferences violate expected utility because, after dropping the common (italicized) term 0.87 U (0) from the



Uncertainty, Fig. 1 A version of the Allais paradox for risk

Circles indicate 'chance nodes'.

expected-utility inequality for Fig. 1a and dropping the common term 0.87 U (25 K) from the expected-utility inequality for Fig. 1b, the two inequalities become the same. Hence, under expected utility either both preferences should be for the safe prospect or both preferences should be for the risky one, and they cannot switch as in Fig. 1. The special preference for safety in the second choice (the *certainty effect*) cannot be captured in terms of utility. Hence, alternative, non-expected utility models have been developed.

Based on the valuable intuition that risk attitude should have something to do with how people feel about probabilities, Quiggin (1982) introduced rank-dependent utility theory for risk. The same theory was discovered independently for the broader and more subtle context of uncertainty by Schmeidler (1989, first version 1982), a contribution that will be discussed later. A *probability weighting function* w: [0, 1] → [0, 1] satisfies $w(0) = 0$, $w(1) = 1$, and is strictly increasing and continuous. It reflects the (in)sensitivity of people towards probability. Assume that the outcomes of a prospect $(p_1 : x_1, \ldots, p_n : x_n)$ are rank-ordered, $x_1 \geq \ldots \geq x_n$. Then its *rank-dependent utility* (*RDU*) is $\sum_{j=1}^{n} \pi_j U(x_j)$, where *utility* U is as before, and π_j, the *decision weight* of outcome x_j, is $w(p_1 + \cdots + p_j) - w(p_1 + \cdots + p_{j-1})$ (which is $w(p_1)$ for $j = 1$).

Tversky and Kahneman (1992) adapted their widely used original prospect theory (Kahneman and Tversky 1979) by incorporating the rank dependence of Quiggin and Schmeidler. Prospect theory generalizes rank dependence by allowing a different treatment of gains from that of losses, which is desirable for empirical purposes. In this article on uncertainty, I focus on gains, in which case prospect theory in its modern version, sometimes called cumulative prospect theory, coincides with RDU.

With rank dependence, we can capture psychological (mis)perceptions of unfavourable outcomes being more likely to arise, in agreement with Lopes's (1987) intuition. We can also capture decision attitudes of deliberately paying more attention to bad outcomes. An extreme example of the latter pessimism concerns worst-case analysis, where all weight is given to the most unfavourable outcome. Rank dependence can explain the Allais paradox because the weight of the 0.07 branch in Fig. 1b may exceed that in Fig. 1a:

$$w(1) - w(0.93) \geq w(0.13) - w(0.06). \quad (1)$$

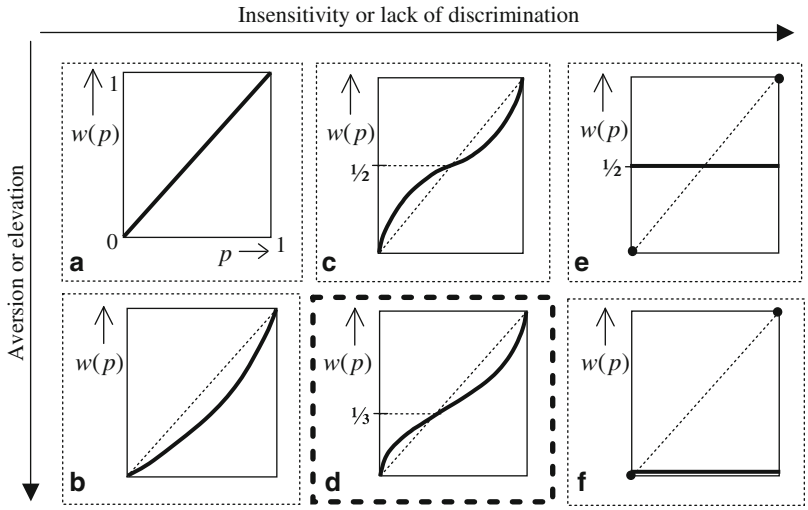
This inequality holds for w-functions that are steeper near 1 than in the middle region, a shape that is empirically prevailing indeed.

The following figures depict some probability weighting functions. Figure 2a concerns expected utility, and Fig. 2b a *convex w*, which means that

$$w(p + r) - w(r) \quad (2)$$

is increasing in r for all $p \geq 0$. This is equivalent to w' being increasing, or w'' being positive. Equation 1 illustrates this property. Equation 2 gives the decision weight of an outcome occurring with probability p if there is an r probability of better outcomes. Under convexity, outcomes receive more weight as they are ranked worse (that is, r is larger), reflecting *pessimism*. It implies low evaluations of prospects relative to sure outcomes, enhancing risk aversion.

Empirical studies have found that usually $w(p) > p$ for small p, contrary to what convexity would imply, and that $w(p) < p$ only for moderate and high probabilities p (inverse-S; Abdellaoui 2000; Bleichrodt and Pinto 2000; Gonzalez and Wu 1999; Tversky and Kahneman 1992), as in Fig. 2c, d. It leads to extremity-oriented behaviour with both best and worst outcomes overweighted. The curves in Fig. 2c, d also satisfy Eq. 1 and also accommodate the Allais paradox. They predict risk seeking for prospects that with a small probability generate a high gain, such as in public lotteries. The inverse-S shape suggests a cognitive insensitivity to probability, generating insufficient response to intermediate variations of probability and then, as a consequence, overreactions to changes from impossible to possible and from possible to certain. These phenomena arise prior to any 'motivational' (value-based) preference or dispreference for risk. Extreme cases of such behaviour are in Fig. 2e, f (where we relaxed the continuity requirement for w).



Uncertainty, Fig. 2 (**a**) Expected utility: linearity; (**b**) Aversion to risk: convexity; (**c**) Insensitivity: inverse-S (**d**) Prevailing empirical finding; (**e**) Extreme insensitivity; (**f**) Extreme aversion and insensitivity

Starmer (2000) surveyed non-expected utility for risk. The main alternatives to the rank-dependent models are the betweenness models (Chew 1983; Dekel 1986), with Gul's (1991) disappointment aversion theory as an appealing special case. Betweenness models are less popular today than the rank-dependent models. An important reason, besides their worse empirical performance (Starmer 2000), is that models alternative to the rank-dependent ones did not provide concepts as intuitive as the sensitivity to probability/information modelled through the probability weighting w of the rank-dependent models. For example, consider a popular special case of betweenness, called weighted utility. The value of a prospect is

$$\frac{\sum_{i=1}^{n} p_i f(x_i) U(x_i)}{\sum_{j=1}^{n} p_j f(x_j)} \tag{3}$$

for a function $f : R \to R^+$. This new parameter f can, similar to rank dependence, capture pessimistic attitudes of overweighting bad outcomes by assigning high values to bad outcomes. It, however, applies to outcomes and not to probabilities. Therefore, it captures less extra variance of the data in the presence of utility than w, because utility also applies to outcomes. For example, for fixed outcomes, Eq. 3 cannot capture the varying sensitivity to small, intermediate and high probabilities found empirically. Both pessimism and marginal utility are entirely specified by the range of outcomes considered without regard to the probabilities involved. It seems more interesting if new concepts, besides marginal utility, concern the probabilities and the state of information of the decision maker rather than outcomes and their valuation. This may explain the success of rank-dependent theories and prospect theory.

Phenomena Under Uncertainty that Naturally Extend Phenomena Under Risk

The first approach to deal with uncertainty was the Bayesian approach, based on de Finetti (1931), Ramsey (1931), and Savage (1954). It assumes that people assign, as well as possible, subjective probabilities $P(E_j)$ to uncertain events E_j. They then evaluate prospects $(E_1 : x_1, \ldots, E_n : x_n)$ through their (subjective) expected utility $\sum_{j=1}^{n} P(E_j) U(x_j)$. This model was the basis of Bayesian statistics and of much of the economics of uncertainty (Greenspan 2004). The empirical measurement of subjective probabilities has been studied extensively (Fishburn 1986; Manski 2004; McClelland and Bolger 1994). We confine our attention in what follows to models that have been introduced since the mid-1980s, models that deviate from

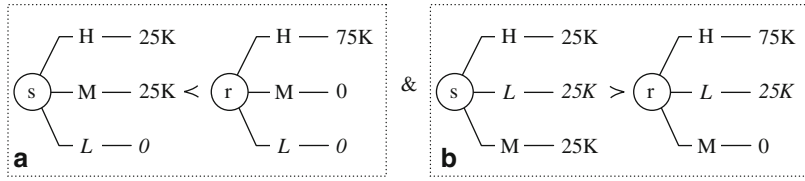
Bayesianism. To Bayesians (including this author) such models are of interest for descriptive purposes.

Machina and Schmeidler (1992) characterized *probabilistic sophistication,* where a decision maker assigns subjective probabilities to events with unknown probabilities and then proceeds as for known probabilities. The decision maker may, however, deviate from expected utility for known probabilities, contrary to the Bayesian approach, and Allais-type behaviour can be accommodated.

The difference between objective, exogenous probabilities and subjective, endogenous probabilities is important. The former are stable, and readily available for analyses, empirical tests and communication in group decisions. The latter can be volatile and can change at any time by mere further thinking by the agent. For descriptive studies, subjective probabilities may become observable only after complex measurement procedures. Hence, I prefer not to lump objective and subjective probabilities together into one category, as has been done in several economic works (Ellsberg 1961, p. 645; Epstein 1999). In this article, the term risk refers exclusively to exogenous objective probabilities. Such probabilities can be considered a limiting case of subjective probabilities, in the same way as decision under risk can be considered a limiting case of decision under uncertainty (Greenspan 2004, pp. 36–7). Under a differentiability assumption for state spaces, Machina (2004) formalized this inpt. Risk, while not occurring very frequently, is especially suited for applications of decision theory.

The Allais paradox is as relevant to uncertainty as it is to risk (MacCrimmon and Larsson 1979, pp. 364–5; Wu and Gonzalez 1999). Figure 3 presents a demonstration by Tversky and Kahneman (1992, section 1.3). The analogy with Fig. 1 should be apparent. The authors conducted the following within-subjects experiment. Let d denote the difference between the closing value of the Dow Jones on the day of the experiment and on the day after, where we consider the events H(igh): $d > 35$, M(iddle): $35 \geq d \geq 30$, L(ow): $30 > d$. The total Dow Jones value at the time of the experiment was about 3,000. The right prospect in Fig. 3b is (H:75K, L:25K, M:0), and the other prospects are denoted similarly. Of 156 money managers during a workshop, 77% preferred the risky prospect r in Fig. 3a, but 68% preferred the safe prospect s in Fig. 3b. The majority preferences violate expected utility, just as they do under risk: after dropping the common terms $P(\mathrm{L})U(0)$ *and* $P(\mathrm{L})U(25\mathrm{K})$ (P denotes subjective probabilities), the same expected-utility inequality results for Fig. 3a as for Fig. 3b. Hence, either both preferences should be for the safe prospect, or both preferences should be for the risky one, and they cannot switch as in Fig. 3. This reasoning holds irrespective of what the subjective probabilities $P(\mathrm{H})$, $P(\mathrm{M})$ and $P(\mathrm{L})$ are.(The condition of expected utility that is falsified here, Savage's (1954) 'sure-thing principle', can be related to the separability preference condition of consumer theory.)

Schmeidler's (1989) *rank-dependent utility* (*RDU*) can accommodate the Allais paradox for uncertainty. We consider a *weighting function* (or non-additive probability or capacity) W that assigns value 0 to the vacuous (empty) event $\varnothing$, value 1 to the universal event, and satisfies monotonicity with respect to set-inclusion (if $A \supset B$ then $W(A) \geq W(B)$). Probabilities are special cases of weighting functions that satisfy *additivity*: $W(A \cup B) = W(A) + W(B)$ for disjoint events A and B. General weighting functions need not satisfy additivity. Assume that the outcomes of a prospect $(E_1 : x_1, \ \ldots \ , E_n : x_n)$ are rank-ordered, $x_1 \geq \ \ldots \ \geq x_n$. Then the prospect's *rank-dependent utility* (*RDU*) is $\sum_{j=1}^{n} \pi_j U(x_j)$ where *utility* U is as before, and π_j, the *decision weight* of outcome x_j, is $W(E_1 \cup \ \ldots \ \cup E_j) - W(E_1 \cup \ \ldots \ \cup E_{j-1})(\pi_1 = W(E_1))$. The decision weight of x_j



Uncertainty, Fig. 3 The certainty effect (Allais paradox) for uncertainty

is the marginal W contribution of E_j to the event of receiving a better outcome.

Quiggin's RDU for risk is the special case with probabilities $p_j = P(E_j)$ given for all events, and $W(E_j) = w(P(E_j))$ with w the probability weighting function. Tversky and Kahneman (1992) improved their 1979 prospect theory not only by avoiding violations of stochastic dominance, but also, and more importantly, by extending their theory from risk to uncertainty, by incorporating Schmeidler's RDU.

Figure 3 can, just as in the case of risk, be explained by a larger decision weight for the M branches in Fig. 3b than in Fig. 3a:

$$W(\mathrm{M} \cup \mathrm{H} \cup \mathrm{L}) - W(\mathrm{H} \cup \mathrm{L}) \geq W(\mathrm{M} \cup \mathrm{H}) - W(\mathrm{H}). \quad (4)$$

This inequality occurs for W-functions that are more sensitive to changes of events near the certain universal event M ∪ H ∪ L than for events of moderate likelihood such as M ∪ H. Although for uncertainty we cannot easily draw graphs of W functions, their properties are natural analogs of those depicted in Fig. 2a–f. W is *convex* if the marginal W contribution of an event E to a disjoint event R is increasing in R, that is,

$$W(E \cup R) - W(R) \quad (5)$$

is increasing in R (with respect to set inclusion) for all E. This agrees with Eq. 4, where increasing R from H to H ∪ L leads to a larger decision weight for $E = M$. Our definition of convexity is equivalent to other definitions in the literature such as $W(A \cup B) + W(A \cap B) \geq W(A) + W(B)$. (Take $E = A - B$, and compare $R = A \cap B$ with the larger $R = A$.)

For probabilistic sophistication ($W(\cdot) = w(P(\cdot))$), convexity of W is equivalent to convexity of w under usual richness conditions, illustrating once more the close similarity between risk and uncertainty. Equation 5 gives the decision weight of an outcome occurring under event E if better outcomes occur under event R. Under convexity, outcomes receive more weight as they are ranked worse (that is, R is larger), reflecting *pessimism*. Theoretical economic studies usually assume convex W's, implying low evaluations of prospects relative to sure outcomes.

Empirical studies have suggested that weighting functions W for uncertainty exhibit patterns similar to Fig. 2d, with unlikely events overweighted rather than, as convexity would have it, underweighted (Einhorn and Hogarth 1986; Tversky and Fox 1995; Wu and Gonzalez 1999). As under risk, we get extremity orientedness, with best and worst outcomes overweighted and lack of sensitivity towards intermediate outcomes (Chateauneuf et al. 2005; Tversky and Wakker 1995).

Phenomena for Uncertainty that Do Not Show Up for Risk: The Ellsberg Paradox

Empirical studies have suggested that phenomena found for risk hold for uncertainty as well, and do so to a more pronounced degree (Fellner 1961, p. 684; Kahn and Sarin 1988, p. 270; Kahneman and Tversky 1979, p. 281), in particular regarding the empirically prevailing inverse-S shape and its extension to uncertainty (Abdellaoui et al. 2005; Hogarth and Kunreuther 1989; Kilka and Weber 2001; Weber 1994, pp. 237–8). It is plausible, for example that the absence of known probabilities adds to the inability of people to sufficiently distinguish between various degrees of likelihood not very close to impossibility and certainty. In such cases, inverse-S shapes will be more pronounced for uncertainty than for risk. This observation entails a within-person comparison of attitudes for different sources of uncertainty, and such comparisons are the main topic of this section.

For Ellsberg's paradox, imagine an urn K with a known composition of 50 red balls and 50 black balls, and an ambiguous urn A with 100 red and black balls in unknown proportion. A ball is drawn at random from each urn, with R_k denoting the event of a red ball from the known urn, and the events B_k, R_a and B_a defined similarly. People prefer to bet on the known urn rather than on the ambiguous urn, and common preferences are:

$$(B_k : 100, R_k : 0) \succ (B_a : 100, R_a : 0) \times \text{and } (B_k : 0, R_k : 100) \succ (B_a : 0, R_a : 100).$$

Such preferences are also found if people can themselves choose the colour to bet on so that there is no reason for suspecting an unfavourable composition of the unknown urn. Under probabilistic sophistication with probability measure P, the two preferences would imply $P(B_k) > P(B_a)$ and $P(R_k) > P(R_a)$. However, $P(B_k) + P(R_k) = 1 = P(B_a) + P(R_a)$ yields a contradiction, because two big numbers cannot give the same sum as two small numbers. Ellsberg's paradox consequently violates robabilistic sophistication and, a fortiori, expected utility. Keynes (1921, p. 75) discussed the difference between the above two urns before Ellsberg did, but did not put forward the choice paradox and deviation from probabilistic sophistication as Ellsberg did. We now analyse the example assuming RDU.

In many studies of uncertainty, such as Schmeidler (1989), expected utility is assumed for risk, primarily for the sake of simplicity. Then, $W(B_k) = W(R_k) = 0.5$ in the above example, with these W values reflecting objective probabilities. Under RDU, the above preferences imply $W(B_a) = W(R_a) < 0.5$, in agreement with convex (or eventwise dominance, or inverse-S; for simplicity of presentation, I focus on convexity hereafter) weighting functions W. This finding led to the widespread misunderstanding that it is primarily the Ellsberg paradox that implies convex weighting functions for unknown probabilities, a condition that was sometimes called 'ambiguity aversion'. I have argued above that it is the Allais paradox, and not the Ellsberg paradox, that implies these conclusions, and I propose another interpretation of the Ellsberg paradox hereafter, following works by Amos Tversky in the early 1990s.

First, it is more realistic not to commit to expected utility under risk when studying uncertainty. Assume, therefore, that

$W(B_k) = W(R_k) = w(P(B_k)) = w(P(R_k)) = w(0.5)$ for a nonlinear probability weighting function. It follows from the Ellsberg paradox that

$W(B_a) = W(R_a) < w(0.5)$. This suggests:

Hypothesis. *In the Ellsberg paradox, the weighting function is more convex for the unknown urn than for the known urn.*

Thus, the Ellsberg paradox itself does not speak to convexity in an absolute sense, and does not claim convexity for known or for unknown probabilities. It speaks to convexity in a relative (within-person) sense, suggesting *more* convexity for unknown probabilities than for known probabilities. It is, for instance, possible that the weighting function is concave, and not convex, for both known and unknown probabilities, but is less concave (and thus more convex) for the unknown probabilities (Wakker 2001, section 6; cf. Epstein 1999, pp. 589–90, or Ghirardato and Marinacci 2002, example 25).

With information only about observed behaviour, and without additional information about the compositions of the urns or the agent's knowledge thereof, we cannot conclude which of the urns is ambiguous and which is not. It would then be conceivable that urn K were ambiguous and urn A were unambiguous, and that the agent satisfied expected utility for A and was optimistic or ambiguity seeking (concave weighting function, Eq. 5 decreasing in R) for K, in full agreement with the Ellsberg preferences. Which of the urns is ambiguous and which is not is based on extraneous information, being our knowledge about the composition of the urns and about the agent's knowledge thereof. This point suggests that no endogenous definition of (un)ambiguity is possible.

The Ellsberg paradox entails a comparison of attitudes of one agent with respect to different sources of uncertainty. It constitutes a *within-agent* comparison. Whereas the Allais paradox concerns violations of expected utility in an absolute sense, the Ellsberg paradox concerns a relative aspect of such violations, finding more convexity (or eventwise dominance, or inverse-S) for the unknown urn than for the known urn. Such a phenomenon cannot show up if we study only risk, because risk is essentially only one source of uncertainty. Apart from some volatile psychological effects (Kirkpatrick and Epstein 1992; Piaget and Inhelder 1975), it seems plausible that people do not distinguish between different ways of generating objective known probabilities.

Uncertain events of particular kinds can be grouped together into sources of uncertainty. Formally, let *sources* be particular algebras of events,

which means that sources are closed under complementation and union, and contain the vacuous and universal events. For example, source $\mathscr{A}$ may concern the performance of the Dow Jones stock index tomorrow, and source $\mathscr{B}$ the performance of the Nikkei stock index tomorrow. Assume that $\mathscr{A}$ from source designates the event that the Dow Jones index goes up tomorrow, and B from source $\mathscr{B}$ the event that the Nikkei index goes up tomorrow. If we prefer (A:100, not-A,0) to (B:100, not-B:0), then this may be caused by a special source preference for $\mathscr{A}$ over $\mathscr{B}$, say, if $\mathscr{A}$ comprises less ambiguity for us than $\mathscr{B}$. However, it may also occur simply because we think that event A is more likely to occur than event B. To examine ambiguity attitudes we have to find a way to 'correct' for differences in perceived levels of likelihood.

One way to detect (*strong*) *source preference* for $\mathscr{A}$ over $\mathscr{B}$ is to find an $\mathscr{A}$ -partition $(A_1, \ldots, A_n)$ and a $\mathscr{B}$ -partition $(B_1, \ldots, B_n)$ of the universal event such that for each j, $(A_j : 100, not - A_j, 0) \succ (B_j : 100, not - B\neg_j : 0)$ (Nehring 2001, definition 4; Tversky and Fox 1995; Tversky and Wakker 1995). Because both partitions span the whole universal event, we cannot have stronger belief in every A_j than B_j (under some plausible assumptions about beliefs), and hence there must be a preference for dealing with $\mathscr{A}$ events beyond belief. Formally, the condition requires that a similar preference of $\mathscr{B}$ over $\mathscr{A}$ is never detected. The Ellsberg paradox is a special case of this procedure.

Under the above approach to source preference, there is a special role for probabilistic sophistication. For a source $\mathscr{A}$ for which not some of its events are more ambiguous than others, it is plausible that $\mathscr{A}$ exhibits source indifference with respect to itself. This condition can be seen to amount to the additivity axiom of qualitative probability (if A_1 is as likely as A_3, and A_2 is as likely as A_4, then $A_1 \cup A_2$ is as likely as $A_3 \cup A_4$ whenever $A_1 \cap A_2 = A_3 \cap A_4 = \emptyset$), which, under sufficient richness, implies probabilistic sophistication for $\mathscr{A}$ under RDU, and does so in general (without RDU assumed) under an extra dominance condition (Fishburn 1986; Sarin and Wakker 2000). The condition also comprises source sensitivity (Tversky and Wakker 1995). Probabilistic sophistication, then, entails a *uniform degree* of ambiguity of a source.

In theoretical economic studies it has usually been assumed that people are averse to ambiguity, corresponding with convex weighting functions. Empirical studies, mostly by psychologists, have suggested a more varied pattern, where different sources of ambiguity can arouse all kinds of emotions. For example, Tversky and Fox (1995) found that basketball fans exhibit source preference for ambiguous uncertain events related to basketball over events with known probabilities, which entails ambiguity seeking. This finding is not surprising in an empirical sense, but its conceptual implication is important: attitudes towards ambiguity depend on many ad hoc emotional aspects, such as a general aversion to deliberate secrecy about compositions of urns, or a general liking of basketball. Uncertainty is a large domain, and fewer regularities can be expected to hold universally for uncertainty than for risk, in the same way as fewer regularities will hold universally for the utility of non-monetary outcomes (hours of listening to music, amounts of milk to be drunk, life duration, and so on) than for the utility of monetary outcomes. It means that there is much yet to be discovered about uncertainty.

Models for Uncertainty Other Than Rank-Dependence

Multiple Priors

An interesting model of ambiguity by Jaffray (1989), with a separation of ambiguity beliefs and ambiguity attitudes, unfortunately has received little attention as yet. A surprising case of unknown probabilities can arise when the expected utility model perfectly well describes behaviour, but utility is state-dependent. The (im) possibility of defining probability in such cases has been widely discussed (Drèze 1987; Grant and Karni 2005; Nau 2006).

The most popular alternative to Schmeidler's RDU is the multiple priors model introduced by Wald (1950). It assumes a set $\mathscr{P}$ of probability measures plus a utility function U, and evaluates

each prospect through its minimal expected utility with respect to the probability distributions contained in $\mathscr{P}$. The model has an overlap with RDU: if W is convex, then RDU is the minimal expected utility over $\mathscr{P}$ where $\mathscr{P}$ is the CORE of W, that is, the set of probability measures that eventwise dominate W. Drèze (1961, 1987) independently developed a remarkable analog of the multiple priors model, where the maximal expected utility is taken over $\mathscr{P}$, and $\mathscr{P}$ reflects moral hazard instead of ambiguity. Drèze also provided a preference foundation. Similar functionals appear in studies of robustness against model misspecification in macroeconomics (Hansen and Sargent 2001).

Variations of multiple priors, combining pessimism and optimism, employ convex combinations of the expected utility minimized over $\mathscr{P}$ and the expected utility maximized over $\mathscr{P}$ (Ghirardato et al. 2004, proposition 19). Such models can account for extremity orientedness, as with inverse-S weighting functions and RDU. Arrow and Hurwicz (1972) proposed a similar model where a prospect is evaluated through a convex combination of the minimal and maximal utility of its outcomes (corresponding with $\mathscr{P}$ being the set of all probability measures). This includes maximin and maximax as special cases. Their approach entails a level of ambiguity so extreme that no levels of belief other than 'sure-to-happen', 'sure-not-to-happen' and 'don't know' play a role, similar to Fig. 2e, f, and suggesting a three-valued logic. Other non-belief-based approaches, including minimax regret, are in Manski (2000) and Savage (1954), with a survey in Barberà et al. (2004).

Other authors proposed models where for each single event a separate interval of probability values is specified (Budescu and Wallsten 1987; Kyburg 1983; Manski 2004). Such interval-probability models are mathematically different from multiple priors because there is no unique relation between sets of probability measures over the whole event space and intervals of probabilities separately for each event. The latter models are more tractable than multiple priors because probability intervals for some relevant event are easier to specify than probability measures over the whole space, but these models did not receive a preference foundation and never became popular in economics. Similar models of imprecise probabilities received attention in the statistics field (Walley 1991).

Wald's multiple priors model did receive a preference axiomatization (Gilboa and Schmeidler 1989), and consequently became the most popular alternative to RDU for unknown probabilities. The evaluating formula is easier to understand at first than RDU. The flexibility of not having to specify precisely what 'the' probability measure is, while usually perceived as an advantage at first acquaintance, can turn into a disadvantage when applying the model. We then have to specify exactly what 'the' set of probability distributions is, which is more complex than exactly specifying only one probability measure (cf. Lindley 1996).

The simple distinction between probability measures that are either possible (contained in $\mathscr{P}$) or impossible (not contained in $\mathscr{P}$), on the one hand adds to the tractability of the model, but on the other hand cannot capture cognitive states where different probability measures are plausible to different degrees. To the best of my knowledge, the multiple priors model cannot yet be used in quantitative empirical measurements today, and there are no empirical assessments of sets of priors available in the literature to date. Multiple priors are, however, well suited for general theoretical analyses where only general properties of the model are needed. Such analyses are considered in many theoretical economic studies, where the multiple priors model is very useful.

The multiple priors model does not allow deviations from expected utility under risk, and a desirable extension would obviously be to combine the model with non-expected utility for risk. Promising directions for resolving the difficulties of the multiple priors model are being explored today (Maccheroni et al. 2005).

Model-Free Approaches to Ambiguity

Dekel et al. (2001) considered models where outcomes of prospects are observed but the state space has not been completely specified, as relevant to incomplete contracts. Similar approaches with ambiguity about the underlying states and

events appeared in psychology in repeated-choice experiments by Hertwig et al. (2003), and in support theory (Tversky and Koehler 1994). This section discusses two advanced attempts to define ambiguity in a model-free way that have received much attention in the economic literature.

In a deep paper, Epstein (1999) initiated one such approach, continued in Epstein and Zhang (2001). Epstein sought to avoid any use of known probabilities and tried to endogenize (un)ambiguity and the use of probabilities. (He often used the term uncertainty as equivalent to ambiguity.) For example, he did not define risk neutrality with respect to known probabilities, as we did above, but with respect to subjective probabilities derived from preferences as in probabilistic sophistication (Epstein 1999, Eq. 2). He qualified probabilistic sophistication as ambiguity neutrality (not uniformity as done above). Ghirardato and Marinacci (2002) used another approach that is similar to Epstein's. They identified absence of ambiguity not with probabilistic sophistication, as did Epstein, but, more restrictively, with expected utility.

The above authors defined an agent as ambiguity averse if there exists another, hypothetical, agent who behaves the same way for unambiguous events, but who is ambiguity neutral for ambiguous events, and such that the real agent has a stronger preference than the hypothetical agent for sure outcomes (or unambiguous prospects, but these can be replaced by their certainty equivalents) over ambiguous prospects. This definition concerns traditional between-agent within-source comparisons as in Yaari (1969). The stronger preferences for certainty are, under rank-dependent models, equivalent to eventwise dominance of weighting functions, leading to non-emptiness of the CORE (Epstein 1999, lemma 3.4; Ghirardato and Marinacci 2002, corollary 13). These definitions of ambiguity aversion are not very tractable because of the 'there exists' clause. It is difficult to establish which ambiguity neutral agent to take for the comparisons. To mitigate this problem, Epstein 1999, section 4) proposed eventwise derivatives as models of local probabilistic sophistication. Such derivatives exist only for continua of events with a linear structure, and are difficult to elicit. They serve their purpose only under restrictive circumstances (ambiguity aversion throughout plus constancy of the local derivative, called coherence; see Epstein's Theorem 4.3).

In both above approaches, ambiguity and ambiguity aversion are inextricably linked, making it hard to model attitudes towards ambiguity other than aversion or seeking (such other attitudes include insensitivity), or to distinguish between ambiguity-neutrality or -absence (Epstein 1999, p. 584, 1st para; Epstein and Zhang 2001, p. 283; Ghirardato and Marinacci 2002, p. 256, 2nd para). Both approaches have difficulties distinguishing between the two Ellsberg urns. Each urn in isolation can be taken as probabilistically sophisticated with, in our inpt, a uniform degree of ambiguity, and Epstein's definition cannot distinguish which of these is ambiguity neutral (cf. Ghirardato and Marinacci 2002, middle of p. 281). Ghirardato and Marinacci's definition does so, but only because it selects expected utility (and the urn generating such preferences) as the only ambiguity-neutral version of probabilistic sophistication. Any other form of probabilistic sophistication, that is, any non-expected utility behaviour under risk, is then either mismodelled as ambiguity attitude (Ghirardato and Marinacci 2002, pp. 256–7), or must be assumed not to exist.

We next discuss in more detail a definition of (un) ambiguity by Epstein and Zhang (2001), whose aim was to make (un)ambiguity endogenously observable by expressing it directly in terms of a preference condition. They called an event E *unambiguous* if

$$\begin{aligned}&(E:c,E_2:\gamma,E_3:\beta,E_4:x_4,\ \ldots,E_n:x_n)\\ &\times\succcurlyeq(E:c,E_2:\beta,E_3:\gamma;E_4:x_4,\ \ldots,E_n:x_n)\\ &\times\text{implies}\ (E:c',E_2:\gamma,E_3:\beta;E_4:x_4,\ \ldots,E_n:x_n)\\ &\times\succcurlyeq(E:c',E_2:\beta;E_3:\gamma,E_4:x_4,\ \ldots,E_n:x_n)\end{aligned}\tag{6}$$

for all partitions E_2 , … , E_n of not-E, and all outcomes c, c' , x_4 , … , x_n . . . , γ · $\succcurlyeq$ β, with a similar condition imposed on not-E. In words, changing a common outcome c into another common outcome c' under E does not affect preference, but this is imposed only if the preference concerns nothing other than to which event (E_2 or

E_3) a good outcome γ is to be allocated instead of a worse outcome β. Together with some other axioms, Eq. 6 implies that probabilistic sophistication holds on the set of events satisfying this condition, which in the interpretation of the authors designates absence of ambiguity (rather than uniformity). As we will see next, it is not clear why Eq. 6 would capture the absence of ambiguity.

Example Assume that events are subsets of [0,1), $E = [0, 0.5)$, not $-E = [0.5, 1)$, and E has unknown probability π. Every subset A of E has probability $2\pi\lambda(A)$ (λ is the usual Lebesgue measure, that is, the uniform distribution over [0,1)) and every subset B of not-E has probability $2(1 - \pi)\lambda(B)$. Then it seems plausible that event E and its complement not-E are ambiguous, but conditional on these events ('within them') we have probabilistic sophistication with respect to the conditional Lebesgue measure and without any ambiguity. In Schmeidler (1989), the ambiguous events E and not-E are called horse events, and the unambiguous events conditional on them are called roulette events. Yet, according to Eq. 6, events E and not-E themselves are unambiguous, both preferences in Eq. 6 being determined by whether λ (E_2) exceeds λ (E_3).

In the example, the definition in Eq. 6 erroneously ascribes the unambiguity that holds for events conditional on E, so 'within E', to E as a whole. Similar examples can be devised where E and not-E themselves are unambiguous, there is 'non-uniform' ambiguity conditional on E, this ambiguity is influenced by outcomes conditional on not-E through non-separable interactions typical of non-expected utility, and Eq. 6 erroneously ascribes the ambiguity that holds within E to E as a whole.

A further difficulty with Eq. 6 is that it is not violated in the Ellsberg example with urns A and K as above (nor if the uncertainty regarding each urn is extended to a 'uniform' continuum as in Example 5.8ii of Abdellaoui and Wakker 2005), and cannot detect which of the urns is ambiguous. The probabilistic sophistication that is obtained in Epstein and Zhang (2001, Theorem 5.2) for events satisfying Eq. 6, and that rules out the two-urn Ellsberg paradox and its continuous extension of Abdellaoui and Wakker (2005), is mostly driven by their Axioms 4 and 6 (the latter is not satisfied by all rank-dependent utility maximizers contrary to the authors' claim at the end of their section 4; their footnote 18 is incorrect) and the necessity to consider also intersections of different-urn events (see their Appendix E). This imposes, in my terminology, a uniformity of ambiguity over the events satisfying Eq. 6 that, rather than Eq. 6 itself, rules out the above counterexs.

Multi-stage Approaches to Ambiguity

Several authors have considered two-stage approaches with intersections of first-stage events A_i, $i = 1,\cdots,\ell$ and second-stage events K_j, $i = 1,\cdots,k$, so that $n = \ell k$ events A_iK_j result, and prospects $\left(A_iK_j : x_{ij}\right)_{i=1j=1}^{\ell k}$ are considered. It can be imagined that in a first stage it is determined which event A_i obtains, and then in a second stage, conditional on A_i, which event K_j obtains. Many authors considered such two-stage models with probabilities given for the events in both stages, the probabilities of the first stage interpreted as ambiguity about the probabilities of the second stage, and non-Bayesian evaluations used (Levi 1980; Segal 1990; Yates and Zukowski 1976).

Other authors considered representations

$$\sum_{i=1}^{\ell} Q(A_i)\phi\left(\sum_{j=1}^{k} P\left(K_j\right)U\left(x_{ij}\right)\right) \qquad (7)$$

for probability measures P and Q, a utility function U, and an increasing transformation ϕ. For ϕ the identity or, equivalently, ϕ linear, traditional expected utility with backwards induction results. Nonlinear ϕ's give new models. Kreps and Porteus (1979) considered Eq. 7 for intertemporal choice, interpreting nonlinear ϕ's as non-neutrality towards the timing of the resolution of uncertainty. Ergin and Gul (2004) and Nau (2006) reinterpreted the formula, where now the second-stage events are from a source of different ambiguity than the first-stage events. A concave ϕ, for instance, suggests stronger preference for certainty, and more ambiguity aversion, for the first-stage uncertainty than for the second.

Klibanoff et al. (2005) considered cases where the decomposition into A- and K-events is

endogenous rather than exogenous. This approach greatly enlarges the scope of application, but their second-order acts, that is, prospects with outcomes contingent on aspects of preferences, are hard to implement or observe if those aspects cannot be related to exogenous observables.

Equation 7 has a drawback similar to Eq. 3. All extra mileage is to come from the outcomes, to which also utility applies, so that there will not be a great improvement in descriptive performance or new concepts to be developed.

Conclusion

The Allais paradox reveals violations of expected utility in an absolute sense, leading to convex or inverse-S weighting functions for risk and, more generally, for uncertainty. The Ellsberg paradox reveals deviations from expected utility in a relative sense, showing that an agent can deviate more from expected utility for one source of uncertainty (say one with unknown probabilities) than for another (say, one with known probabilities). It demonstrates the importance of within-subject between-source comparisons.

The most popular models for analysing uncertainty today are based on rank dependence, with multiple priors a popular alternative in theoretical studies. The most frequently studied phenomenon is ambiguity aversion. Uncertainty is, however, a rich empirical domain with a wide variety of phenomena, where ambiguity aversion and ambiguity insensitivity (inverse-S) are prevailing but are not universal patterns. The possibility of relating the properties of weighting functions for uncertainty to cognitive inpts such as insensitivity to likelihood-information makes RDU and prospect theory well suited for links with other fields such as psychology, artificial intelligence (Shafer 1976) and neuroeconomics (Camerer et al. 2004).

See Also

- ▶ Allais Paradox
- ▶ Allais, Maurice (Born 1911)
- ▶ Ambiguity and Ambiguity Aversion
- ▶ Bernoulli, Daniel (1700–1782)
- ▶ Certainty Equivalence
- ▶ De Finetti, Bruno (1906–1985)
- ▶ Decision Theory in Econometrics
- ▶ Expected Utility Hypothesis
- ▶ Extremal Quantiles and Value-at-Risk
- ▶ Kahneman, Daniel (Born 1934)
- ▶ Non-expected Utility Theory
- ▶ Rational Behaviour
- ▶ Rational Expectations
- ▶ Revealed Preference Theory
- ▶ Risk
- ▶ Risk Aversion
- ▶ Savage, Leonard J. (Jimmie) (1917–1971)
- ▶ Savage's Subjective Expected Utility Model
- ▶ Separability
- ▶ Statistical Decision Theory
- ▶ Stochastic Dominance
- ▶ Tversky, Amos (1937–1996)
- ▶ Utility

Han Bleichrodt, Chew Soo Hong, Edi Karni, Jacob Sagi and Stefan Trautmann made useful comments.

Bibliography

Abdellaoui, M. 2000. Parameter-free elicitation of utilities and probability weighting functions. *Management Science* 46: 1497–1512.

Abdellaoui, M., and P.P. Wakker. 2005. The likelihood method for decision under uncertainty. *Theory and Decision* 58: 3–76.

Abdellaoui, M., F. Vossmann, and M. Weber. 2005. Choice-based elicitation and decomposition of decision weights for gains and losses under uncertainty. *Management Science* 51: 1384–1399.

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

Arrow, K., and L. Hurwicz. 1972. An optimality criterion for decision making under ignorance. In *Uncertainty and expectations in economics: Essays in honour of G.L.S. Shackle*, ed. C. Carter and J. Ford. Oxford: Basil Blackwell.

Barberà, S., Bossert, W. Pattanaik, P. 2004. Ranking sets of objects. *Handbook of utility theory, vol.* 2, Extensions, S. Barberà, P. Hammond C. Seidl. Dordrecht: Kluwer Academic Publishers.

Bernoulli, D. 1738. Specimen theoriae novae de mensura sortis. *Commentarii Academiae Scientiarum Imperialis Petropolitanae* 5: 175–192.

Bleichrodt, H., and J. Pinto. 2000. A parameter-free elicitation of the probability weighting function in medical decision analysis. *Management Science* 46: 1485–1496.

Budescu, D., and T. Wallsten. 1987. Subjective estimation of precise and vague uncertainties. In *Judgmental forecasting*, ed. G. Wright and P. Ayton. New York: Wiley.

Camerer, C., and M. Weber. 1992. Recent developments in modelling preferences: Uncertainty and ambiguity. *Journal of Risk and Uncertainty* 5: 325–370.

Camerer, C., G. Loewenstein, and D. Prelec. 2004. Neuroeconomics: Why economics needs brains. *Scandinavian Journal of Economics* 106: 555–579.

Chateauneuf, A. 1991. On the use of capacities in modeling uncertainty aversion and risk aversion. *Journal of Mathematical Economics* 20: 343–369.

Chateauneuf, A., Eichberger, J. and Grant, S. 2005. Choice under uncertainty with the best and worst in mind: NEO-additive capacities. CERMSEM, CEM, University of Paris I.

Chew, S.H. 1983. A generalization of the quasilinear mean with applications to the measurement of income inequality and decision theory resolving the Allais paradox. *Econometrica* 51: 1065–1092.

de Finetti, B. 1931. Sul significato soggettivo della probabilità. *Fundamenta Mathematicae* 17: 298–329.

Dekel, E. 1986. An axiomatic characterization of preferences under uncertainty: Weakening the independence axiom. *Journal of Economic Theory* 40: 304–318.

Dekel, E., B. Lipman, and A. Rustichini. 2001. Representing preferences with a subjective state space. *Econometrica* 69: 891–934.

Drèze, J. 1961. Les fondements logiques de l'utilité cardinale et de la probabilité subjective. La Décision, Colloques Internationaux du CNRS, 73–83.

Drèze, J. 1987. *Essays on economic decision under uncertainty.* Cambridge: Cambridge University Press.

Einhorn, H., and R. Hogarth. 1986. Decision making under ambiguity. *Journal of Business* 59: S225–S250.

Ellsberg, D. 1961. Risk, ambiguity and the savage axioms. *Quarterly Journal of Economics* 75: 643–669.

Epstein, L. 1999. A definition of uncertainty aversion. *Review of Economic Studies* 66: 579–608.

Epstein, L., and J. Zhang. 2001. Subjective probabilities on subjectively unambiguous events. *Econometrica* 69: 265–306.

Ergin, H. and Gul, F. 2004. A subjective theory of compound lotteries. Working paper, MIT.

Fellner, W. 1961. Distortion of subjective probabilities as a reaction to uncertainty. *Quarterly Journal of Economics* 75: 670–689.

Fishburn, P. 1986. The axioms of subjective probability. *Statistical Science* 1: 335–358.

Ghirardato, P., and M. Marinacci. 2002. Ambiguity made precise: A comparative foundation. *Journal of Economic Theory* 102: 251–289.

Ghirardato, P., F. Maccheroni, and M. Marinacci. 2004. Differentiating ambiguity and ambiguity attitude. *Journal of Economic Theory* 118: 133–173.

Gilboa, I., ed. 2004. *Uncertainty in economic theory: Essays in honor of David Schmeidler's 65th birthday.* London: Routledge.

Gilboa, I., and D. Schmeidler. 1989. Maxmin expected utility with a non-unique prior. *Journal of Mathematical Economics* 18: 141–153.

Gonzalez, R., and G. Wu. 1999. On the shape of the probability weighting function. *Cognitive Psychology* 38: 129–166.

Grant, S., and E. Karni. 2005. Why does it matter that beliefs and valuations be correctly represented? *International Economic Review* 46: 917–934.

Greenspan, A. 2004. Innovations and issues in monetary policy: The last fifteen years. *American Economic Review: Papers and Proceedings* 94: 33–40.

Gul, F. 1991. A theory of disappointment aversion. *Econometrica* 59: 667–686.

Hansen, L., and T. Sargent. 2001. Robust control and model uncertainty. *American Economic Review: Papers and Proceedings* 91: 60–66.

Hertwig, R., G. Barron, E. Weber, and I. Erev. 2003. Decisions from experience and the effect of rare events in risky choice. *Psychological Science* 15: 534–539.

Hogarth, R., and H. Kunreuther. 1989. Risk, ambiguity, and insurance. *Journal of Risk and Uncertainty* 2: 5–35.

Jaffray, J.-Y. 1989. Linear utility theory for belief functions. *Operations Research Letters* 8: 107–112.

Kahn, B., and R. Sarin. 1988. Modeling ambiguity in decisions under uncertainty. *Journal of Consumer Research* 15: 265–272.

Kahneman, D., and A. Tversky. 1979. Prospect theory: An analysis of decision under risk. *Econometrica* 47: 263–291.

Keynes, J.M. 1921. *A treatise on probability.* London: Macmillan.

Kilka, M., and M. Weber. 2001. What determines the shape of the probability weighting function under uncertainty. *Management Science* 47: 1712–1726.

Kirkpatrick, L., and S. Epstein. 1992. Cognitive-experiential self-theory and subjective probability: Further evidence for two conceptual systems. *Journal of Personality and Social Psychology* 63: 534–544.

Klibanoff, P., M. Marinacci, and S. Mukerji. 2005. A smooth model of decision making under ambiguity. *Econometrica* 73: 1849–1812.

Knight, F. 1921. *Risk, uncertainty, and profit*. New York: Houghton Mifflin.

Köbberling, V., C. Schwieren, and P.P. Wakker. 2006. *Prospect-theory's diminishing sensitivity versus economics' intrinsic utility of money: How the introduction of the euro can be used to disentangle the two empirically.* Mimeo: Department of Economics, University of Maastricht.

Kreps, D., and E. Porteus. 1979. Dynamic choice theory and dynamic programming. *Econometrica* 47: 91–100.

Kyburg, H. Jr. 1983. *Epistemology and Inference*. Minneapolis: University of Minnesota Press.

Levi, I. 1980. *The enterprise of knowledge*. Cambridge, MA: MIT Press.

Lindley, D. 1996. Discussion of Walley (1996). *Journal of the Royal Statistical Society B* 58: 47–48.

Lopes, L. 1987. Between hope and fear: The psychology of risk. *Advances in Experimental Psychology* 20: 255–295.

Maccheroni, F., Marinacci, M. Rustichini, A. 2005. Ambiguity aversion, malevolent nature, and the variational representation of preferences. Mimeo, Istituto di Metodi Quantitativi, Università Bocconi, Milan.

MacCrimmon, K., and S. Larsson. 1979. Utility theory: Axioms versus 'paradoxes'. In *Expected utility hypotheses and the Allais paradox*, ed. M. Allais and O. Hagen. Dordrecht: Reidel.

Machina, M. 2004. Almost-objective uncertainty. *Economic Theory* 24: 1–54.

Machina, M., and D. Schmeidler. 1992. A more robust definition of subjective probability. *Econometrica* 60: 745–780.

Manski, C. 2000. Identification problems and decisions under ambiguity: Empirical analysis of treatment response and normative analysis of treatment choice. *Journal of Econometrics* 95: 415–442.

Manski, C. 2004. Measuring expectations. *Econometrica* 72: 1329–1376.

McClelland, A., and F. Bolger. 1994. The calibration of subjective probabilities: Theories and models 1980–1994. In *Subjective probability*, ed. G. Wright and P. Ayton. New York: Wiley.

Mukerji, S., and J.-M. Tallon. 2001. Ambiguity aversion and incompleteness of financial markets. *Review of Economic Studies* 68: 883–904.

Nau, R. 2006. Uncertainty aversion with second-order utilities and probabilities. *Management Science* 52: 136–145.

Nehring, K. 2001. *Ambiguity in the context of probabilistic beliefs*. Mimeo: UC Davis.

Piaget, J., and B. Inhelder. 1975. *The origin of the idea of chance in children*. New York: Norton.

Quiggin, J. 1982. A theory of anticipated utility. *Journal of Economic Behavior and Organization* 3: 323–343.

Ramsey, F. 1931. Truth and probability. In *The foundations of mathematics and other logical essays*. London: Routledge and Kegan Paul.

Sarin, R., and P.P. Wakker. 2000. Cumulative dominance and probabilistic sophistication. *Mathematical Social Sciences* 40: 191–196.

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

Schmeidler, D. 1982. Subjective probability without additivity, Working paper. Foerder Institute of Economic Research, Tel Aviv University.

Schmeidler, D. 1989. Subjective probability and expected utility without additivity. *Econometrica* 57: 571–587.

Segal, U. 1990. Two-stage lotteries without the reduction axiom. *Econometrica* 58: 349–377.

Shafer, G. 1976. *A mathematical theory of evidence*. Princeton: Princeton University Press.

Starmer, C. 2000. Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of Economic Literature* 38: 332–382.

Tversky, A., and C. Fox. 1995. Weighing risk and uncertainty. *Psychological Review* 102: 269–283.

Tversky, A., and D. Kahneman. 1992. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty* 5: 297–323.

Tversky, A., and D. Koehler. 1994. Support theory: A nonextensional representation of subjective probability. *Psychological Review* 101: 547–567.

Tversky, A., and P.P. Wakker. 1995. Risk attitudes and decision weights. *Econometrica* 63: 1255–1280.

Wakker, P.P. 2001. Testing and characterizing properties of nonadditive measures through violations of the sure-thing principle. *Econometrica* 69: 1039–1059.

Wald, A. 1950. *Statistical decision functions*. New York: Wiley.

Walley, P. 1991. *Statistical reasoning with imprecise probabilities*. London: Chapman and Hall.

Weber, E. 1994. From subjective probabilities to decision weights: The effects of asymmetric loss functions on the evaluation of uncertain outcomes and events. *Psychological Bulletin* 115: 228–242.

Wu, G., and R. Gonzalez. 1999. Nonlinear decision weights in choice under uncertainty. *Management Science* 45: 74–85.

Yaari, M. 1969. Some remarks on measures of risk aversion and on their uses. *Journal of Economic Theory* 1: 315–329.

Yates, J., and L. Zukowski. 1976. Characterization of ambiguity in decision making. *Behavioral Science* 21: 19–25.

Uncertainty and General Equilibrium

Mukul Majumdar and Roy Radner

Abstract

This article reviews alternative approaches to incorporating uncertainty in Walrasian models. It begins with a sketch of the Arrow–Debreu model of complete markets. An extension of this framework allowing for economic agents to have different information about the environment is followed by a critique. When markets are incomplete and trades take place sequentially, several types of equilibrium concept arise according to the hypotheses we make about the way traders form their expectations. We present conditions for the existence of equilibria for two such equilibrium concepts,

and discuss the possible failure to attain Paretian welfare optima.

Keywords

Arrow–Debreu model; Bounded rationality; Budget constraints; Competitive equilibrium; Conditional probability; Consumption possibility set; Equilibrium; Existence of competitive equilibrium; Expectation formation; Expected utility hypothesis; General equilibrium; Incomplete information; Incomplete markets; Indicative planning; Inside information; Limited liability; Moral hazard; Nonprice information; Optimality of competitive equilibrium; Pareto efficiency; Perfect foresight; Production possibility set; Pseudo-equilibrium; Rational expectations; Rational expectations equilibrium; Risk; Sequential trading; Steady state; Temporary (or momentary) equilibrium; Uncertainty; Walras's law

JEL Classifications

D58

One of the notable intellectual achievements of economic theory during the second half of the 20th century has been the rigorous elaboration of the Walras–Pareto theory of value; that is, the theory of the existence and optimality of competitive equilibrium. Although many economists and mathematicians contributed to this development, the resulting edifice owes so much to the pioneering and influential work of Arrow and Debreu that in this paper we shall refer to it as the 'Arrow–Debreu theory'. (For comprehensive treatments, together with references to previous work, see Debreu 1959; Arrow and Hahn 1971.)

The Arrow–Debreu theory was not originally put forward for the case of uncertainty, but an ingenious device introduced by Arrow (1953), and further elaborated by Debreu (1953), enabled the theory to be reinterpreted to cover the case of uncertainty about the availability of resources and about consumption and production possibilities. (see Debreu 1959, Ch. 7, for a unified treatment of time and uncertainty.)

Subsequent research has extended the Arrow–Debreu theory to take account of (*a*) differences in information available to different economic agents, and the 'production' of information, (*b*) the incompleteness of markets, and (*c*) the sequential nature of markets. The consideration of these complications has stimulated the developments of new concepts of equilibrium, two of which will be elaborated in this article under the headings: (*a*) equilibrium of plans, prices, and price expectations (EPPPE) and (*b*) rational expectations equilibrium (REE). The exploration of these features of real-world markets has also made possible a general-equilibrium analysis of money and securities markets, institutions about which the original Arrow–Debreu theory could provide only limited insights. It has also led to a better understanding of the limits to the ability of the 'invisible hand' in attaining a Pareto optimal allocation of resources.

Review of the Arrow–Debreu Model of a Complete Market for Present and Future Contingent Delivery

In this section, we review the approach of Arrow (1953) and Debreu (1959) to incorporating uncertainty about the environment into a Walrasian model of competitive equilibrium. The basic idea is that commodities are to be distinguished, not only by their physical characteristics and by the location and dates of their availability and/or use, but *also by the environmental event in which they are made available and/or used.* For example, ice cream made available (at a particular location on a particular date) if the weather is hot may be considered to be a different commodity from the same kind of ice cream made available (at the same location and date) if the weather is cold. We are thus led to consider a list of 'commodities' that is greatly expanded by comparison with the corresponding case of certainty about the environment. The standard arguments of the theory of competitive equilibrium, applied to an economy with this expanded list of commodities, then require that we envisage a 'price' for each commodity, the resulting set of price ratios

specifying the market rate of exchange between each pair of commodities.

Just what institutions could, or do, effect such exchanges is a matter of interpretation that is, strictly speaking, outside the model. We shall present one straightforward inpt, and then comment briefly on an alternative inpt.

First, however, it will be useful to give a more precise account of concepts of environment and event that we shall be employing. The description of the 'physical world' is decomposed into three sets of variables: (*a*) decision variables, which are controlled (chosen) by economic agents; (*b*) environmental variables, which are not controlled by any economic agent; and (*c*) all other variables, which are completely determined (Possibly jointly) by decisions and environmental variables. A *state* of the environment is a complete specification (history) of the environmental variables from the beginning to the end of the economic system in qst. An *event* is a set of states; for example, the event 'the weather is hot in New York on 1 July 1970' is the set of all possible histories of the environment in which the temperature in New York during the day of 1 July 1970 reaches a high of at least (say) 75 °F. Granted that we cannot know the future with certainty, at any given date, there will be a family of *elementary observable* (knowable) *events,* which can be represented by a partition of the set of all possible states (histories) into a family of mutually exclusive subsets. It is natural to assume that the partitions corresponding to successive dates are successively finer, which represents the accumulation of information about the environment.

We shall imagine that a 'market' is organized before the beginning of the physical history of the economic system. An elementary contract in this market will consist of the purchase (or sale) of some specified number of units of a specified commodity to be delivered at a specified location and date, if and only if a specified elementary event occurs. Payment for this purchase is to be made now (at the beginning), in 'units of account', at a specified price quoted for that commodity-location-date-event combination. Delivery of the commodity in more than one elementary event is obtained by combining a suitable set of elementary contracts. For example, if delivery of one quart of ice cream (at a specified location and date) in hot weather costs \$1.50 (now) and delivery of one-quart in non-hot weather costs \$1.10, then sure delivery of one quart (that is, whatever the weather) costs \$1.50 + \$1.10 = \$2.60.

There are two groups of economic agents in the economy: *producers* and *consumers.* A producer chooses a production plan, which determines his inpt and/or output of each commodity at each date in each elementary event (we shall henceforth suppress explicit reference to location, it being understood that the location is specified in the term 'commodity'). For a given list of prices, the present value of a production plan is the sum of the values of outputs minus the sum of the values of inputs. Each producer is characterized by a set of production plans that are (given the technological know-how) feasible for him: his production possibility set.

A consumer chooses a consumption plan, which specifies his consumption of each commodity at each date in each elementary event. Each consumer is characterized by: (*a*) a set of consumption plans that are (Physically, psychologically, and so on) feasible for him: his consumption possibility set; (*b*) preferences among the alternative plans that are feasible for him; (*c*) his endowment of physical resources, that is, a specification of the quantity of each commodity, for example, labour, at each date in each event, with which he is exogenously endowed; and (*d*) his shares in each producer, that is, the fraction of the present value of each producer's production plan that will be credited to the consumer's account. (for any one producer, the sum of the consumers' shares is unity.) For given prices and given production plans of all the producers, the present net worth of a consumer is the total value of his resources plus the total value of his shares of the present values of producers' production plans.

An equilibrium of the economy is a list of prices, a set of production plans (one for each producer), and a set of consumption plans (one for each consumer), such that (*a*) each producer's plan has maximum present value in his production possibility set; (*b*) each consumer's plan maximizes his preferences within his consumption

possibility set, subject to the additional (budget) constraint that the present cost of his consumption plan not exceed his present net worth; (*c*) for each commodity at each date in each elementary event, the total demand equals the total supply: that is, the total planned consumption equals the sum of the total resource endowments and the total planned net output (where inputs are counted as negative outputs).

Notice that (*a*) producers and consumers are 'price takers'; (*b*) for given prices there is no uncertainty about the present value of a production plan or of given resource endowments, nor about the present cost of a consumption plan; (*c*) therefore, for given prices and given producers' plans, there is no uncertainty about a given consumer's present net worth; (*d*) since a consumption plan may specify that, for a given commodity at a given date, the quantity consumed is to vary according to the event that actually occurs, a consumer's preferences among plans will reflect not only his 'taste' but also his subjective beliefs about the likelihoods of different events and his attitude towards risk (Savage 1954).

It follows that beliefs and attitudes towards risk play no role in the assumed behaviour of producers. On the other hand, beliefs and attitudes towards risk do play a role in the assumed behaviour of consumers, although for given prices and production plans each consumer knows his (single) budget constraint with certainty.

We shall call the model just described an 'Arrow–Debreu' economy. One can demonstrate, under 'standard conditions': (*a*) the existence of an equilibrium, (*b*) the Pareto optimality of an equilibrium, and (*c*) that, roughly speaking, every Pareto optimal choice of production and consumption plans is an equilibrium relative to some price system for some distribution of resource endowments and shares (Debreu 1959). In another direction of research initiated by Debreu (1970), the focus was to identify properties (like local uniqueness, finiteness) of Walrasian equilibria that were *generic* (typical or robust in a given context or in a class models). In what follows we shall use the term 'generic' informally and invite the reader to verify the exact definition from the original reference.

In the above interpretation of the Arrow–Debreu economy, all accounts are settled before the history of the economy begins, and there is no incentive to revise plans, reopen the markets or trade in shares. There is an alternative inpt, which will be of interest in connection with the rest of this article, but which corresponds to exactly the same formal model. In this second inpt, there is a single commodity at each date – let us call it 'gold' – that is taken as a numeraire at that date. A 'price system' has two parts: (1) for each date and each elementary event at that date, there is a price, to be paid in gold at the beginning date, for one unit of gold to be delivered at the specified date and event; (2) for each commodity, date, and event at that date, a price, to be paid in gold at that date and event, for one unit of the commodity to be delivered at that same date and event. The first part of the price system can be interpreted as 'insurance premiums' and the second part as 'spot prices' at the given date and event. The insurance interpretation is to be made with some reservation, however, since there is no real object being insured and no limit to the amount of insurance that an individual may take out against the occurrence of a given event. For this reason, the first part of the price system might be better interpreted as reflecting a combination of betting odds and interest rates.

Although the second part of the price system might be interpreted as spot prices, it would be a mistake to think of the determination of the equilibrium values of these prices as being deferred in real time to the dates to which they refer. The definition of equilibrium requires that the agents have access to the complete system of prices when choosing their plans. In effect, this requires that at the beginning of time all agents have available a (common) forecast of the equilibrium spot prices that will prevail at every future date and event.

Extension of the Arrow–Debreu Model to the Case in Which Different Agents Have Different Information

In an Arrow–Debreu economy, at any one date each agent may have incomplete information

about the state of the environment, but all the agents will have the *same* information. This last assumption is not tenable if we are to take good account of the effects of uncertainty in an economy. We shall now sketch how, by a simple reinpt of the concepts of production possibility set and consumption possibility set, we can extend the theory of the Arrow–Debreu economy to allow for differences in information among the economic agents.

For each date, the information that will be available to a given agent at that date may be characterized by a partition of the set of states of the environment. To be consistent with our previous terminology, we should assume that each such information partition must be at least as coarse as the partition that describes the elementary events at that date; that is, each set in the information partition must contain a set in the elementary event partition for the same date. For example, each set in the elementary event partition at a given date might specify the high temperature at that date, whereas each set in a given agent's information partition might specify only whether this temperature was higher than 75 °F, or not.

An agent's information restricts his set of feasible plans in the following manner. Suppose that at a given date the agent knows only that the state of the environment lies in a specified set A (one of the sets in his information partition at that date), and suppose (as would be typical) that the set A contains several of the elementary events that are in principle observable at that date. Then any action that the agent takes at that date must necessarily be the same for all elementary events in the set A. In particular, if the agent is a consumer, then his consumption of any specified commodity must be the same in all elementary events contained in the information set A; if the agent is a producer, then his input or output of any specified commodity must be the same for all events in A. (We are assuming that consumers know what they consume and producers what they produce at any given date.)

Let us call the sequence of information partitions for a given agent his information structure and let us say that this structure is fixed if it is given independent of the actions of himself or any other agent. Furthermore, in the case of a fixed information structure, let us say that a given plan (consumption or production) is compatible with that structure if it satisfies the conditions described in the previous paragraph, at each date.

Suppose that consumption and production possibility sets of the Arrow–Debreu economy are interpreted as characterizing, for each agent, those plans that would be feasible if he had 'full information' (that is, if his information partition at each date coincided with the elementary event partition at that date). The set of feasible plans for any agent with a fixed information structure can then be obtained by restricting him to those plans in the full information possibility set that are also compatible with his given information structure.

From this point on, all of the machinery of the Arrow–Debreu economy (with some minor technical modifications) can be brought to bear on the present model. In particular, we get a theory of existence and optimality of competitive equilibrium relative to fixed structures of information for the economic agents. We shall call this the 'extended Arrow–Debreu economy'. *We should add that differences among information structures of the agents may lead to a significant reduction of the number of active markets.* (for a fuller treatment, see Radner 1968, 1982.)

Choice of Information

There is no difficulty in principle in incorporating the choice of information structure into the extended Arrow–Debreu economy. We doubt, however, that it is reasonable to assume that the technological conditions for the acquisition and use of information generally satisfy the hypotheses of the standard theorems on the existence and optimality of competitive equilibrium.

The acquisition and use of information about the environment typically require the expenditure of goods and services, that is, of commodities. If one production plan requires more information for its implementation than another (that is, requires a finer information partition at one or more dates), then the list of (commodity) inputs should reflect the increased inputs for information. In this

manner a set of feasible production plans can reflect the possibility of choice among alternative information structures.

Unfortunately, the acquisition of information often involves a 'set-up cost', that is, the resources needed to obtain the information may be independent of the scale of the production process in which the information is used. This set-up cost will introduce a non-convexity in the production possibility set, and thus one of the standard conditions in the theory of the Arrow–Debreu economy will not be satisfied (Radner 1968).

Even without set-up costs, *there is a general tendency for the value of information to exhibit 'increasing returns', at least at low levels,* provided that the structure of information varies smoothly with its cost. This striking phenomenon leads to discontinuities in the demand for information. (for a precise statement, see Radner and Stiglitz 1984).

Critique of the Extended Arrow–Debreu Economy

If the Arrow–Debreu model is given a literal inpt, then it clearly requires that the economic agents possess capabilities of imagination and calculation that exceed reality by many orders of magnitude. Related to this is the observation that the theory requires in principle a complete system of insurance and futures markets, which system appears to be too complex, detailed, and refined to have practical significance. A further obstacle to the achievement of a complete insurance market is the phenomenon of 'moral hazard' (Arrow 1965).

A second line of criticism is that the theory does not take account of at least three important institutional features of modern capitalist economies: money, the stock market, and active markets at every date.

These two lines of criticism have an important connection, which suggests how the Arrow–Debreu theory might be improved. If, as in the Arrow–Debreu model, each production plan has a sure unambiguous present value at the beginning of time, then consumers have no interest in trading in shares, and there is no point in a stock market. If all accounts can be settled at the beginning of time, then there is no need for money during the subsequent life of the economy; in any case, the standard motives for holding money are not applicable.

On the other hand, once we recognize explicitly that there is a sequence of markets, one for each date, and not one of them complete (in the Arrow–Debreu sense), then certain phenomena and institutions not accounted for in the Arrow–Debreu model become reasonable.

First, there is uncertainty about the prices that will hold in future markets, as well as uncertainty about the environment.

Second, producers do not have a clear-cut natural way of comparing net revenues at different dates and states. Stockholders have an incentive to establish a stock exchange since it enables them to change the way their future revenues depend on the states of the environment. As an alternative to selling his shares in a particular enterprise, a stockholder may try to influence the management of the enterprise in order to make the production plan conform better to his own subjective probabilities and attitude towards risk.

Third, consumers will typically not be able to discount all of their 'wealth' at the beginning of time, because (*a*) their shares of producers' future (uncertain) net revenues cannot be so discounted and (*b*) they cannot discount all of their future resource endowments. Consumers will be subject to a sequence of budget constraints, one for each date (rather than to a single budget constraint relating present cost of his consumption plan to present net worth, as in the Arrow–Debreu economy).

Fourth, economic agents may have an incentive to speculate on the prices in future markets, by storing goods, hedging, and so on. Instead of storing goods, an agent may be interested in saving part of one date's income, in units of account, for use on a subsequent date, if there is an institution that makes this possible. There will thus be a demand for 'money' in the form of demand deposits.

Fifth, agents will be interested in forecasting the prices in markets at future dates. These prices will be functions of both the state of the

environment and the decisions of (in principle, all) economic agents up to the date in qst.

Sixth, if traders have different information at a particular date, then the equilibrium prices at that date will reflect the pooled information of the traders, albeit in a possibly complicated way. Hence, traders who have a good model of the market process will be able to infer something about other traders' information from the market prices.

Expectations and Equilibrium in a Sequence of Markets

Consider now a sequence of markets at successive dates. Suppose that no market at any one date is complete in the Arrow–Debreu sense: that is, at every date and for every commodity there will be some future dates and some events at those future dates for which it will not be possible to make current contracts for future delivery contingent on those events. In such a model, several types of 'equilibrium' concept suggest themselves, according to the hypotheses we make about the way traders form their expectations.

Let us place ourselves at a particular date-event pair; the excess supply correspondence at that date-event pair reflects the traders' information about past prices and about the history of the environment up through that date. If a given trader's excess supply correspondence is generated by preference satisfaction, then the relevant preferences will be conditional upon the information available. If, furthermore, the trader's preferences can be scaled in terms of utility and subjective probability, and conform to the expected utility hypothesis, then the relevant probabilities are the conditional probabilities given the available information. These conditional probabilities express the trader's expectations regarding the future. Although a general theoretical treatment of our problem does not necessarily require us to assume that traders' preferences conform to the expected utility hypothesis, it will be helpful in the following heuristic discussion to keep in mind this particular interpretation of expectations.

A trader's expectations concern both future environmental events and future prices. Regarding expectations about future environmental events, there is no conceptual problem. According to the expected utility hypothesis, each trader is characterized by a subjective probability measure on the set of complete histories of the environment. Since, by definition, the evolution of the environment is exogenous, a trader's conditional subjective probability of a future event, given the information to date, is well defined.

It is not so obvious how to proceed with regard to traders' expectations about future prices. We shall contrast two possible approaches. In the first, which we shall call the *perfect foresight* approach, let us assume that the behaviour of traders is such as to determine, for each complete history of the environment, a unique corresponding sequence of price systems, say $\varphi_t^*(e_t)$, where e_t is the particular event at date t. If the 'laws' governing the economic system are known to all, then every trader can calculate the sequence of functions φ_t^*. In this case, at any date–event pair a trader's expectations regarding future prices are well defined in terms of the functions φ_t^* and his conditional subjective probability measures on histories of the environment, given his current information. Traders need not agree on the probabilities of future environmental events, and therefore they need not agree on the probability distribution of future prices, *but they must agree on which future prices are associated with which events.* We shall call this last type of agreement the condition of *common price expectation functions*.

Thus, the perfect foresight approach implies that, in equilibrium, traders have common price expectation functions. These price expectation functions indicate, for each date-event pair, what the equilibrium price system would be in the corresponding market at that date-event pair. Pursuing this line of thought, it follows that, in equilibrium, the traders would have strategies (Plans) such that, if these strategies were carried out, the markets would be cleared at each date–event pair.

Call such plans *consistent*. A set of common price expectations and corresponding consistent plans is called an *equilibrium of plans, prices and price expectations* (*EPPPE*).

This model of equilibrium can be extended to cover the case in which different traders have different information, just as the Arrow–Debreu model was so extended. In particular, one could express in this way the hypothesis that a trader cannot observe the individual preferences and resource endowments of other traders. Indeed, one can also introduce into the description of the state of the environment variables that, for each trader, represent his alternative hypotheses about the 'true laws' of the economic system. In this way the condition of common price expectation functions can lose much of its apparent restrictiveness.

The situation in which traders enter the market with different nonprice information presents an opportunity for agents to learn about the environment from prices, since current market prices reflect, in a possibly complicated manner, the nonprice information signals received by the various agents. To take an extreme example, the 'inside information' of a trader in a securities market may lead him to bid up the price to a level higher than it otherwise would have been. In this case, an astute market observer might be able to infer that an insider has obtained some favourable information, just by careful observation of the price movement. More generally, *an economic agent who has a good understanding of the market is in a position to use market prices to make inferences about the* (*nonprice*) *information received by other agents*.

These inferences are derived, explicitly or implicitly, from an individual's 'model' of the relationship between the nonprice information received by market participants and the market prices. On the other hand, the true relationship is determined by the individual agents' behaviour, and hence by their individual models. Furthermore, economic agents have the opportunity to revise their individual models in the light of observations and published data. Hence, there is a feedback from the true relationship to the individual models. An equilibrium of this system, in which the individual models are identical with the true model, is called *rational expectations equilibrium* (*REE*).

This concept of equilibrium is more subtle, of course, than the ordinary concept of the equilibrium of supply and demand. In a rational expectations equilibrium, not only are prices determined so as to equate supply and demand, but individual economic agents correctly perceive the true relationship between the nonprice information received by the market participants and the resulting equilibrium market prices. This contrasts with the ordinary concept of equilibrium in which the agents respond to prices but do not attempt to infer other agents' nonprice information from the actual market prices.

Although it is capable of describing a richer set of institutions and behaviour than is the Arrow–Debreu model, the perfect foresight approach is contrary to the spirit of much of competitive market theory in that it postulates that individual traders must be able to forecast, in some sense, the equilibrium prices that will prevail in the future under all alternative states of the environment. Even if one grants the extenuating circumstances mentioned in previous paragraphs, this approach still seems to require of the traders a capacity for imagination and computation far beyond what is realistic. An equilibrium of plans and price expectations might be appropriate as a conceptualization of the ideal goal of indicative planning, or of a long-run steady state towards which the economy might tend in a stationary environment.

These last considerations lead us in a different direction, which we shall call the *bounded rationality* approach. This approach is much less well defined, but expresses itself in terms of various retreats from the hypothesis of 'fully rational' behaviour by traders, for example, by assuming that the trader's planning horizons are severely limited, or that their expectation formation follows some simple rules-of-thumb. An example of the bounded-rationality approach is the theory of *temporary* (*or momentary*) *equilibrium.*

In the evolution of a sequence of temporary equilibria, each agent's expectations will be successively revised in the light of new information about the environment and about current prices (see Grandmont 1987). Therefore, the evolution of the economy will depend upon the rules or processes of expectation formation and revision used by the agents. In particular, there might be interesting conditions under which such a

sequence of temporary equilibria would converge, in some sense, to a (stochastic) steady state. This steady state, for example, a stationary probability distribution of prices, would constitute a fourth concept of equilibrium (see Bhattacharya and Majumdar 2007).

Of the four concepts of equilibrium, the first two are perhaps the closest in the spirit to the Arrow–Debreu theory. How far do some of the conclusions of the Arrow–Debreu theory extend to this new situation? We turn now to this qst. The literature subsequent to the publication of Radner (1967, 1968, 1972) is already voluminous. The interested reader is referred to the reviews by Magill and Shafer (1991), Geanakopolos (1990), Shafer (1998), and the books by Duffie (1988), Magill and Quinzii (1996), LeRoy and Werner (2001).

Equilibrium of Plans, Prices and Price Expectations

Consider now the model of perfect-foresight equilibrium sketched above, in which the agents have common information at every date–event pair (for a precise description of the model, see Radner 1972). Three features of the situation are different from the Arrow–Debreu model: (1) there is a sequence of markets (or rather a 'tree' of markets), one for each date–event pair, no one of which is complete: (2) for each agent, there is a separate budget constraint corresponding to each date–event pair; (3) even if there is a natural bound on consumption and production, there is no single natural bound on the *positions* that traders can take in the markets for securities, if short sales are permitted, (4) there is no obvious objective for each firm to pursue, since each firm's profit is defined only for each date-event pair.

To deal with points (3) and (4), consider the following assumptions. Regarding (3), although there is no *single* natural bound on traders' positions, *some* bound is natural; for example, a commitment to deliver a quantity of a commodity vastly greater than the total supply would not be credible to moderately well-informed traders. Regarding (4), assume that the manager of each firm has preferences on the sequence of net revenues that can be represented by a continuous, strictly concave utility function. We elaborate on variations these and other assumptions. In other respects, we make the 'standard' assumptions of the Arrow–Debreu model.

A Canonical Model of Sequential Trading

For ease of exposition we sketch a matchbox model of sequential trading, variations and extensions of which have provided useful building blocks in the formal development beyond the Arrow–Debreu framework. Our exposition draws upon the excellent introduction to the literature by Shafer (1998). There are two periods, 0 and 1, with S states of nature in period l. In each| period, ℓ commodities are traded, with the trades in period 1 being contingent on the realized state s: thus, there, are $L \equiv \ell(s + 1)$ contingent commodities in the model (that is, the commodity space is R^L. We have $I \geq 2$ agents, each characterized by a utility function u_i (representing the preferences) and an endowment vector w_i (in $R^L{}_{++}$, the set of all strictly positive vectors in R^L). Denote by $w = (w_1,\ldots,w_I)$ the list of endowment vectors. The utility functions are assumed to have the appropriate smoothness and boundary conditions, strict monotonicity and strict quasiconcavity properties. Each agent is supposed to know his own characteristics and each observes the true state when it occurs in period 1. We write a vector y in R^L in the form $y = (y(0), y(1), \ldots, y(s))$ with each $y(s)$ in R^ℓ. A *spot price system* is a vector p in $R^L{}_{++}$. See Magill and Shafer (1991) for a more complete description and examples.

We first review the Arrow–Debreu competitive equilibrium in this model in which all trades and prices are decided in period 0. We denote a list of prices by $P = (P(0), P(s))$; $P(0)$ is the vector of prices for goods consumed in period 0, and $P(s)(s \geq 1)$ is the price vector in period 0 for delivery of goods in period 1 contingent on state s being the realized state. The ith agent's optimization problem is maximize$_x$ $u_i(x)$

subject to: $P(0)(x(0) - w_i(0)) + \sum_{s \geq 1} P(s)[x(s) - w_i(s)] = 0$.

Note that the agent *faces a single budget constraint.* A *competitive equilibrium* is a collection $\left((x_i)_{i \leq 1}, \overline{P}\right)$ such that

(i) x_i solves agent i's optimization problem at $\overline{P}$; and
(ii) $\sum_i [x_i(s) - w_i(s)] = 0$ for s = 0, . . ., S.

To introduce the model of *sequential trading,* we suppose there are J *assets* (or *securities*) which are traded in period 0 and return dividends in period 1. A unit of asset j will cost q_j units of account payable in 0 ($q = (q_1, \ldots, q_J)$) in R^J) and return $V_j(s)$ units of account in period 1 in state s. An asset is called *nominal* if the returns are given exogenously; it is called *real* if the return at state s is the market value of a commodity vector, that is, if $V_j(s) = P(s)a_j(s)$ for some vector $a_j(s)$ in R^ℓ. Of course, mixtures are possible, but, for our matchbox model, we consider only the pure real asset case or the pure nominal asset case. Denote by V the $S \times J$ matrix of returns that has in row s and column j the dividend, $V_j(s)$, of asset j in state s, and let $v(s)$ denote the vector of the returns in state s. In the real asset case, the list of vectors $a = (a_j(s))$ parameterizes the asset structure, and the returns matrix $V(P)$ is a function of p. In the nominal case the returns matrix V itself parametrizes the asset structure. Denote by Z_j the amount purchased of asset j, with $z = (z_1, \ldots, z_J)$ being the portfolio of assets. The amount z_j may be positive or negative; the assets are considered to be in zero net supply.

We now apply to this model the concept of an *equilibrium of plans, prices, and price expectations.* At a spot price system p and an asset price system q, define an agent's optimization problem as

$$\text{maximize}_{x,z}\ u_i(x)$$
$$\text{subject to: } P(0)\,x(0) - w_i(0) = -qz,$$
$$P(s)(x(s) - w_i(s)) = \Sigma_j\, v_j(s) z_j,\ s = 1, \ldots, S.$$

We should stress that the agent faces *a multiplicity of budget constraints.* The first constraint listed is that the net expenditure on goods plus the cost of the portfolio of assets must sum to zero. The constraints for $s \geq 1$ indicate that if s is the realized state in period 1, the *net* expenditure on goods must equal the dividends of the asset portfolio. The purchase of the assets in period 0 and their dividends in period 1 provides a means both for transferring income between period 0 and period 1 and for transferring income across the potential states in period 1. Note that these constraints preclude the agent from planning bankruptcy in any state; implicit in the constraints is an infinite penalty for bankruptcy.

An equilibrium of plans, prices, and price expectations is a list $((x_i, z_i)_{I \geq 1}, (P, q))$ such that:

(i) $(\overline{x}_i, \overline{z}_i)$solves age|nt i's optimization problem at $(\overline{p}, \overline{q})$;
(ii) $\sum_i (\overline{x}_i(s) - w_i(s)) = 0, s = 0, \ldots, S$; and
(iii) $\sum_i \overline{z}_i = 0$.

The interpretation of the EPPPE concept is as follows. In period 0, each agent i observes the current spot prices $\overline{p}(0)$and the asset prices $\overline{q}$, the 'prices'. Then, based on 'price expectations' about spot prices in period 1, say $p^e(s)$, $s = 1, \ldots, S$, the agent solves the optimization problem, forming the 'plans' $(\overline{x}_i, \overline{z}_i)$. If it turns out that the price expectations of all the agents are the *same* and the *common* expectation $\overline{p}(s)(s = 1, \ldots, S)$ is such that, together with the observed prices $\overline{p}(0)$ and $\overline{q}$, all markets clear, then we are in an EPPPE equilibrium.

The information requirements of this model are quite strict. Each agent is fully informed in the sense that he or she will be able to verify the true state once it occurs. In addition, each agent knows exactly the distribution of returns of each asset across the states (that is, each agent knows each $a_j(\cdot)$ in the real asset case and each $v_j(\cdot)$ in the nominal case).

We now define *completeness of markets* in this model. The formal definition is that, for any possible vector of units of account across the S states of nature, an agent can form a portfolio of assets that gives this distribution of returns. That is, for any S-vector y of net expenditures on goods in

period 1($y_s = P(s)(x(s) - w(s))$), there is some portfolio z such that $y = Vz$. In our model, for the nominal case this is equivalent to the *returns matrix* V having the rank S, so that the *column vectors* of V *span all* of R^S. In particular, there must be $J \geq S$ assets. In the case of real assets the rank of the returns matrix $V(P)$ is a function of p, and is thus endogenous to the model. However, *since* $V(\cdot)$ *is linear in p, it has a 'generic' rank* (see Magill and Shafer 1990, for a precise formulation), and this rank is the maximum rank the returns matrix can take on at any p. The real asset structure is defined to be complete if *this generic rank is S.* Again *this requires* $J \geq S$. We note for later reference that, if one imposes restrictions on the size of trades an agent can make in the asset market, then these markets cannot be complete regardless of the number of assets, since by restricting asset trades z we cannot in general expect to express every vector in R^S in the form Vz.

A very useful implication of completeness will now be stated. Suppose the market is complete. Then given a competitive equilibrium $\left((\overline{x}_i)_{i\geq 1}\overline{P}\right)$ one can construct a canonical model with an appropriate EPPPE equilibrium $\left((x_i z_i)_{i\geq 1},\overline{p},\overline{q}\right)$: it should be stressed that the *allocation* $(\overline{x}_i)_{i\geq 1}$ *is the same.* Conversely, given an EPPPE $\left((x_i z_i)_{i\geq 1},\overline{p},q\right)$:, one can construct a competitive equilibrium $\left((\overline{x}_i)_{i\geq 1},\overline{P}\right)$. This 'translation' of the frameworks has been exploited in the formal analysis, and has clear implications for the *optimality* of an EPPPE (see Magill and Shafer 1991; Shafer 1998).

Radner (1972) demonstrated that an equilibrium exists in a more general model provided we impose bounds on the size of trades in the asset markets.

Existence Theorem 1 *In the canonical model described above, if each agent's optimization problem is modified with an additional constraint of the form* $z \geq b$, *then an equilibrium of plans, prices, and price expectations exists.*

Reasonable or not, the 'ad hoc' constraint in this result posed an intellectual challenge for subsequent researchers. Moreover, as Hart (1975) discovered, without exogenous bounds on the size of trades in the asset markets, equilibria may fail to exist. The main characteristic of Hart's example is that the *returns matrix changes rank with p*, and one approach has been to restrict attention to asset structures that do not exhibit this behaviour. Cass (2006) and Werner (1985) showed that, if one restricts attention to *pure nominal asset structures,* then equilibria exist *without* imposing lower bounds. Similarly, Geanakoplos and Polemarchakis (1986) observed that, in the *real asset case,* if all assets are denominated in terms of the market value of a single good, then the returns matrix $V(P)$ will have constant rank and – just as in the nominal case – equilibria always exist. In general, then, we have the following theorem.

Existence Theorem 2 *In the canonical model, if the asset structure is such that the matrix of returns has constant rank, then an equilibrium of plans, prices, and price expectations exists.*

To get some insights into the non-existence issue, we provide a simple example. Consider the Radner model with one state in period 1, one asset, two consumers, and two goods, with the following data. Endowments are $w_i(s) = (1, 1)$ for s = 0, 1. Utility functions are $u_1(x) = v_1(x(0)) + v_1(x(1))$ with.

$v_1(x) = (1/3)\ell n x_1 + (2/3)\ell n x_2$ for agent 1 and $u_2(x) = u_2(x(0)) + (1/2)v_2(x(1))$ with $v_2(x) = (2/3)\ell n x_1 + (1/3)\ell n x_2$ for agent 2. The competitive equilibrium prices can be easily computed in this log-linear economy; they are P_1 (0) = 11/36, $P_2(0) = 10/36$, $P_1(1) = 7/36$, and $P_2(1) = 8/36$. Now consider a Radner version of the model, with one real asset given by $a_1 = (8, -7)$. The return on this asset in state 1 is $V = P_1(1)8 - P_2(1)7$, so investing in this asset is essentially a bet that the relative price of good 1 in terms of good 2 is greater than 7/8. Since there is one state and one asset, this is the complete markets case. Note that the 1×1 returns matrix drops rank precisely in the case the relative price is 7/8 in state 1.

We now show that this model does not have an EPPPE equilibrium. First, we try for an equilibrium with the return $V \neq 0$. If such an equilibrium existed, it would have to coincide with the

competitive equilibrium since V has rank 1, but in the unique competitive equilibrium the period-1 ratio is 7/8, so $V = 0$, a contradiction. Second, consider the possibility of an equilibrium with $V = 0$. Then there would be no transfers of income between periods 0 and 1, so the period-1 equilibrium would have to coincide with the static competitive equilibrium with the utility functions v_1 and v_2. But it is easy to see from the symmetry of the functions and the equal endowments that the relative price ratio in this case would be 1, and thus $V \neq 0$, again a contradiction. Thus no equilibrium exists. Note, however, that this example is *not* robust: alter the asset a small amount so that $V \neq 0$ at the price ratio 7/8 and then the complete markets case will work; or alter endowments or utility parameters a little so that the period-1 price ratio is no longer 7/8, and the complete markets case again works.

This example gives a clue on how to proceed for the existence problem with real assets when markets are complete. One aims at generic results. In this case, remember, $V(P)$ has constant rank S on an open set of full measure in the space of prices. As we have mentioned above, an EPPPE at which $V(P)$ has rank S is equivalent to a competitive equilibrium with a complete set of contingent commodity contracts. That is, the allocations are the same, and there is an easy correspondence between competitive equilibrium prices and the corresponding EPPPE prices. Thus a natural approach is first to obtain a competitive equilibrium, which always exists in our model, and then to construct the corresponding EPPPE prices. If at these prices $V(P)$ has rank S, then we have an EPPPE.

Kreps (1982) made the critical observation that, if the rank of the returns matrix is less that S, then a perturbation of the returns structure a will restore $V(P)$ to full rank (as in our preceding ex), and thus we will have an EPPPE. That is, *generically* in a, an EPPPE exists. Similarly, Magill and Shafer (1990) observed that a small perturbation of endowment would cause the competitive equilibrium prices to move into the region where V(P) has full rank, and thus an equilibrium exists generically in endowments.

Generic Existence Theorem 1 *In the canonical model:*

1. *if $J \geq S$ then, for each $w \in \mathrm{R}^{L}_{++}$, an equilibrium of plans, prices, and price expectations exists for almost all (asset structure) a in $R^{\ell \mathrm{JS}}$*
2. *for each asset structure a for which $V(\cdot)$ has generic rank S, an equilibrium of plans, prices, and price expectations exists for almost all endowment lists w in $R^{LI}{}_{++}$.*

In the case where both V(P) *can change rank with* p *and markets are not complete* (*in particular, if $J < S$*), the trick of first obtaining a competitive equilibrium and then converting it to an EPPPE equilibrium is no longer available. Nevertheless, by defining a 'pseudo' equilibrium concept that replaces the competitive equilibrium in the argument for the complete market case, Duffie and Shafer (1985) were able to show that an EPPPE equilibrium exists generically in both a and w.

Generic Existence Theorem 2 *In the canonical model with all real assets, an equilibrium of plans, prices, and price expectations exists for almost all (a, w) in $R^{\ell \mathrm{JS}} \times R^{L}{}_{++}$*

We now look at the issue of *local uniqueness.* In what follows, in 'counting' equilibria we are counting the number of equilibrium allocations, since there may be certain redundancies in equilibrium prices. In the complete market case this is fairly straightforward, requiring only an adaptation of Debreu's (1970) argument, since competitive equilibria and EPPPE equilibria coincide when the returns matrix has full rank. In the case of incomplete markets and all real assets, an argument similar to Debreu's applied to the 'pseudo' equilibrium also works.

Local Uniqueness Theorem 1 *In the canonical model, if the asset structure is such that markets are complete, or if all assets are real, then for almost all w in the space of endowment lists there exist a finite number of EPPPE, and each equilibrium is locally a smooth function of endowment lists w and asset structures a or V.*

The situation with nominal assets and incomplete markets is, however, completely different. In

this case there is 'serious indeterminacy' as the following result suggests.

Local Uniqueness Theorem 2 *In the canonical model with nominal assets, let the returns matrix V satisfy $J < S$ and $I > J$. Then, for almost all w in the space of endowment lists, the set of allocations of an EPPPE contains a set homeomorphic to R^{S-1}.* (See Geanakoplos and Mas-Colell 1989, and a similar result by Balasko and Cass 1989.)

Next, we turn to Pareto optimality. At an EPPPE equilibrium at which the returns matrix has rank S, the resulting equilibrium allocation will also be a competitive equilibrium allocation, and thus fully Pareto efficient. This leads to the following theorem.

Pareto Optimality Theorem 1 *For the canonical model:*

(1) *in the nominal asset case, if V has rank S then every EPPPE equilibrium allocation is Pareto efficient*;
(2) *in the real asset case, if the asset structure is such that the generic rank of $V(\cdot)$ is S then, for almost all w in the space of endowment lists, EPPPE equilibrium allocations are Pareto efficient.*

In case of incomplete markets, one certainly does not expect full allocative efficiency. The following result (Genakopolos and Polemarchakis 1986) emphasizes the failure of the first fundamental theorem.

Non-Optimality Theorem 1 *For the canonical model, if $J < S$ then, generically in (w, a) in R^{L}_{++}, EPPPE equilibrium allocations are not Pareto efficient.*

One might hope that, in an appropriate sense, EPPPE equilibria are constrained efficient. There is still no generally accepted notion of what the correct definition of 'constrained efficiency' might be in this case; some argue that the concept cannot be properly defined unless the reasons for incompleteness of markets are endogenously embedded into the model. Nevertheless, we can discuss certain efficiency properties of the equilibria. First, there are *robust* examples of the model with multiple equilibria in which two of the EPPPE equilibria have the property that one Pareto dominates the second (Shafer 1998). As a consequence of such robust examples, any efficiency property that the incomplete market EPPPE equilibria may possess must be 'weak'. One approach to constrained efficiency is to follow the Lange-Lerner tradition: if a central planner were permitted to choose the asset portfolios for the agents, and then allow agents to trade freely on competitive markets for commodities, could the planner improve upon an EPPPE equilibrium? The answer is, in an appropriate generic sense, 'yes' (Geanakopolos and Polemarchakis 1986). This is, of course, not possible if markets are complete.

Another 'natural' question to ask: is there a connection between how inefficient the EPPPE equilibria are and how incomplete the markets may be? One measure of incompleteness is $S - J$, assuming the J assets give the returns matrix a generic rank J. By introducing a new asset, which reduces the incompleteness in this sense, does efficiency improve? The answer is 'no', again due to an example of Hart (1975), in which a new asset is introduced but the new EPPPE equilibrium allocation is Pareto dominated by the original EPPPE equilibrium allocation. This suggests that perhaps this notion of 'almost complete' is at fault.

Production

We first discuss the question of existence of equilibrium, but before paraphrasing the existence theorem we must define what we shall call a *pseudo-equilibrium*.

The definition of pseudo-equilibrium is obtained from the definition of equilibrium by replacing the requirement of consistency of plans by the condition that each date and each event the difference between total saving and total investment (by consumers) is smaller at the pseudo-equilibrium prices than at any other prices.

One can prove (Radner 1972) that under assumptions about technology and consumer preferences similar to those used in the Arrow–Debreu theory, and with the additional assumptions sketched above: (*a*) there exists a pseudo-

equilibrium; (*b*) if in a pseudo-equilibrium the current and future prices on the stock market are all strictly positive, then the pseudo-equilibrium is an equilibrium.

The crucial difference between this theorem and the corresponding one in the Arrow–Debreu theory seems to be due to the form taken by Walras's Law, which in this model can be paraphrased by saying that saving must be at least equal to investment at each date in each event. This form derives from the replacement of a single budget constraint (in terms of present value) by a sequence of budget constraints, one for each date-event pair.

In the above model with production, the 'shareholders' have unlimited liability, and therefore have a status more like that of partners than of shareholders, as these terms are usually understood. One way to formulate limited liability for shareholders is to impose the constraint on producers that their net revenues be non-negative at each date-event pair. However, in this case producers' correspondences may not be upper semi-continuous. This is analogous to the problem that arises when, for a given price system, the consumer's budget constraints force him to be on the boundary of his consumption set. In the case of the consumer, this situation is avoided by some assumption (see Debreu 1959, notes to ch. 5, pp. 88–9; Debreu 1962). However, for the case of the producer, it is not considered unusual in the standard theory of the firm that, especially in equilibrium, the maximum profit achievable at the given price system could be zero (for example, in the case of constant returns to scale).

What are conditions on the producers and consumers that would directly guarantee the existence of an equilibrium, not just a pseudo-equilibrium? In other words, under what conditions would the share markets be cleared at every date-event pair? Notice that, if there is an excess supply of shares of a given producer j at a date-event pair (t, e), then at date $(t + 1)$ only part of the producer's revenue will be 'distributed'. One would expect this situation to arise only if his revenue is to be negative in at least one event at date $t + 1$; thus, at such a date-event pair the producer would have a deficit covered neither by 'loans (that is, not offset by forward contracts) nor by shareholders' contributions. In other words, the producer would be 'bankrupt' at that point.

One approach might be to eliminate from a pseudo-equilibrium all producers for whom the excess supply of shares is not zero at some date–event pair, and then search for an equilibrium with the smaller set of producers, and so on, successively reducing the set of producers until an equilibrium is found. This procedure has the trivial consequence that an equilibrium always exists, since it exists for the case of pure exchange (the set of producers is empty)! This may not be the most satisfactory resolution of the problem, but it does point up the desirability of having some formulation of the possibility of 'exit' for producers who are not doing well.

Although the above model with production does not allow for 'exit' of producers (except with the modification described in the preceding paragraph), it does allow for 'entrance' in the following limited sense. A producer may have zero production up to some date, but plans to produce thereafter; this is not inconsistent with a positive demand for shares at preceding dates.

The creation of new 'equity' in an enterprise is also allowed for in a limited sense. A producer may plan for a large investment at a given date-event pair, with a negative revenue. If the total supply of shares at the preceding date-event pair is nevertheless taken up by the market, this investment may be said to have been 'financed' by shareholders.

The above assumptions describe a model of producer behaviour that is not influenced by the shareholders or (directly) by the prices of shares. A common alternative hypothesis is that a producer tries to maximize the current market value of this enterprise. There seems to us to be at least two difficulties with this hypothesis. First, there are different market values at different date-event pairs, so it is not clear how these can be maximized simultaneously. Second, the market value of an enterprise at any date-event pair is a price, which is supposed to be determined, along with other prices, by an equilibrium of supply and demand. The 'market-value-maximizing' hypothesis would seem to require the producer to predict, in some sense, the effect of a change in his plan on

a price *equilibrium*: in this case, the producers would no longer be price-takers, and one would need some sort of theory of general equilibrium for monopolistic competition.

There is one circumstance in which the value of the firm can be defined unambiguously, given the system of present prices and common expectations about future prices. Call a price system *arbitrage-free* if it is not possible to make a sure, positive cash flow from trading, without a positive investment. An equilibrium price system is, a fortiori, arbitrage-free. One can show (see Radner 1967; Harrison and Kreps 1979; Duffie and Shafer 1985 that an arbitrage-free price system implicitly determines a system of 'insurance premiums' for a corresponding family of events. This means that by suitable trading one can insure oneself against the occurrence of any of these events. If these events include all of the uncertain events that may affect the (uncertain) revenues of the firm, then they can be used in a natural way to define a present value of the firm at any date–event pair, for any production plan of the firm, and no probability judgements are needed to calculate the value. On the other hand, if the family of 'insurable events' is not rich enough, then the value is a random variable, and stockholders may not agree on its probability distribution.

A survey of results on the generic existence of equilibrium with production (and stock markets) is given in Magill and Shafer (1991) (see also Duffie 1988, ch. 2).

Rational Expectations Equilibrium

The formal study of rational expectations equilibrium was introduced by Radner (1967); it was taken up independently by Lucas (1972) and Green (1973), and further investigated by Grossman, Allen, Jordan, and others. We should emphasize that we are concerned here with the aspect of 'rational expectations' in which traders make inferences from market prices about other traders' information, a phenomenon that is of interest only when traders do not all have the same nonprice information. The term 'rational expectations equilibrium' (REE) has also been used to describe a situation in which traders correctly forecast (in some sense or other) the probability distribution of future prices. (see Radner 1982, for references to the work of Muth and others on this topic.)

The concept of REE has been used to make a number of interesting predictions about the behaviour of markets: see, for example, Futia (1979, 1981) and the references cited there. A sound foundation for such applications requires the investigation of conditions that would ensure the existence and stability of REE.

We adopt the convention that the future utility of the commodities to each trader depends on the *state of the environment.* With this convention, we can model the inferences that a trader makes from the market prices and his own nonprice information signal by a family of conditional probability distributions of the environment given the market prices and his own nonprice information. We shall call such a family of conditional distributions the trader's *market model.* Given such a market model, the market prices will influence a trader's demand in two ways: first, through his budget constraint, and second, through his conditional expected utility function. It is this second feature, of course, that distinguished theories of rational expectations equilibrium from earlier models of market equilibrium.

Given the traders' market models, the equilibrium prices will be determined by the equality of supply and demand in the usual way, and thus will be a deterministic function of the joint nonprice information that the traders bring to the market. In order for the market models of the traders to be 'rational', they must be consistent with that function. To make this idea precise, it will be useful to have some formal notation. Let p denote the vector of market prices, e denote the (utility-relevant) state of the environment, and s_i denote traders i's nonprice information signal ($i = 1, \ldots, I$). The joint nonprice information of all traders together will be denoted by $s = (s_1,...,s_I)$. We shall call s the 'joint signal'. (The term 'state of information' is also commonly applied to this array.) Trader i's market model, say m_i, is a family of conditional probability distributions of e, given s_i and p. Given the traders' market models, the equilibrium price vector will be

some (measurable) function of the joint nonprice information, say $p = \varphi(s)$.

To model the required rationality of the traders' models, suppose that, for each i, trader i has (subjective) prior beliefs about the environment and the information signals that are expressed by a joint probability distribution, say Q_i of e and s. These prior beliefs need not, of course, be the same for all traders. Given the price function φ, a *rational* market model for trader i would be the family of conditional probability distributions of e, given s_i and p, that are derived from the distribution Q_i and the price function φ; thus (supposing e and s to be discrete variables),

$$m_i\left(e' \mid s_i', p'\right) = \mathrm{Pr}obQ_1(e = e \backslash si \text{ and } P(s) = p'). \tag{1}$$

A given price function φ, together with the rationality condition (1), would determine the total market excess supply for each price vector p and each joint information signal s, say $Z(P,s,\varphi)$. Note that the excess supply for any p and s *depends also on the price function* φ, since (in principle) the entire price function is used to calculate the conditional distribution in (1). We can now define a *rational expectations equilibrium* (REE) to be a price function φ^* such that, for (almost) every s, excess supply is zero at the price vector $\varphi(s)$, that is,

$$Z(\varphi^*(s), s, \varphi^*) = 0, \text{for almost every s}. \tag{2}$$

If markets are incomplete, the existence of REE is not assured by the 'classical' conditions of ordinary general equilibrium analysis. Even under such conditions, if traders condition their expected utilities on market prices, then their demands can be discontinuous in the price function. Specific examples of the nonexistence of REE due to such discontinuities were given by Kreps (1977), Green (1977), and others. These examples naturally led theorists to question whether the absence of REE is pervasive or is confirmed to a 'negligible' set of such examples. The work of Radner, Allen, and Jordan (see Jordan and Radner 1982; Allen 1986, for references) provided – in a certain context – an essentially complete answer, which can be loosely summarized in the statement that *REE exists generically except when the dimension of the space of private information is equal to the dimension of the price space*. (Recall that REE exists *generically* in a given model if, for any vector of parameters values for which REE does not exist, there are arbitrarily small perturbations of the parameters for which REE does exist.) Furthermore, if the dimension of the space of private information is strictly *less* than the dimension of the price space, then generically there is a REE that is *fully revealing*, that is, in which the price reveals to each trader all the nonprice information used by *all* traders (Radner 1979; Allen 1981).

Equilibrium and Learning with Imperfect Price Models

A more stringent 'rational expectations' requirement would concern the opportunities that traders might have for learning from experience. For example, suppose that there is a market at each of a succession of dates t, and that the successive exogenous vectors (e_t, s_t) are independent and identically distributed. Suppose further that at the beginning of date t trader i knows the *past* history of environments, prices, and his own nonprice information. On the basis of this history he updates his initial market model to form a current market model. These current market models, together with the nonprice information signals at date t, then determine an equilibrium price at date t, say p_t^*, as above. The updating of models constitutes the *learning* process of the traders. For a given learning process, one might ask whether the process converges in any useful sense, and if so, whether the models are asymptotically consistent with the (endogenously determined) actual relationship between signals and equilibrium prices, that is whether they converge to a REE. In this case one would say that the REE is *stable* (relative to the learning process).

Thus far, answers to this question are only fragmentary. Bray (1982) has studied a simpler linear asset-market model in which, at each date, each trader i updates his model by calculating an

ordinary least-squares estimate of the regression of e on p and s_i using all the past values (e_t, p_t*, s_{it}). For this example, Bray proves stability.

On the other hand, Blume and Easley (1982) present a somewhat less optimistic view of the possibility of learning rational expectations. They define a class of learning procedures by which traders use successive observations to form their subjective models, where the term model for trader i means a conditional distribution of s, given s_i and p. They show that rational expectations equilibria are at least 'locally stable' under learning, but that learning possesses may also get stuck at a profile of subjective models that is not an REE. The learning procedures defined by Blume and Easley are applied to a fairly general class of stochastic exchange environments that do not possess the special linear structure of the above example. However, to accommodate this additional generality, Blume and Easley constrain traders to choose their subjective models from a fixed finite set of models and convex combinations thereof. Hence, for some profiles of subjective models, market clearing may result in a 'true' model that lies outside the admissible set. It is then intuitively plausible that a natural learning procedure could get stuck at a profile of subjective models that differs from the resulting true model but is in some sense the best admissible approximation to the true model, even if the admissible set contains an REE model. This phenomenon is illustrated in Section 5 of their paper. For a review of a number of themes related to learning, see Blume and Easley (1998).

Bibliography

Allen, B. 1981. Generic existence of completely revealing equilibria for economies with uncertainty when prices convey information. *Econometrica* 49: 1173–1199.

Allen, B. 1986. General equilibrium with rational expectations. In *Contributions to mathematical economics: Essays in honor of Gerard Debreu*, ed. A. Mas-Colell and W. Hildenbrand. Amsterdam: North-Holland.

Arrow, K.J. 1953. Le rôle de valeurs boursières pour la repartition la meilleure des risques. *Econometrie* 11, 41–8. Trans. as 'The role of securities in the optimal allocation of risk-bearing', Review of Economic Studies 31 (1964): 91–6.

Arrow, K.J. 1965. *Aspects of the theory of risk-bearing*. Yijo Johansson Foundation: Helsinki.

Arrow, K.J., and F.H. Hahn. 1971. *General competitive analysis*. San Francisco: Holden-Day.

Balasko, Y., and D. Cass. 1989. The structure of financial equilibrium I: Exogenous yields and unrestricted participation. *Econometrica* 57: 135–162.

Bhattacharya, R., and M.K. Majumdar. 2007. *Random dynamical systems: Theory and applications*. Cambridge: Cambridge University Press.

Blume, L.E., and D. Easley. 1982. Learning to be rational. *Journal of Economic Theory* 26: 318–339.

Blume, L.E., and D. Easley. 1998. Rational expectations and rational learning. In *Organizations with incomplete information*, ed. M. Majumdar. Cambridge: Cambridge University Press.

Bray, M. 1982. Learning, estimation and the stability of rational expectations. *Journal of Economic Theory* 26: 318–339.

Burke, J. 1986. Existence of equilibrium for incomplete market economies with production and stock trading. Unpublished manuscript. Texas A&M University.

Cass, D. 2006. Competitive equilibrium with incomplete financial markets. *Journal of Mathematical Economics* 42: 384–405.

Debreu, G. 1953. *Une économie de l'incertain. Mimeo, Electricité de France, Paris*. Trans. as 'Economics under Uncertainty', in G. Debreu, *Mathematical Economics, Twenty Papers of Gerard Debreu.* Cambridge: Cambridge University Press, 1983.

Debreu, G. 1959. *Theory of value*. New York: Wiley.

Debreu, G. 1962. New concepts and techniques for equilibrium analysis. *International Economic Review* 3: 257–273.

Debreu, G. 1970. Economies with a finite set of equilibria. *Econometrica* 38: 387–392.

Duffie, D. 1988. *Security markets: Stochastic models*. Boston: Academic Press.

Duffie, D., and W. Shafer. 1985. Equilibrium in incomplete markets I: A basic model of generic existence. *Journal of Mathematical Economics* 14: 285–300.

Duffie, D., and W. Shafer. 1987. Equilibrium in incomplete markets II: Generic existence in stochastic economies. *Journal of Mathematical Economics* 15: 199–216.

Futia, C.A. 1979. Stochastic business cycles. Unpublished manuscript. Murray Hill: AT&T Bell Laboratories.

Futia, C.A. 1981. Rational expectations in stationary linear models. *Econometrica* 49: 171–192.

Geanakopolos, J. 1990. An introduction to general equilibrium with incomplete asset markets. *Journal of Mathematical Economics* 19: 1–38.

Geanakoplos, J., and A. Mas-Colell. 1989. Real indeterminacy with financial assets. *Journal of Economic Theory* 47: 22–38.

Geanakoplos, J., and H. Polemarchakis. 1986. Existence, regularity, and constrained suboptimality of competitive allocations when the asset market is incomplete. In *Uncertainty, information and communication: Essays in honor of Kenneth Arrow*, ed. W.P. Heller, R.M. Starr, and

D.A. Starrett, vol. 3. Cambridge: Cambridge University Press.
Grandmont, J.-M. 1987. Temporary equilibrium. In *The new palgrave*, ed. J. Eatwell, M. Milgate, and P. Newman, vol. 4. London: Macmillan.
Green, J. 1973. Information, inefficiency, and equilibrium. Discussion Paper No. 284, Harvard Institute of Economic Research.
Green, J. 1977. The nonexistence of informational equilibria. *Review of Economic Studies* 44: 451–463.
Harrison, J.M., and D.M. Kreps. 1979. Martingales and arbitrage in multiperiod securities markets. *Journal of Economic Theory* 20: 381–408.
Hart, O. 1975. On the optimality of equilibrium when the market structure is incomplete. *Journal of Economic Theory* 11: 418–443.
Jordan, J.S., and R. Radner. 1982. Rational expectations in microeconomic models: An overview. *Journal of Economic Theory* 26: 201–223.
Kreps, D.M. 1977. A note on 'fulfilled expectations' equilibria. *Journal of Economic Theory* 14: 32–43.
Kreps, D.M. 1982. Multi-period securities and the efficient allocation of risk: A comment on the Black–Scholes option pricing model. In *The economics of uncertainty and information*, ed. J. McCall. Chicago: University of Chicago Press.
LeRoy, S.E., and J. Werner. 2001. *Principles of financial economics*. Cambridge: Cambridge University Press.
Lucas, R.E. 1972. Expectations and the neutrality of money. *Journal of Economic Theory* 4: 103–124.
Magill, M., and M. Quinzii. 1996. *Theory of incomplete markets*. Cambridge, MA: MIT Press.
Magill, M., and W. Shafer. 1990. Characterization of generically complete real asset structures. *Journal of Mathematical Economics* 19: 167–194.
Magill, M., and W. Shafer. 1991. Incomplete markets. In *Handbook of mathematical economics*, ed. W. Hildenbrand and H. Sonnenschein, vol. 4. New York: Elsevier.
Radner, R. 1967. Equilibrie des marchés à terme et au comptant en cas d'incertitude. *Cahiers d'Econometrie* 4, 35–52. Paris: CNRS.
Radner, R. 1968. Competitive equilibrium under uncertainty. *Econometrica* 36: 31–58.
Radner, R. 1972. Existence of equilibrium of plans, prices, and price expectations in a sequence of markets. *Econometrica* 40: 289–304.
Radner, R. 1979. Rational expectations equilibrium: Generic existence and the information revealed by prices. *Econometrica* 47: 655–657.
Radner, R. 1982. Equilibrium under uncertainty. In *Handbook of mathematical economics*, ed. K.J. Arrow and M.D. Intriligator, vol. 2. Amsterdam: North-Holland.
Radner, R., and J.E. Stiglitz. 1984. A nonconcavity in the value of information. In *Bayesian models of economic theory*, ed. M. Boyer and R.E. Kihlstrom. Amsterdam: North-Holland.
Savage, L.J. 1954. *The foundations of statistics*. New York: Wiley.
Shafer, W. 1998. Equilibrium with incomplete markets in a sequence economy. In *Organizations with incomplete information*, ed. M. Majumdar. Cambridge: Cambridge University Press.
Werner, J. 1985. Equilibrium in economies with incomplete financial markets. *Journal of Economic Theory* 36: 110–119.
Younes, Y. 1985. Competitive equilibrium and incomplete market structures. Unpublished manuscript. Department of Economics, University of Pennsylvania (revised October 1986).

Uncovered Interest Parity

Peter Isard

Abstract

This article provides an overview of the uncovered interest parity assumption. It traces the history of the concept, summarizes evidence on the empirical validity of uncovered interest parity, and discusses different interpretations of the evidence and the implications for macroeconomic analysis. The uncovered interest parity assumption has been an important building block in multi-period models of open economies and, although its validity is strongly challenged by the empirical evidence, at least at short time horizons, its retention in macroeconomic models is supported on pragmatic grounds by the lack of much empirical support for existing models of the exchange risk premium.

Keywords

Arbitrage; Capital controls; Covered interest parity (CIP); Credit risk; Exchange market intervention; Exchange rate dynamics; Exchange rate expectations; Exchange risk premium; Forward exchange; Incomplete information; Inflation expectations; Interest rate differentials; Jensen's inequality; Keynes, J. M.; Peso problem; Portfolio balance models; Prediction bias; Rational bubbles; Rational expectations; Rational learning; Risk

premium; Spot exchange; Unbiasedness hypothesis; Uncovered interest parity (UIP)

JEL Classifications
F31

The assumption of uncovered interest parity (UIP) is an important building block for macroeconomic analysis of open economies. It provides a simple relationship between the interest rate on an asset denominated in any one country's currency unit, the interest rate on a similar asset denominated in another country's currency, and the expected rate of change in the spot exchange rate between the two currencies.

The theory of interest parity received prominence from expositions by Keynes (for example, 1923, pp. 115–139), whose attention had been captured by the rapid expansion of organized trading in forward exchange following the First World War (Einzig 1962, pp. 239–241, 275). Although an understanding of the forward exchange market must have developed within various banking circles during the second half of the 19th century, apart from an isolated exposition by a German economist, Walther Lotz (1889), the 19th-century literature on foreign exchange theory apparently dealt only with spot exchange rates (Einzig 1962, pp. 214–215). Forward exchange trading gave rise to the notion of covered interest parity (CIP), which related the differential between domestic and foreign interest rates to the percentage difference between forward and spot exchange rates. Since it was clear that forward rates also reflected perceptions about future spot rates, it was a short step to the assumption of UIP, which builds on the theory of CIP by essentially postulating that market forces drive the forward exchange rate into equality with the expected future spot exchange rate.

Basic Concepts

The concept of interest parity recognizes that portfolio investors at any time t have the choice of holding assets denominated in domestic currency, offering the own rate of interest r_t between times t and $t + 1$, or of holding assets denominated in foreign currency, offering the own rate of interest r_t^*. Thus, an investor starting with one unit of domestic currency should compare the option of accumulating $1 + r_t$ units with the option of converting at the spot exchange rate into st units of foreign currency, investing in foreign assets to accumulate $s_t(1 + r_t^*)$ units of foreign currency at time $t + 1$, and then reconverting into domestic currency. If the domestic and foreign assets differ only in their currencies of denomination, and if investors have the opportunity to cover against exchange rate uncertainty by arranging at time t to reconvert from foreign to domestic currency one period later at the forward exchange rate f_t (in units of foreign currency per unit of domestic currency), then market equilibrium requires the condition of CIP:

$$1 + r_t = s_t(1 + r_t^*)/f_t. \tag{1}$$

If condition (1) did not hold, profitable market arbitrage opportunities could be exploited without incurring any risks.

Investors also have the opportunity to leave their foreign currency positions uncovered at time t and to wait until time t+1 to make arrangements to reconvert into domestic currency at the spot exchange rate s_{t+1}. Unlike f_t, the value of s_{t+1} is unknown at time t, and so the attractiveness of holding an uncovered position must be assessed in terms of the probabilities of different outcomes for s_{t+1}. The assumption of UIP postulates that markets will equilibrate the return on the domestic currency asset with the expected value at time $t(E_t)$ of the yield on an uncovered position in foreign currency:

$$\begin{aligned} 1 + r_t &= E_t\left[s_t(1 + r_t^*)/s_{t+1}\right] \\ &= s_t(1 + r_t^*)E_t(1/s_{t+1}). \end{aligned} \tag{2}$$

This is essentially equivalent to combining the CIP condition with the assumption that exchange rates are driven, at the margin, by risk-neutral market participants who stand ready to take uncovered spot or forward positions whenever

the forward rate deviates from the expected future spot rate.

By manipulating condition (1), it is easily seen that CIP implies

$$\frac{f_t - s_t}{s_t} = \frac{1 + r_t^*}{1 + r_t} - 1 \tag{3}$$

Hence, as a first approximation (for values of $1 + r_t$ in the vicinity of 1):

$$r_t^* - r_t \approx (f_t - s_t)/s_t \tag{4}$$

In addition, when Jensen's inequality – that is, the difference between $E_t(1/s_{t+1})$ and $1/E_t(1/s_{t+1})$ – is ignored, the assumption of UIP can be approximated as

$$\begin{aligned} r_t^* - r &\approx E_t[(s_{t+1} - s_t)/s_t] \\ &= (E_t s_{t+1} - s_t)/s_t. \end{aligned} \tag{5}$$

The assumption of UIP adds an element of dynamics to the CIP condition by hypothesizing a relationship between the observed values of variables at time t and the value of the spot exchange rate that market participants expect at time t to prevail at time $t + 1$. As such, UIP has been embedded in many multi-period models of open economies. The CIP and UIP conditions can be written for any duration of the time period between t and $t + 1$. Thus, if the UIP assumption was valid at all horizons, the observed values of the spot exchange rate and the term structures of domestic and foreign interest rates could be used to infer the expected future time path of the spot exchange rate (Porter 1971).

In addition to playing an important role in the development of multi-period models of open economies, the UIP condition has been a central focal point in the policy debate over the effectiveness of official intervention in exchange markets (Henderson and Sampson 1983). To the extent that UIP was valid at short time horizons, official intervention could not succeed in changing the spot exchange rate relative to the expected future spot rate unless the authorities chose to allow interest rates to change. In this sense, exchange market intervention could not be viewed as providing the authorities with an effective policy instrument in addition to interest rates. Thus, the case for intervention has been considered by some to depend on whether the empirical evidence rejects UIP.

Empirical Evidence

The theory leading to the CIP condition – and hence also to the UIP assumption – abstracts entirely from any credit risks, capital controls, or explicit taxes on domestic and foreign currency investments. Keynes (1923, pp. 126–127) was well aware that investor choices between foreign and domestic assets do not depend on interest rates and exchange rates alone:

> . . . the various uncertainties of financial and political risk . . . introduce a further element which sometimes quite transcends the factor of relative interest. The possibility of financial trouble or political disturbance, and the quite appreciable probability of a moratorium in the event of any difficulties arising, or of the sudden introduction of exchange regulations which would interfere with the movement of balances out of the country, and even sometimes the contingency of a drastic demonetization, – all these factors deter . . . [market participants], even when the exchange risk proper is eliminated, from maintaining large . . . balances at certain foreign centres.

In those circumstances where it is valid to abstract from the types of considerations cited by Keynes, the CIP condition has been generally confirmed. As one source of evidence, interviews at large banks have established that the CIP condition is used as a formula for determining the exchange rates and interest rates at which trading is actually conducted. Foreign exchange traders use Eurocurrency interest rate differentials to determine the forward exchange rates (in relation to spot rates) that they quote to customers, while traders in Eurocurrency deposits use the spreads between forward and spot exchange rates to set the spreads between the interest rates that their banks offer on domestic and foreign currency deposits (Herring and Marston 1976; Levich 1985). As additional evidence, Taylor (1989) has constructed a database of the bid and offer rates quoted contemporaneously for exchange rates and

interest rates by foreign exchange and money market brokers, as recorded on the 'pad' of the chief dealer at the Bank of England. The data include observations on one- two-, three-, six-, and twelve-month maturities during selected intervals between 1967 and 1987. Taylor's study found no evidence of unexploited profit opportunities during relatively calm periods in foreign exchange and money markets, although potentially exploitable profitable arbitrage opportunities did 'occasionally occur' during periods of market turbulence, where the frequency, size and persistence of such opportunities were positively related to length of maturity. Consistently, in circumstances when it is not valid to abstract from capital controls and risks, empirical research has confirmed that deviations from CIP can be related systematically to the effective taxes imposed by capital controls and to non-currency-specific risk premiums associated with prospective controls (Dooley and Isard 1980).

The UIP assumption is more difficult to test than the CIP condition, since market expectations of future exchange rates are not directly observable. Accordingly, UIP has generally been tested jointly with the assumption that exchange market participants form rational expectations, such that future realizations of the exchange rate will equal the value expected at time t plus an error term that is uncorrelated with all information known at time t. Together the two assumptions imply that

$$s_{t+1} = f_t + u_{t+1} \tag{6}$$

and hence

$$s_{t+1} = s_t = r_t - r_t^* + u_{t+1} \tag{7}$$

where u represents a prediction error. This has led economists to assess the UIP assumption empirically by estimating the values of the a and b coefficients in the specification forms

$$s_{t+1} = a_0 + a_1 f_t + u_{t+1} \tag{8}$$

and

$$s_{t+1} - s_t = b_0 + b_1\left(r_t - r_t^*\right) + u_{t+1} \tag{9}$$

where it is assumed that the error terms have zero means and are serially uncorrelated.

Empirical assessments of UIP as a framework for predicting the future spot exchange rate have distinguished two issues: the size of the prediction errors, and the question of whether the predictions are systematically biased. On the first issue, it has become widely acknowledged that interest differentials explain only a small proportion of subsequent changes in exchange rates (Isard 1978; Mussa 1979; Frenkel 1981). This finding has been generally interpreted as implying that observed changes in exchange rates are predominantly the result of unexpected information or 'news' about economic developments, policies or other relevant factors.

The issue of whether predictions are systematically biased can be assessed by testing the hypothesis of unbiasedness – namely, that $(a_0, a_1) = (0, 1)$ in Eq. (8) or $(b_0, b_1) = (0, 1)$ in Eq. (9). Notably, the test that the slope coefficient is unity receives strong support from studies based on (8) but is soundly rejected by studies based on (9) – at least for prediction horizons of a year or less. However, the apparent conflict between the two sets of regression evidence has been resolved in favour of the latter finding, as it is now accepted that (8) is not a legitimate regression equation (Meese 1989). The explanation is based on the fact that the sample variances of the spot rate and forward rate are essentially equal.

Although the empirical evidence strongly rejects the unbiasedness hypothesis at prediction horizons of up to 1 year, the evidence is much more favourable to unbiasedness at horizons of 5–20 years. In particular, when data for industrial countries are pooled, and when annual exchange rate changes and interest differentials (for each country relative to a numeraire country) are averaged over non-overlapping 5- to 20-year periods, the slope coefficients in Eq. (9) become insignificantly different from unity (Flood and Taylor 1997, who note that the average one-year change over n years is equivalent to the change over n years multiplied by a scale factor; see also Chinn and Meredith 2004).

Does Prediction Bias Refute the UIP Assumption?

Economists have not resolved how to interpret the strong rejection of the unbiasedness hypothesis at short prediction horizons. Several possible explanations have been suggested, with different implications for UIP.

One interpretation rejects the UIP hypothesis but not the rational expectations assumption. According to this view, the finding of systematic prediction bias suggests that market participants are risk averse and require risk premiums to hold uncovered foreign currency positions. The prediction bias is thus perceived as an omitted variable problem that can be addressed, in concept, by extending the righthand side of Eq. (9) to include an expression for the risk premium. A second interpretation of prediction bias abandons the assumption that market participants are fully rational.

Other possible explanations do not require rejection of either UIP or the rational expectations hypothesis. These include explanations based on the 'peso problem', simultaneity bias, incomplete information with rational learning, and self-fulfilling prophecies or rational 'bubbles'.

The suggestion that prediction bias reflects a 'peso problem' is generally attributed to Rogoff (1980) and Krasker (1980), who drew attention to an episode in which the Mexican peso sold at a forward discount for a prolonged period prior to its widely anticipated devaluation in 1976. Although market expectations eventually proved correct and may well have been rational *ex ante*, the fact that the devaluation did not occur immediately after it became anticipated made the forward rate a biased predictor over finite data samples that included the pre-devaluation period. The general point is that, even if market participants are risk neutral and form rational expectations, the forward rate can be biased as a predictor of the future spot rate – and the interest rate differential biased as a predictor of the change in the spot rate – whenever market participants repeatedly expect the spot rate to change in response to a policy action or some other event that fails to materialize over a relatively long series of observations.

The suggestion that rejection of the unbiasedness hypothesis reflects simultaneity bias was alluded to by Isard (1988) and later emphasized by McCallum (1994). In particular, given that the monetary authorities in most countries rely on a short-term interest rate as a policy instrument that they are prepared to adjust, *inter alia*, in response to undesired exchange rate movements, the estimates of b1 may be biased by the failure to estimate (9) simultaneously with a second relationship between the interest rate differential and the change in the exchange rate.

As suggested by Lewis (1988, 1989), prediction bias can also emerge under UIP and rational expectations if market participants lack complete information but engage in a process of rational learning. This explanation is analogous to the peso problem in so far as it provides an interpretation in which market participants are risk neutral and fully rational but prone to make repeated mistakes.

Yet another possibility consistent with UIP is the conjecture that prediction bias arises from the self-fulfilling prophecies of rational, risk-neutral market participants. Such prophecies, which are often referred to as 'rational bubbles', have received attention as logical possibilities; but few economists, if any, consider them to have much plausibility as empirical phenomena (Mussa 1990).

Where Things Stand

Because the validity of the UIP hypothesis cannot be tested directly and is not resolved by the rejection of the unbiasedness hypothesis, economists have resorted to indirect tests as a means of obtaining suggestive evidence. In particular, survey data on exchange rate expectations have been collected by several different sources since the early 1980s, and a number of studies have shown that exchange rate expectations, as measured by the average forecasts of sample respondents, deviate considerably from prevailing forward exchange rates (Frankel and Froot 1987; Takagi 1991; Chinn and Frenkel 2002). To the extent that survey measures of average

expectations are meaningful, this would appear to be strong evidence against UIP.

That said, it also needs to be recognized that intertemporal models of open-economy macroeconomics require equations that link current spot exchange rates to expected future exchange rates. Thus, on pragmatic grounds, the case for abandoning the UIP hypothesis depends on how well economists can model the deviation from UIP – namely, the difference between the forward exchange rate and the expected future spot rate, which is generally referred to as the exchange risk premium.

Behavioural hypotheses about the exchange risk premium can be tested by embedding them in models of observable exchange rates. The first conceptual models of the exchange risk premium were based on a portfolio balance framework in which financial claims were distinguished by currencies of denomination but not by the countries obligated to meet the claims (see, for example, Dooley and Isard 1983). Empirical tests of this class of portfolio balance model have explained at most a small portion of the variation over time in the exchange risk premium (Tryon 1983; Boughton 1987). More sophisticated behavioural hypotheses have recognized – in the spirit of the quotation above from Keynes – that exchange risks and credit risks are interrelated, and that the magnitudes of these risks reflect the relative macroeconomic and political conditions, prospects, and uncertainties of the countries that have issued the portfolio claims (Dooley and Isard 1983; Isard 1988). While casual evidence suggests that this type of hypothesis is broadly capable of explaining the empirical behaviour of exchange rates (Dooley and Isard 1991), formal empirical tests that capture the many factors contributing to exchange rate risk are difficult to design, and economists have not yet provided a well-specified replacement for the UIP assumption.

Accordingly, many intertemporal open-economy macroeconomic models continue to impose the UIP assumption – or the assumption of UIP adjusted by an exogenous exchange risk premium that provides a mechanism for analyzing the effects of 'exogenous' changes in risk perceptions or asset preferences. However, consistent with the evidence that rejects the unbiasedness hypothesis, it has proved difficult to mimic the observed behaviour of key macroeconomic variables with models that impose the UIP assumption and also treat exchange rate expectations as fully model-consistent. Thus, models that impose the UIP assumption tend to treat exchange rate expectations as not completely rational. One fairly common practice, for example, is to treat exchange rate expectations (and inflation expectations) as having both forward-looking (model-consistent) and backward-looking components.

Quite apart from ongoing debates over the validity of the UIP assumption as an *ex ante* hypothesis, and the usefulness of incorporating the UIP assumption into macroeconomic models, there is abundant evidence, as noted above, that the changes in spot exchange rates that are expected *ex ante* are generally dominated by unexpected changes. Thus, regardless of the usefulness of UIP as an *ex ante* hypothesis for macroeconomic modelling, it is quite clear that UIP by itself provides a very inaccurate framework for predicting the changes in exchange rates that are observed *ex post*.

See Also

► Exchange Rate Dynamics

Bibliography

Boughton, J. 1987. Tests of the performance of reduced-form exchange rate models. *Journal of International Economics* 23: 41–56.

Chinn, M., and J. Frenkel. 2002. Survey data on exchange rate expectations: More currencies, more horizons, more tests. In *Monetary policy, capital flows, and financial market developments in an era of financial globalization: Essays in honour of Max Fry*, ed. W. Allen and D. Dickinson. London: Routledge.

Chinn, M., and G. Meredith. 2004. Monetary policy and long horizon uncovered interest parity. *IMF Staff Papers* 51: 409–430.

This article draws extensively on Isard (1991, 1995). The views expressed are those of the author and do not necessarily reflect those of the International Monetary Fund.

Dooley, M., and P. Isard. 1980. Capital controls, political risk and deviations from interest-rate parity. *Journal of Political Economy* 88: 370–384.

Dooley, M., and P. Isard. 1983. The portfolio-balance model of exchange rates and some structural estimates of the risk premium. *IMF Staff Papers* 30: 683–702.

Dooley, M., and P. Isard. 1991. A note on fiscal policy, investment location decisions, and exchange rates. *Journal of International Money and Finance* 10: 161–168.

Einzig, P. 1962. *The history of foreign exchange*. London: Macmillan.

Flood, R., and M. Taylor. 1997. Exchange rate economics: What's wrong with the conventional macro approach? In *The microstructure of foreign exchange markets*, ed. J. Frankel, G. Galli, and A. Giovannini. Chicago: University of Chicago Press.

Frankel, J., and K. Froot. 1987. Using survey data to test standard propositions regarding exchange rate expectations. *American Economic Review* 77: 133–153.

Frenkel, J. 1981. Flexible exchange rates, prices and the role of 'news': Lessons from the 1970s. *Journal of Political Economy* 89: 665–705.

Henderson, D., and S. Sampson. 1983. Intervention in foreign exchange markets: A summary of ten staff studies. *Federal Reserve Bulletin* 69: 830–836.

Herring, R., and R. Marston. 1976. The forward market and interest rates in the Eurocurrency and national money markets. In *Eurocurrencies and the international monetary system*, ed. C. Stern, J. Makin, and D. Logue. Washington, DC: American Enterprise Institute.

Isard, P. 1978. *Exchange-rate determination: A survey of popular views and recent models*, Princeton studies in international finance. Vol. 42. Princeton: International Finance Section, Department of Economics, Princeton University.

Isard, P. 1988. Exchange rate modeling: An assessment of alternative approaches. In *Empirical Macroeconomics for interdependent economies*, ed. R. Bryant et al. Washington, DC: Brookings Institution.

Isard, P. 1991. Uncovered interest parity. In *The new Palgrave dictionary of money and finance*, 1st ed., ed. P. Newman, M. Milgate, and J. Eatwell. Basingstoke: Palgrave Macmillan. Also issued as Working paper WP/91/51, Washington, DC: International Monetary Fund, 1991.

Isard, P. 1995. *Exchange rate economics*. Cambridge: Cambridge University Press.

Keynes, J.M. 1923. *A tract on monetary reform*. London: Macmillan.

Krasker, W. 1980. The peso problem in testing the efficiency of forward exchange markets. *Journal of Monetary Economics* 6: 269–276.

Levich, R. 1985. Empirical studies of exchange rates: Price behavior, rate determination and market efficiency. In *Handbook of International Economics*, ed. R. Jones and P. Kenen, Vol. 2. Amsterdam: North-Holland.

Lewis, K. 1988. The persistence of the 'peso problem' when policy is noisy. *Journal of International Money and Finance* 7: 5–21.

Lewis, K. 1989. Changing beliefs and systematic rational forecast errors with evidence from foreign exchange. *American Economic Review* 79: 621–636.

Lotz, W. 1889. Die Währungsfrage in Österreich-Ungarn. *Schmollers Jahrbuch* 13: 34–35.

McCallum, B. 1994. A reconsideration of the uncovered interest parity relationship. *Journal of Monetary Economics* 33: 105–132.

Meese, R. 1989. Empirical assessment of foreign currency risk premiums. In *Financial risk: Theory, evidence and implications*, ed. C. Stone. Boston: Kluwer.

Mussa, M. 1979. Empirical regularities in the behavior of exchange rates and theories of the foreign exchange market. *Carnegie-Rochester Conference Series on Public Policy* 11: 9–57.

Mussa, M. 1990. *Exchange rates in theory and reality*, Essays in international finance. Vol. 179. Princeton: International Finance Section, Department of Economics, Princeton University.

Porter, M. 1971. A theoretical and empirical framework for analyzing the term structure of exchange rate expectations. *IMF Staff Papers* 18: 613–642.

Rogoff, K. 1980. Chapter 1: An empirical investigation of the martingale property of foreign exchange futures prices. *Essays on expectations and exchange rate volatility*. Doctoral dissertation, Cambridge, MA: MIT. Online. Available at http://post.economics.harvard.edu/faculty/rogoff/papers/ChapterOne.pdf. Accessed 4 May 2006.

Takagi, S. 1991. Exchange rate expectations: A survey of survey studies. *IMF Staff Papers* 38: 156–183.

Taylor, M. 1989. Covered interest arbitrage and market turbulence. *Economic Journal* 99: 376–391.

Tryon, R. 1983. *Small empirical models of exchange market intervention: A review of the literature*, Staff studies. Vol. 134. Washington, DC: Board of Governors of the Federal Reserve System.

Underconsumptionism

Michael Schneider

Keywords

Acceleration principle; Aggregate demand; Baran, P. A.; Catchings, W.; Clark, J. B.; Disproportionate production; Emmanuel, A.; Harrod–Domar growth model; Hoarding; Hobson, J. A.; Imperialism; Kautsky, K.; Keynesianism; Lauderdale, Eighth Earl of;

Lenin, V. I.; Luxemburg, R.; Malthus, T. R.; Marx, K. H.; Redistribution of income; Rodbertus, J. K.; Saving equals investment; Sismondi, J. C. L. S. de; Sweezy, P.M.; Underconsumption; Underconsumptionism; Wages fund

JEL Classifications
B1

'Underconsumption' is the label given to theories which attribute the failure of the total output of an economy to continue to be sold at its cost of production (including normal profit) to too low a ratio of consumption to output. According to underconsumption theories, such deficient consumption leads either to goods being able to be sold only at below-normal rates of profit, or to goods not being able to be sold at all. These effects are seen as leading in turn to cutbacks in production and increases in unemployment. Underconsumption theories are thus amongst those which seek to explain cyclical or secular declines in the rate of economic growth.

Where underconsumption exists in the sense that the ratio of consumption to output is below the optimum level, it follows that the ratio of 'unconsumed' output to total output must be too high. For the period in which underconsumption breaks out, underconsumptionists in general both identify this 'unconsumed' output with saving, and equate saving with investment. Thus Haberler, to whose 1937 analysis of underconsumption and related theories the reader should turn for the best extended treatment of the subject, wrote that in 'its best-reasoned form . . ., the underconsumption theory uses "under-consumption" to mean "over-saving"', and that in 'the under-consumption or over-saving theory . . . savings are, as a rule, invested . . .' (Haberler 1937, pp. 115 and 117).

While the theories advanced by underconsumptionists overlap with some other macroeconomic theories in certain ways, their basic characteristics make them distinct in other respects. Underconsumption theories share with Keynesian theories, for example, the characteristic that they are 'demand-side' (as opposed to 'supply side') theories. However, there is a fundamental difference between the two, in that Keynesian theories attribute the failure of total output to reach the full employment level to a deficiency of *aggregate* demand. The two types of theory consequently have different implications. As Robbins succinctly put it, with reference to the underconsumptionist J.A. Hobson, for 'Mr Keynes, one way out of the slump would be a revival of investment; for Mr Hobson, this would simply make matters worse' (Robbins 1932, p. 420). A further difference between the two lies in the fact that, by contrast with Keynesian theories, hoarding plays no part in underconsumption theories. Underconsumptionists in general confine their analyses to the real sector of the economy, and where monetary factors are discussed at all they are treated as secondary.

There are also some connections between underconsumption theories and the accelerator theory of investment. As Haberler pointed out, the acceleration principle can be used 'in support of a special type of the under-consumption theory of the business cycle' (Haberler 1937, p. 30). More importantly, a variant of the principle can be seen as underlying all underconsumption theories. When first expounded by J.M. Clark, the acceleration principle was used to explain the level of activity in the investment goods sector of an economy by changes in the demand for finished goods. In essence, underconsumptionists base their theories on the idea that changes in the demand for consumption goods determine the future level of activity in the investment goods sector. By this means they draw the conclusion that the level of activity in the economy as a whole is wholly determined by consumption demand.

In a review of Harrod's *Towards a Dynamic Economics,* Joan Robinson suggested that 'Mr Harrod's analysis provides the missing link between Keynes and Hobson' (Robinson 1949, p. 80). The resemblance of underconsumption theories to growth models of the Harrod–Domar type is in fact greater than their resemblance to Keynesian theories. As Domar pointed out in the *American Economic Review* article (1947) expounding his growth model, he shared with Hobson a concern with the capacity-creating effect of investment, a

question which Keynes hardly touched on in the *General Theory*. The essential features of underconsumption theories can in fact be captured by a growth model of the Harrod–Domar type, in which however the driving force is provided not by the rate of growth of investment but by the rate of growth of consumption. Such a model is particularly appropriate in the case of those theories which treat underconsumption as a secular rather than a cyclical phenomenon.

There are connections too between underconsumption theories and explanations of 'economic crises' in terms of 'disproportionate production', to use Marx's terminology. By 'disproportionate production' Marx meant an allocation of labour time between sectors or industries other than that required to satisfy social need as reflected in demand. Now underconsumption involves an allocation of too few resources to the consumption goods sector and too many resources to the investment goods sector. But as Haberler pointed out, such 'vertical disproportion' should be distinguished from 'horizontal disproportion'. And unlike cases of horizontal disproportion (if optimal stock levels are ignored), vertical disproportion, involving industries not equidistant from consumption goods industries, cannot be rectified immediately by a return to 'proportionate' production. For the excessive production of investment goods consequent upon underconsumption leaves a legacy in the form of excessive productive capacity. Underconsumption theories thus should be distinguished from the more general category of 'disproportionality' theories; the disproportionality element they incorporate is specific and has distinctive consequences.

Over-investment theories provide a different example of vertical disproportion. As they are defined by Haberler, over-investment theories offer an explanation of the excessive aggregate demand characteristic of an economy during the upswing of a trade cycle. Therefore they also belong to a different category from underconsumption theories, even though the deficiency of consumption characteristic of the latter is accompanied by excessive investment.

Despite their basic similarities, underconsumption theories differ as to the cause of, and hence remedies for, underconsumption. A view to be found in the writings of some underconsumptionists, especially the less well-known ones, is that underconsumption is due to total purchasing power falling short of the value of output. Since all the value of output accrues to the owner of one factor of production or another, this proposition as it stands cannot be sustained. This view as to the cause of underconsumption should not be confused, however, with a superficially similar view relating to its effects, which is at least implicit in all underconsumptionist thinking. This is the idea that income is generated not by production but by purchases of what is produced. In underconsumptionist writings, by contrast for example with Keynesian writings, income may fall short of the value of output.

How this may be so is perhaps best seen in terms of period analysis. An outbreak of underconsumption will lead in the first period to excessive saving accompanied by excessive investment. In the second period the resulting additional capacity will be used, and unless there is an increase in consumption the level of output will exceed the demand for it; hence in this period the income generated by purchases will fall short of the value of output, while at the same time saving, if it is defined as that part of income (as opposed to output) not consumed, will just match investment demand. The deficient demand in the second period will lead in the third period to actual output falling short of potential output, that is to excess capacity, with saving however continuing to equal investment.

One underconsumptionist who clearly did not attribute underconsumption to lack of purchasing power is Malthus. In his correspondence with Ricardo, Malthus instead took the position that 'a nation must certainly have the power of purchasing all that it produces, but I can easily conceive it not to have the will' (Sraffa 1952, p. 132). Like some other underconsumptionists, notably Sismondi and Hobson, Malthus believed that one cause of underconsumption is to be found in the limited capacity of human beings to expand their wants, at least in the short run. It was Malthus's view that men have a tendency towards indolence once their needs for necessaries are

satisfied. If in the face of such limited growth in human wants capital accumulation continues apace, the resulting increase in productive capacity will fail to be matched by an equal increase in consumer demand. The remedy for this state of affairs, suggested Malthus, is an increase in commerce, both domestic and foreign, so as to stimulate tastes by exposing the population to new products.

Most commonly, however, underconsumptionists find the cause of underconsumption in a maldistribution of income. The underlying argument is simple. If different economic classes have different propensities to consume, the distribution of an excessive share of income to classes with a relatively low propensity to consume will result in underconsumption. Underconsumptionists agree that a remedy cannot be found by a redistribution of income towards the capitalist class, which they see as having a relatively high propensity to save. They differ, however, on the question of the class to which income should in cases of underconsumption be redistributed. The earliest underconsumptionists ruled out a redistribution of income towards workers, perhaps partly because it was incompatible with their adherence to the wages fund doctrine, according to which total wages are fixed by the capital set aside in advance to pay them. They advocated rather a redistribution of income towards landlords. The first underconsumptionist to advocate a redistribution of income towards workers was Sismondi, whose example was followed by most later underconsumptionists.

Underconsumption theories were first put forward in the 19th century. While some 17th- and 18th-century writers, most notably Mandeville and the Physiocrats, advocated an increase in expenditure on consumption goods, none of them linked this with a corresponding reduction in investment. Therefore although they may be seen as predecessors of Keynes, they should not be classified as underconsumptionists. The first to advance an underconsumption theory in the sense outlined above was Lauderdale, in *An Inquiry into the Nature and Origin of Public Wealth* (1804). Perhaps the best known of the subsequent underconsumptionists are Malthus, Sismondi, Rodbertus, Hobson and Rosa Luxemburg. For a fuller (and sometimes different) account of the theories of these and other underconsumptionists than is possible here, the reader should turn to Bleaney (1976) or Nemmers (1956). Further, additional light has been shed on the overall nature of underconsumptionist theories by the several attempts that have been made to express the theory put forward by Malthus in the form of a model, notably by Eagly (1974), Eltis (1980) and Costabile and Rowthorn (1985).

While some underconsumption theories were largely prompted by current or expected economic events, in other cases the inspiration was mainly intellectual. Both factors seem to have been important in the case of Lauderdale, the earliest underconsumptionist. Lauderdale was in part reacting against the praise of parsimony by Adam Smith, but he was also alarmed at the prospect of the British government using its revenue after the end of the Napoleonic wars for the purpose of capital accumulation in place of wartime consumption. More generally, as a precaution against underconsumption, Lauderdale advocated a lessening of the current inequality of wealth, as Malthus was also to do. By contrast Spence, in *Britain Independent of Commerce* (1807), developed an underconsumption theory on the basis of Physiocratic ideas. His solution for underconsumption was encouragement of consumption by landlords, so as to restore the income of the manufacturing class to its former level.

His correspondence with Ricardo shows that Malthus had developed underconsumptionist views by 1814. This fact is doubly significant. It proves both that Malthus's underconsumptionism preceded the depressed economic conditions which followed the ending of the Napoleonic wars in 1815, rather than being a response to them, and that Marx's charge that Malthus plagiarized Sismondi is unfounded. The underconsumptionist elements in Malthus's thinking are to be found not only in his correspondence with Ricardo, but also in his *Principles of Political Economy Considered with a View to their Practical Application* (1820). The latter had an influence on the underconsumption theory put forward in Chalmers' *Political Economy* (1832).

It may also have provided a stimulus for the underconsumption theory advanced in a pamphlet entitled *Considerations on the Accumulation of Capital* (1822). Published anonymously, this pamphlet was written by Cazenove, the friend of Malthus who was later to edit (also anonymously) the second edition of Malthus's *Principles.*

Like Lauderdale, Sismondi reacted against Adam Smith's views on parsimony, and like Malthus he had become an underconsumptionist by the end of the Napoleonic wars, as is evidenced by the material contained in the article entitled 'Political Economy' which Sismondi wrote in 1815 for Brewster's *Edinburgh Encyclopaedia.* A complete account of Sismondi's underconsumption theory is only to be found, however, in his *Nouveaux principes d'économie politique* (1819). Here Sismondi argued that where producers supply a large anonymous market, competition for profits leads each of them on the one hand to overestimate the demand for the commodity he produces and overaccumulate capital accordingly, and on the other hand so to depress wages that they grow at a slower rate than profits. Sismondi's remedies for underconsumption include organization of industry on a local basis, and a redistribution of income towards wages.

For a discussion of possible sources of the underconsumptionist elements in the writings of Robert Owen and the Ricardian Socialists the reader is referred to King (1981). A more comprehensive underconsumption theory than in those writings is to be found in Rodbertus's 'second letter' to von Kirchmann, published in 1850–51. Rodbertus was reacting against the ideas of Jean-Baptiste Say and his followers. His own view was that in a laissez-faire economy underconsumption must inevitably emerge and worsen, because 'natural' laws will ensure that an ever increasing productivity of labour will be accompanied by an ever decreasing share of income going to wages. His remedy was 'rational' intervention in the economy to counteract these 'natural' laws.

The emphasis in Marx's economic theory on the necessity in a capitalist economy for value not only to be generated in production but also realized by sale makes that theory well adapted to use to the development of an underconsumption theory. Marx himself gave substantial praise to Sismondi for his exposition of such a theory, and there are several passages in Marx's own writings which put forward an underconsumptionist view. On the other hand, there is a well-known passage in volume 2 of *Capital* which condemns underconsumption theories in no uncertain terms, and in any case there are other elements in his economic theory which are so much more important to Marx that he is not usually classified as an underconsumptionist. Many, though by no means all, of his followers have in fact condemned underconsumption theories. Examples of such condemnation are to be found in some of Lenin's writings, notably his pamphlet entitled *A Characterisation of Economic Romanticism (Sismondi and our Native Sismondists),* written in 1897. This pamphlet was particularly directed at the underconsumptionist views of the Russian 'Populists', or 'Narodniks', who had argued that capitalism could not survive in Russia without the consumer markets provided by its then dwindling peasant economy. It was Lenin's view that for the development of capitalism expansion of the market for investment goods is more important than expansion of the market for consumption goods.

Amongst Marx's earlier followers, those who most strongly supported the underconsumptionist element in Marx's thinking were Kautsky and Rosa Luxemburg. Rosa Luxemburg's main arguments were set out in *The Accumulation of Capital* (1913). Contrasting the over-growing generation of value in a capitalist economy with the inability of workers and unwillingness of capitalists to realize that value by increasing their consumption, she crossed swords with Tugan Baranovski, who had argued that capitalists 'see to it that ever more machines are built for the sake of building – with their help – ever more machines' (Luxemburg 1913, p. 335). Rosa Luxemburg took the same view as that advanced by J.B. Clark, in his introduction to the English translation of Rodbertus's 'second letter' to von Kirchmann, namely that 'this case presents no glut: but it is an unreal case' (Rodbertus 1898, p. 15). She concluded that because it was inevitably faced by increasing underconsumption, a capitalist economy could only survive as long as it

was able to dispose of its surplus to non-capitalist consumers, either at home or abroad, the latter accounting in her view for policies of imperialism. Apart from Rosa Luxemburg, others who have both drawn on Marx's ideas and made use of underconsumption theory include Sweezy in *The Theory of Capitalist Development* (1942), Baran and Sweezy in *Monopoly Capital* (1966) and Emmanuel in *Unequal Exchange* (1969).

A causal connection between underconsumption and policies of imperialism was also argued to exist by the non-Marxist writer J.A. Hobson, in *Imperialism: A Study* (1902). Jointly with A.F. Mummery, Hobson had reacted to the depression in trade in the 1880s by putting forward an underconsumption theory in *The Physiology of Industry* (1889), which was the first underconsumptionist work actually to use the term 'underconsumption'. In this book Mummery and Hobson argued that the sole source of demand for investment goods is demand for consumption goods. From this they drew the conclusion, as Malthus had done, that there exists an optimum ratio between saving (investment) and spending (consumption). Like Sismondi, they stressed the role of competition in causing supply to exceed demand. They went beyond the earlier underconsumptionists, however, in specifically arguing that neither a fall in the rate of interest nor a fall in the price level could remedy a state of depression brought about by underconsumption. Hobson's subsequent restatement of this theory, with various amplifications, made him the most influential 20th-century exponent of underconsumption theories.

In *The Physiology of Industry* Mummery and Hobson drew the policy conclusion that 'where Under-consumption exists, Savings should be taxed' (Mummery and Hobson 1889, p. 205). In his later works, however, from *The Problem of the Unemployed* (1896) on, Hobson laid most stress on a redistribution of income from what he called 'unearned income' (income unrelated to effort) to wages as the main remedy for underconsumption. The most comprehensive expositions of Hobson's underconsumption theory are to be found in *The Industrial System* (1909), which is characterized by a more extensive treatment of underconsumption in a growing economy, *The Economics of Unemployment* (1922), and *Rationalisation and Unemployment* (1930).

Other 20th-century exponents of underconsumption theories include Foster and Catchings, in a number of jointly written books. The theories of Major Douglas, however, with their lack of reference to over-investment and their emphasis on the role of money and credit, do not fit well into the underconsumptionist category.

Underconsumption theories have never been acceptable to orthodox economists, perhaps partly because underconsumptionists in general have lacked rigour in the exposition of their ideas, and partly because underconsumption theories have been seen as a threat to the saving necessary for economic growth in particular, and to capitalism in general. They have also attracted less attention since 1936 than before, because Keynes's *General Theory* satisfied the needs of many of those whose intuitions led them to seek a 'demand-side' explanation of economic depression. However, underconsumption theories can be argued still to provide a useful supplement to Keynesian theories, as a reminder that there is a limit to the extent to which employment can be increased by increases in investment alone. There is perhaps some recognition of this in the distinction which is now commonly made as to whether the current need is for an 'investment-led' or a 'consumption-led' recovery.

See Also

- ▶ Hobson, John Atkinson (1858–1940)
- ▶ Keynes, John Maynard (1883–1946)

Bibliography

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

Bleaney, M. 1976. *Underconsumption theories: A history and critical analysis*. London: Lawrence & Wishart.

Cazenove, J. 1822. *Considerations on the accumulation of capital and its effects on profits and on exchangeable value*. London: J.M. Richardson. Published anonymously.

Chalmers, T. 1832. *On political economy, in connexion with the moral state and moral prospects of society*. Glasgow: William Collins.

Costabile, L., and R.E. Rowthorn. 1985. Malthus's theory of wages and growth. *Economic Journal* 95: 418–437.
Domar, E.D. 1947. Expansion and employment. *American Economic Review* 37: 34–55.
Eagly, R.V. 1974. *The structure of classical economic theory*. Oxford: Oxford University Press.
Eltis, W.A. 1980. Malthus's theory of effective demand and growth. *Oxford Economic Papers* 32: 19–56.
Emmanuel, A. 1969. *Unequal exchange: A study of the imperialism of trade*. New York: Monthly Review Press. 1972.
Haberler, G. 1937. *Prosperity and depression*. Geneva: League of Nations.
Hobson, J.A. 1896. *The problem of the unemployed*. London: Methuen.
Hobson, J.A. 1902. *Imperialism: A study*. London: Nisbet.
Hobson, J.A. 1909. *The industrial system*. New York: Longmans.
Hobson, J.A. 1922. *The economics of unemployment*. London: Allen & Unwin.
Hobson, J.A. 1930. *Rationalisation and unemployment*. London: Allen & Unwin.
Keynes, J.M. 1936. *The general theory of employment, interest and money*. London: Macmillan.
King, J.E. 1981. Perish commerce! Free trade and underconsumption in early British radical economics. *Australian Economic Papers* 20 (37): 235–257.
Lauderdale, Eighth Earl of. 1804. *An inquiry into the nature and origin of public wealth*. Edinburgh. Reprinted, New York: Augustus M. Kelley, 1962.
Lenin, V.I. 1897. A characterisation of economic romanticism (Sismondi and our native Sismondists). In *Collected works*, ed. V.I. Lenin, vol. 2. Moscow: Foreign Languages Publishing House. 1962.
Luxemburg, R. 1913. *The accumulation of capital*. Trans. A. Schwarzschild, with an introduction by J. Robinson. London: Routledge & Kegan Paul, 1951.
Malthus, T.R. 1820. *Principles of political economy considered with a view to their practical application*. London: Murray.
Malthus, T.R. 1836. *Principles of political economy considered with a view to their practical application*. 2nd ed. London: William Pickering.
Marx, K. 1885. *Capital*, vol. 2. Moscow: Foreign Languages Publishing House, 1957.
Mummery, A.F., and J.A. Hobson. 1889. *The physiology of industry*. London: Murray.
Nemmers, E.E. 1956. *Hobson and underconsumption*. Amsterdam: North-Holland.
Robbins, L. 1932. Consumption and the trade cycle. *Economica* 12: 413–430.
Robinson, J. 1949. Mr Harrod's dynamics. *Economic Journal* 59: 68–85.
Rodbertus, K. 1898. *Overproduction and crises*. London: Swan Sonnenschein.
Sismondi, J.C.L. 1815. *Political economy*. New York: Kelley. 1966.
Sismondi, J.C.L. 1819. *Nouveaux principles d'économie politique*. Paris: Delaunay.
Spence, W. 1807. Britain independent of commerce. In *Tracts on political economy*. London: Longman, Hurst, Orme & Brown. 1822.
Sraffa, P., ed. 1952. *Works and correspondence of David Ricardo*. Vol. 6. Cambridge: Cambridge University Press.
Sweezy, P.M. 1942. *The theory of capitalist development*. New York: Monthly Review Press.

Underemployment Equilibria

P. Jean-Jacques Herings

Abstract

The standard model of general equilibrium is extended by allowing for expectations about supply opportunities by households and firms. In this framework there is typically a 1-dimensional continuum of underemployment equilibria that range from equilibria with arbitrarily pessimistic expectations to equilibria with rather optimistic expectations. An example illustrates the model and highlights some features of underemployment equilibria. The multiplicity of equilibria has a natural interpretation as being the result of coordination failures. The results in this framework are compared with those of the fixprice literature. Extensions to a monetary economy are discussed.

Keywords

Agent optimization; Animal spirits; Arrow–Debreu model of general equilibrium; Cobb–Douglas functions; Competitive equilibrium; Coordination failures; Economics of general disequilibrium; Excess capacity; Excess demand; Fixprice models; Game theory; General equilibrium; General equilibrium models of coordination failures; Incomplete markets; Involuntary unemployment; Keynes, J.M.; Keynesianism; Market clearing; Neoclassical model; Non-market clearing prices; Pareto optimality; Path connectedness; Positive externalities; Price rigidities; Rational

expectations; Rationing; Seigniorage; Self-justifying expectations; Spillover effects; Strategic complementarities; Supply opportunities; Temporary equilibrium; Underemployment equilibria; Wage rigidities; Walras's law

JEL Classifications
C62; D51; E24; D5

Underemployment of resources refers to the situation where an increase in the resource utilization rate could lead to a Pareto improvement. Typical examples are involuntary unemployment and idle production capacities. There are two quite distinct views on the underemployment of resources. In the standard neoclassical world of the Arrow–Debreu model, underemployment of resources cannot occur. In the competitive equilibrium, involuntary unemployment does not exist, and production capacities are left idle when only such is Pareto optimal.

The Keynesian tradition, in contrast, builds on wage and price rigidities in its explanation of underemployment of resources. Indeed, Keynes's contribution has been reinterpreted by Clower (1965) and Barro and Grossman (1971) as the economics of general disequilibrium, in which price rigidities lead to quantity constraints for households and firms, which generally have spillover effects in other markets. This lead has been further developed in the fixprice literature, originating in the work of Bénassy (1975), Drèze (1975), and Younès (1975), and in general equilibrium theories on temporary equilibrium (see Grandmont 1977, for a survey).

Although the fixprice literature stresses wage and price rigidities, Keynes himself postulates that it is possible to encounter self-justifying expectations, beliefs which are individually rational but which may lead to socially irrational outcomes (Keynes 1936, ch. 12). We therefore would like to address the question whether underemployment of resources is possible when expectations are rational, agents optimize, and trade takes place at competitive prices. The underlying reason for underutilization of resources comes from coordination failures, self-justifying expectations which are individually rational but socially suboptimal.

In the literature, one may distinguish three classes of models with coordination failures. The first class consists of rather abstract game-theoretic models following the seminal work of Bryant (1983), see Cooper (1999), for a state-of-the art account. An important message coming from this stream of the literature is that coordination failures may occur when there are strategic complementarities and positive externalities. However, these models typically lack a coordinating role for the price mechanism.

Strategic models with a coordinating role for prices constitute the second class. Roberts (1987) presents a model that meets these criteria. It is a simple model of a closed economy that allows for coordination failures in a strategic setting. However, outcomes in the second class of models are often not robust to slightly different specifications of the model (Jones and Manuelli 1992).

The third class of models consists of general equilibrium models of coordination failures. We refer to Citanna et al. (2001) for the most general presentation of these ideas. The third class leads to results that are robust and general. The methodological assumptions are shared with those of the neoclassical model: agent optimization, market clearing, and rational expectations. Underemployment of resources occurs as a consequence of self-confirming, pessimistic expectations about supply opportunities.

Competitive Equilibria

Consider the classical general equilibrium model with H households, F firms and L commodities as described in Debreu (1959). A household h is characterized by its consumption set, for the sake of simplicity equal to $\mathbb{R}^L_+$, a utility function $u^h : \mathbb{R}^L_+ \to \mathbb{R}$, and initial endowments $e^h \in \mathbb{R}^L_+$. The feasible production plans of firm f are described by the production possibility set Y^f. If firm f chooses production plan y^f and the prices at which trade takes place are $p \in \mathbb{R}^L$, then the firm's profits equal $p \cdot y^f$. Household h receives a share θ^{fh} of the profits of firm f.

We assume both households and firms to be price takers. If trade takes place at prices p, then firm f faces the following profit maximization problem:

$$\max_{y^f \in Y^f} p \cdot y^f .$$

Under standard assumptions the firm's profit maximization problem is well-defined. For the remainder we assume the profit maximizing production bundle to be unique. This can also be shown to hold under standard assumptions, mainly requiring the strong assumption of decreasing returns to scale.

The utility maximization problem of household h reads as follows:

$$\max_{x^h \in \mathbb{R}^L_+} u^h\left(x^h\right) \text{s.t.} p \cdot x^h \leq w^h,$$

where w^h equals the value of the household h's initial endowments, $p \cdot e^h$, plus the household's share in the firms' profits, $\sum_f \theta^{fh} p \cdot y^{*f}$, with y^{*f} the profit maximizing production bundle chosen by firm f. Under standard assumptions, the maximization problem of the household has a unique solution x^{*h}.

Under the usual microeconomic methodological premises of agent optimization and market clearing, together with rational expectations, one defines a *competitive equilibrium* as a price system p^* and an allocation $(x^*; y^*) = (x^{*1},...,x^{*H}, y^{*1},..., y^{*F})$ such that at prices p^* households maximize utility by choosing the consumption bundle x^{*h} and firms maximize profits by choosing the production plan y^{*f}.

Rationing

The puzzle as to how competitive equilibria are achieved in real-world economies remains substantial. First, it is well-known that price adjustment processes need not converge to an equilibrium (Debreu 1974; Saari and Simon 1978; Saari 1985). Blad (1978) stresses that convergence, even if it takes place, can take quite some time. Second, in many situations some agents, or coalitions of agents, set prices at levels not compatible with competitive equilibrium. Drèze (1989) models unions that set wages above competitive levels. Herings (1997) and Tuinstra (2000) show that political interference in the market mechanism can be rational from a partisan point of view and might be responsible for sustained deviations from prices that clear markets. Third, Drèze and Gollier (1993), Drèze (1997), and Herings and Polemarchakis (2005) argue that certain price rigidities are a welfare-improving response to market incompleteness. This argument is particularly valid for the two forms of underemployment most frequently encountered, namely, unemployed labour and excess capacities, two examples of commodities for which future markets are hardly developed.

For the moment we maintain the assumption that trade takes place at given prices p. Here, p may or may not be competitive. We are not focusing on a specific theory of non-market clearing prices, but rather are interested in knowing how agents make decisions given that trade takes place at prices p. We deviate from the standard framework and assume that it is not common knowledge whether these prices are competitive or not. Even when all agents know whether prices are competitive or not, it is not necessarily the case that all agents know that all agents know whether prices are competitive or not, and it is even less likely that all agents know that all other agents know that all other agents know whether prices are competitive or not, and so on. Common knowledge of whether prices are competitive requires structural knowledge about the economy, very much at odds with the standard general equilibrium paradigm whereby in a decentralized economy agents only have to maximize utility given the prices that are quoted in the marketplace.

Our price system p may or may not be competitive. Since this fact is not common knowledge, it no longer makes sense for households and firms to express their unconstrained demands and supplies, and they should form expectations about supply and demand opportunities. One possible choice for these expectations is optimistic expectations: all households and firms do not expect to be constrained in either supply or demand. When

prices are competitive, we are back in the situation of competitive equilibrium. The question is: are these the only possible expectations compatible with the microeconomic methodological premises of agent optimization and market clearing, together with rational expectations?

Motivated by the empirical regularity that constraints on the supply side are more common than those on the demand side, as is suggested by unemployment in labour markets or unused capacities in production processes, we follow van der Laan (1980) and Kurz (1982) and restrict attention to constraints on the supply side. Moreover, for the sake of simplicity, we consider point expectations about supply opportunities.

If trade takes place at prices p and firm f expects supply opportunities of at most $\overline{y}^f \in \mathbb{R}^L_+$, then firm f's profit maximization problem becomes:

$$\max_{y^f \in Y^f} p \cdot y^f \text{ s.t.} y^f \leq \overline{y}^f.$$

At this point it should be noted that, if a firm f does not produce a particular commodity l, the value of $\overline{y}^f_l$ is entirely inconsequential. Again, under standard assumptions the firm's profit maximization problem is well-defined. In fact, the constraints related to the expected supply opportunities ensure that the firm's profits are bounded from above, a property that does not hold in general for the competitive model. At prices p and expected supply opportunities of $\overline{y}^f$, the supply of firm f, that is, the profit maximizing production plan of firm f, is denoted by $s^f\left(p, \overline{y}^f\right)$ and the firm's profits by $\pi^f\left(p, \overline{y}^f\right)$. The wealth of household h is then equal to $w^h = p \cdot e^h + \sum_f \theta^{fh} \pi^f\left(p, \overline{y}^f\right)$. The utility maximization problem of household h that trades at prices p, expects supply opportunities equal to $\underline{z}^h$, and has budget w^h equals:

$$\max_{x^h \in \mathbb{R}^L_+} u^h\left(x^h\right) \text{ s.t.} p \cdot x^h \leq w^h \quad \textit{and}$$

$$\underline{z}h \leq x^h - e^h.$$

Under standard assumptions, the maximization problem of the household has a unique solution $d^h\left(p, \underline{z}^h, w^h\right)$.

Since supply may not equal demand, one needs a rule to address discrepancies, called a rationing mechanism. Expected supply opportunities should be related to the rationing mechanism, which determines the allocation in case of excess supplies. For labour markets, one can think of a priority system that determines which worker is the first to become unemployed, who is next, and so on and so forth. Another rationing mechanism would share the burden of unemployment equally among workers, for instance by the imposition of an upper bound on the number of hours worked per week.

For notational convenience we assume the latter rationing mechanism in all markets, implying that in equilibrium rational agents face the same expected supply opportunities, $\overline{y}^1 = \cdots = \overline{y}^F = -\underline{z}^1 = \cdots = -\underline{z}^H$. We denote the commonly expected supply opportunities by $r \in \mathbb{R}^L_+$. At expected supply opportunities r, every firm f faces the constraint $y^f \leq r$ and every household h optimizes utility subject to $-r \leq x^h - e^h$. All the results remain true with appropriate modifications for general rationing systems; see Herings (1996b) for a survey of rationing systems encountered in the literature.

A firm or household is said to be rationed in the market of commodity l' if the expected supply opportunities in this market are binding. More precisely, a firm f is *rationed* in the market of commodity l' at prices p and expected supply opportunities r if $\pi^f(p, \overline{r}) > \pi^f(p, r)$, where $\overline{r}_{l'} = +\infty$ and, for $l \neq l^0$, $\overline{r}_l = r_l$. A household h is *rationed* in the market of commodity l' at prices p and expected supply opportunities r if $u^h(d^h(p, -r, w^h)) < u^h(d^h(p; -\overline{r}, w^h))$, where $\overline{r}$ is related to r as before. There is *rationing* on the market of commodity l' if at least one firm or at least one household is rationed on the market of commodity l'.

Definition An *underemployment equilibrium* of the economy

$E = ((u^h, e^h)_h, ((Y^f, (\theta^{fh})_h)_f)$ at prices p are expected supply opportunities r^* and an allocation (x^*, y^*) such that

(a) for every firm f, $y^{*f} = s^f(p, r^*)$,

(b) for every household h, $x^{*h} = d^h(p, -r^*, w^{*h})$, where $w^{*h} = p \cdot e^h + \sum_f \theta^{fh}\pi^f(p, r^*)$,
(c) $\sum_h x^{*h} = \sum_h e^h = \sum_f y^{*f}$.

An example As a simple example, let us consider an economy with one household, one firm, and two commodities. Let us interpret the commodities as leisure and an aggregate consumption good, and suppose that the household owns initially one unit of leisure and nothing of the consumption good, $e = (1, 0)$. The household's utility function is Cobb–Douglas, $u(x) = x_1, x_2$. The firm transforms the labour input into output by the production function $y_2 = \frac{2}{3}\sqrt{-3y_1}$, where $y_1 \leq 0$. When we normalize the price of output to be 1, turning the wage rate into the real wage rate, it is not hard to verify that the competitive equilibrium is given by $p^* = (1, 1), x^* = \left(\frac{2}{3}, \frac{2}{3}\right)$, and $y^* = \left(-\frac{1}{3}, \frac{2}{3}\right)$.

Now suppose that it is possible to trade at the competitive equilibrium prices, so p = (1, 1), but it is not common knowledge that these prices are competitive, and as a consequence firms and households form point expectations on supply opportunities $r = (r_1, r_2)$. We want to verify whether such expectations can be self-confirming. It is easily verified that for each $r_1^* \in \left[0, \frac{1}{3}\right)$ there is an underemployment equilibrium with expected supply opportunities r^* given by $r_2^* = \frac{2}{3}\sqrt{3r_1^*}, x^* = \left(1 - r_1^*, r_2^*\right)$, and $y^* = \left(-r_1^*, r_2^*\right)$. The household expects a constraint on labor supply equal to $-r_1^*$ yielding labour income r_1^*. The firm expects a constraint on the supply of output equal to $r_2^* = \frac{2}{3}\sqrt{3r_1^*}$. It optimally demands an amount of labour equal to r_1^*, leading to profits $\frac{2}{3}\sqrt{3r_1^*} - r_1^*$. Notice that the optimal labour demand by the firm equals the constraint on labour supply anticipated by the household. The household's capital income is equal to $\frac{2}{3}\sqrt{3r_1^*} - r_1^*$, leading to total income of $\frac{2}{3}\sqrt{3r_1^*}$, to be spent on the aggregate consumption good. The optimal demand of the household for the aggregate consumption good equals the supply opportunities expected by the firm, thereby confirming those expectations. There is rationing in both markets. The household is rationed in the labour market and the firm in its output market.

Finally, every $\left(r_1^*, r_2^*\right)$ with $r_1^* \geq \frac{1}{3}$ and $r_2^* \geq \frac{2}{3}$ sustains an underemployment equilibrium that coincides with a competitive equilibrium in terms of the allocation reached, $x^* = \left(\frac{2}{3}, \frac{2}{3}\right)$, $y^* = \left(-\frac{1}{3}, \frac{2}{3}\right)$. In this case, there is no market with rationing.

In the example, two extreme underemployment equilibria stand out. One is the underemployment equilibrium with completely pessimistic expectations about supply opportunities, $r^* = (0, 0)$, $x^* = (1, 0)$, $y^* = (0, 0)$. The other is the underemployment equilibrium with expectations about supply opportunities that are sufficiently optimistic to obtain the competitive allocation; the minimally optimistic expectations to achieve this are $r^* = \left(\frac{1}{3}, \frac{2}{3}\right)$ These extreme underemployment equilibria are connected by a set of underemployment equilibria with more moderate expectations on supply opportunities.

In the example, trade was supposed to take place at competitive prices to highlight underemployment caused by mis-coordination of expectations and not by relative prices that are incompatible with competitive equilibrium. One may argue that it is a probability zero event that trade takes place at competitive prices. The crucial features of the example remain unchanged when trade takes place at non-competitive prices. Suppose that trade takes place at a real wage p_1 above the competitive wage rate of 1. It can be verified that there is still a no-trade equilibrium sustained by completely pessimistic expectations on expected supply opportunities. Although the competitive allocation is no longer feasible when the real wage is above the competitive level, it can be verified that there is also an underemployment equilibrium where the firm does not face rationing and the household observes rationing of its labour supply; the minimally optimistic expectations on supply opportunities that sustain this equilibrium are given by

$$r^* = \left(\frac{1}{3(p_1)}, \frac{2}{3p_1}\right)$$

leading to consumption and production

$$x^* = \left(1 - \frac{1}{3(p_1)^2}, \frac{2}{3p_1}\right), \left(\frac{-1}{3(p_1)^2}, \frac{2}{3p_1}\right).$$

The same underemployment allocation is sustained by more optimistic expectations

$$r_1^* = \frac{1}{3(p_1)^2} \quad \text{and} \quad r_2^* \geq \frac{2}{3p_1}.$$

Again, the two extreme equilibria are connected by a continuum of underemployment equilibria, with expectations ranging from completely pessimistic to expectations that sustain an underemployment equilibrium without rationing of the firm and with rationing of the labour supply of the household.

When the real wage p_1 is below the competitive level, there is still an underemployment equilibrium with completely pessimistic expectations about supply opportunities and no trade. There is also an underemployment equilibrium without rationing of the supply of the household but with rationing of the firm's supply of output. Let $\overline{r}_2$ be equal to $4(\sqrt{1+3(p_1)^2}/6p_1$. Notice that $\overline{r}_2$ is below 2/3 when the real wage p_1 is below 1. The minimally optimistic expectations that sustain an equilibrium without rationing of the household are given by $r^* = (1-\overline{r}_2/p_1, \overline{r}_2)$ leading to an allocation $x^* = (\overline{r}_2/p_1, \overline{r}_2)$, $y^* = (\overline{r}_2/p_1 - 1\overline{r})$. The same underemployment equilibrium allocation is sustained by the more optimistic expectations $r_1^* \geq 1-\overline{r}_2/p_1$ and $r_2^* = \overline{r}_2$. The two extreme underemployment equilibria are connected by a continuum of underemployment equilibria with more moderate expectations on supply opportunities.

Animal Spirits

The example suggests that in general there is a 1-dimensional continuum of equilibria, ranging from a no-trade equilibrium with completely pessimistic expectations to an equilibrium with rather optimistic expectations and without rationing in at least one market. This result is almost true, except that the case with completely pessimistic supply expectations leads to zero income for the households, a case that is well-known to be problematic for equilibrium existence. Inspired by preliminary results in van der Laan (1982), Herings (1996a, 1998) and Citanna et al. (2001) provide conditions under which the following result holds.

Theorem *Under standard assumptions, the economy* $\mathcal{E} = ((u^h, e^h)_h, (Y^f, (\theta^{fh})_h)_f)$ *where trade takes place at prices p possesses a connected set of underemployment equilibria E such that for every* $\rho \in (0, \rightarrow)$, *there is an equilibrium* (r^*, x^*, y^*) *in E with* $\max_l r_l^* = \rho$. (A set is path-connected if any two points in the set can be connected by a path that does not leave the set. Path-connectedness implies connectedness, a slightly weaker topological property of sets, which loosely speaking means that the set consists of one piece.)

The theorem gives general equilibrium underpinnings to the Keynesian ideas that changes in expectations, also referred to as animal spirits, can affect equilibrium economic activity, in particular the level of output and employment. The theorem rules out the case with completely pessimistic expectations, corresponding to $\max_l r_l^* = 0$. It shows that the set E links equilibria with arbitrarily pessimistic expectations ($\max_l r_l^*$ arbitrarily small, but positive) to equilibria with rather optimistic expectations ($\max_l r_l^*$ arbitrarily large). Notice that the condition $\max_l r_l^*$ arbitrarily large only implies that for one market expectations are sufficiently optimistic to rule out rationing. The expectations on supply opportunities on other markets might still be completely pessimistic.

In the absence of production, the statement of the theorem can be simplified somewhat. It is shown in Herings (1998), under standard assumptions, that the economy $\mathcal{E} = ((u^h, e^h)_h)$ where trade takes place at prices p possesses a connected set of underemployment equilibria E such that for every $\rho \in [0, \rightarrow)$ there is an equilibrium (r^*, x^*, y^*) in E with $\max_l r_l^* = \rho$. In exchange economies, the connected set includes an underemployment equilibrium with completely pessimistic expectations.

The theorem demonstrates that the set of equilibria is at least 1-dimensional. In general, one should expect the dimension of the set of equilibria to be exactly equal to 1. The reason is that the model postulates L free variables corresponding to the expected supply opportunities r and

L market clearing conditions. Let E be an economy where trade takes place at prices p and let $z : R^L_+ \to R^L$ denote the excess demand function of the economy, a function of expected supply opportunities r. Because of Walras's law, it holds that for every $r \in R^L_+, p \cdot z(r) = 0$. This implies that there are only L - 1 independent market clearing conditions. Since there are L free variables, this leaves a 1-dimensional solution set.

At this point it should be observed that the same reasoning also applies to the standard competitive model. And indeed, in general the set of competitive equilibria is 1-dimensional too. Whenever (p^*, x^*, y^*) constitutes a competitive equilibrium, so does $(\lambda p^*, x^*, y^*)$ for λ positive. However, in the standard competitive model, the excess demand function is homogeneous of degree 1, implying that the same allocation is sustained by p^* and λp^*. The homogeneity property holds for prices but not for expectations. In general, the excess demand function z is not homogeneous of any degree, and it is not the case that λr^* is part of an underemployment equilibrium when r^* is. Generically, the set E of the theorem contains a 1-dimensional set of distinct equilibrium allocations.

Coordination Failures

The theorem makes clear that a multiplicity of equilibria results even when prices are competitive. The interpretation of the multiplicity of equilibria as coordination failures and the link to the macroeconomic literature on coordination failures were made in Drèze (1997). It would be tempting to conclude that, when trade takes place at competitive prices, then the connected set of underemployment equilibria contains the competitive equilibrium allocation. Although it is true that the competitive equilibrium allocation is an underemployment equilibrium allocation sustained by trade at competitive prices coupled with sufficiently optimistic expectations, it is possible to produce examples where it is outside the connected set of the theorem (Citanna et al. 2001).

Under an additional assumption akin to gross substitutability, the following result holds. If trade takes place at competitive equilibrium prices p, then the economy $\mathscr{E} = ((u^h, e^h)_h, (Y^f, (\theta^{fh})_h)_f)$ possesses a connected set of underemployment equilibria E such that for every $\rho \in (0, \to)$ there is an equilibrium (r^*, x^*, y^*) in E with $\max_l r^*_l = \rho$, and for every $\rho \in (0, \to)$ there is an equilibrium (r^*, x^*, y^*) in E with $\min_l r^*_l = \rho$. More precisely, the following assumption on the aggregate excess demand function suffices. If $r, \overline{r} \in \mathbb{R}^L_+$ with $r \leq \overline{r}$ and $r_{l'} = \overline{r}_{l'}$, then $z_{l'}(r) \leq z_{l'}(\overline{r})$. The interpretation of the assumption is the following. A weak increase in expected supply opportunities in markets different from l' should lead to a weak increase in the excess demand for commodity l'.

This assumption, though strong, is not unreasonable. On the household side, a household may lower its supply of commodity l' in exchange for more supply of another commodity, for instance if the household switches to a more attractive job. A household may also increase its demand for commodity l' as a consequence of higher income. Indeed, more expected supply opportunities of commodities different from l' weakly increase the household's income, which will lead to more demand of commodity l' if it is a normal good. On the producer side, if l' is an output for some firm, then increased supply opportunities of other goods, lead to a weakly lower supply of commodity l', when the firm directs more inputs to the production of the other goods. If l' is an input for a firm, then increased supply opportunities of other goods naturally lead to more production, and thereby an increased input demand, in particular for commodity l'. Notice that the assumption needs to hold only in the aggregate.

Efficiency

It is not hard to argue that underemployment equilibria are not Pareto optimal in general. As soon as there are two commodities, l and l', and two households h and h', such that household h' is rationed in the market for commodity l' but not in the market for commodity l, whereas household h is not rationed in the market for commodity l', then it follows almost immediately that the marginal rate of substitution of commodity l for l' is

not the same for households h and h'. This contradicts Pareto optimality.

It has been argued before that price rigidities may emerge for efficiency reasons. The argument of the previous paragraph makes clear that such is not the case in a complete markets setting when coordination failures are absent. In an incomplete market setting, however, Drèze and Gollier (1993) and Herings and Polemarchakis (2005) show that price rigidities may lead to equilibria that are Pareto superior to competitive equilibria. In general, it will depend on the magnitude of the coordination failures, whether or not welfare improvements result.

Extensions

For illustrative purposes, we have so far considered the simplest case where trade takes place at predetermined prices for all commodities. It is not hard to generalize this set-up substantially, and allow for general lower bounds $\underline{p}$ and upper bounds $\overline{p}$ that define the set of admissible prices at which trade may take place. The notion of underemployment equilibrium should then be extended by the requirement that only when the price of a commodity l equals its lower bound is rationing of the supply of commodity l allowed for. In such a more general setting, it may be interesting to also allow for demand rationing when the price of a commodity equals its upper bound.

Allowing for demand rationing will enlarge the set of equilibria. By taking all $\underline{p}_l$ equal to $-\infty$ and all $\overline{p}_l$ to ∞, we obtain the notion of competitive equilibrium as a special case. The results we mentioned before remain true in this more general set-up.

The existence of a continuum of underemployment equilibria is a robust phenomenon. This seems to be in striking contrast with the conclusion of the fixprice literature, where equilibria are typically locally unique. The reason for this apparent disparity is that the fixprice literature puts one additional constraint on the equilibrium set. It is assumed that there is no rationing in the market of an a priori determined numeraire commodity, called money. Not only is the interpretation of one of the commodities as money controversial, it is also misleading as far as the indeterminacy of equilibrium is concerned.

Suppose we follow Drèze and Polemarchakis (2001) and extend the model by a model of a central bank, which produces money at no cost. Households and firms need money to make transactions, a particular example being a cash-in-advance transactions technology. The central bank sets the nominal interest rate at which it accommodates all money demand. The central bank redistributes the revenues from seigniorage to the households in the form of dividends. Money does not enter into the utility function of the households.

This model fits exactly into the framework discussed so far. Without loss of generality, commodity L may serve as money. The central bank can be modelled as firm F, which can produce the output money in arbitrary amounts without using inputs. The profits of firm F are equal to the seigniorage of the central bank. The price of money is equal to the interest rate. Since the central bank accommodates money demand, the money supply of the central bank is equal to that of a profit-maximizing firm F that expects a constraint on the supply of commodity L equal to the aggregate demand for commodity L (for strictly positive interest rates). Our theorem on the existence of a connected set of underemployment equilibria therefore applies to the case where money is explicitly introduced, thereby contradicting the determinacy results obtained in the fixprice literature.

We have shown how underemployment equilibria result in a general equilibrium model where agents are allowed to form expectations on expected supply opportunities. We analyse whether such expectations can be self-confirming and argue that, even at competitive prices, a continuum of equilibria results, including an equilibrium with approximately no trade and a competitive equilibrium. Such equilibria also arise at other price systems, but are then a consequence of both self-confirming pessimistic expectations and of prices incompatible with competitive equilibrium. Expected supply opportunities in

underemployment equilibria bear a close resemblance to the self-justifying expectations of Keynes (1936), beliefs which are individually rational but socially suboptimal. The further study of underemployment equilibria in models with time and uncertainty, incomplete markets, price-setting agents, and a monetary authority features prominently on the research agenda, as it would allow for explicit links with the modern macroeconomic literature on inflation, output and unemployment.

See Also

- ▶ Determinacy and Indeterminacy of Equilibria
- ▶ Existence of General Equilibrium
- ▶ Fixprice Models
- ▶ General Equilibrium (New Developments)
- ▶ Involuntary Unemployment
- ▶ Money and General Equilibrium
- ▶ Rationing

Bibliography

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

Bénassy, J.-P. 1975. Neo-Keynesian disequilibrium theory in a monetary economy. *Review of Economic Studies* 42: 503–523.

Blad, M. 1978. On the speed of adjustment in the classical tatonnement process: A limit result. *Journal of Economic Theory* 19: 186–191.

Bryant, J. 1983. A simple rational expectations Keynes-type model. *Quarterly Journal of Economics* 98: 525–528.

Citanna, A., H. Crès, J. Drèze, P. Herings, and A. Villanacci. 2001. Continua of underemployment equilibria reflecting coordination failures, also at Walrasian prices. *Journal of Mathematical Economics* 36: 169–200.

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

Cooper, R. 1999. *Coordination games – complementarities and macroeconomics*. Cambridge: Cambridge University Press.

Debreu, G. 1959. *Theory of value*. New Haven: Yale University Press.

Debreu, G. 1974. Excess demand functions. *Journal of Mathematical Economics* 1: 15–21.

Drèze, J. 1975. Existence of an exchange equilibrium under price rigidities. *International Economic Review* 16: 301–320.

Drèze, J. 1989. *Labour management, contracts and capital markets: A general equilibrium approach*. Oxford: Basil Blackwell.

Drèze, J. 1997. Walras–Keynes equilibria-coordination and macroeconomics. *European Economic Review* 41: 1735–1762.

Drèze, J., and C. Gollier. 1993. Risk sharing on the labour market. *European Economic Review* 37: 1457–1482.

Drèze, J., and H. Polemarchakis. 2001. Monetary equilibria. In *Economics essays: A Festschrift for Werner Hildenbrand*, ed. G. Debreu, W. Neuefeind, and W. Trockel. Heidelberg: Springer.

Grandmont, J.-M. 1977. Temporary general equilibrium theory. *Econometrica* 45: 535–572.

Herings, P. 1996a. Equilibrium existence results for economies with price rigidities. *Economic Theory* 7: 63–80.

Herings, P. 1996b. *Static and dynamic aspects of general disequilibrium theory. Theory and decision library series C: Game theory, mathematical programming and operations research*. Dordrecht: Kluwer Academic Publishers.

Herings, P. 1997. Endogenously determined price rigidities. *Economic Theory* 9: 471–498.

Herings, P. 1998. On the existence of a continuum of constrained equilibria. *Journal of Mathematical Economics* 30: 257–273.

Herings, P., and H. Polemarchakis. 2005. Pareto improving price regulation when the asset market is incomplete. *Economic Theory* 25: 135–154.

Jones, L., and R. Manuelli. 1992. The coordination problem and equilibrium theories of recessions. *American Economic Review* 82: 451–471.

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

Kurz, M. 1982. Unemployment equilibrium in an economy with linked prices. *Journal of Economic Theory* 26: 100–123.

Roberts, J. 1987. An equilibrium model with involuntary unemployment at flexible, competitive prices and wages. *American Economic Review* 77: 856–874.

Saari, D. 1985. Iterative price mechanisms. *Econometrica* 53: 1117–1131.

Saari, D., and C. Simon. 1978. Effective price mechanisms. *Econometrica* 46: 1097–1125.

Tuinstra, J. 2000. The emergence of political business cycles in a two-sector general equilibrium model. *European Journal of Political Economy* 16: 509–534.

van der Laan, G. 1980. Equilibrium under rigid prices with compensation for the consumers. *International Economic Review* 21: 63–73.

van der Laan, G. 1982. Simplicial approximation of unemployment equilibria. *Journal of Mathematical Economics* 9: 83–97.

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.

Unemployment

Robert Topel

Abstract

The *unemployed* are individuals who are without work but who are actively seeking employment. The unemployment rate is the percentage of the labour force – the total number of people either working or seeking work – that is unemployed. The evidence suggests that the 'natural rate' of unemployment (or non-employment) is not a constant towards which the labour market converges; rather, it varies with labour market fundamentals.

Keywords

Disability insurance; Labour market institutions; Layoffs; Natural rate of unemployment; Phillips curve; Search models of unemployment; Unemployment; Unemployment insurance

JEL Classifications

J6

The *unemployed* are individuals who are without work but who are actively seeking employment. The *unemployment rate* is the percentage of the labour force – the total number of people either working or seeking work – that is unemployed. Economists and others are interested in unemployment because it says something – we are not sure exactly what – about the economic conditions generally and the success or failure of economic policy.

Some amount of unemployment is both inevitable and efficient because economic fundamentals are stochastic and information is costly. This point was memorialized in the *natural rate* hypothesis of Friedman (1968), Phelps (1974), and Alchian (1969). As Friedman put it:

> 'The natural rate of unemployment'... is the level that would be ground out by the Walrasian system of general equilibrium equations, provided there is embedded in them the actual structural characteristics of labor and commodity markets, including market imperfections, stochastic variability in demands and supplies, the cost of gathering information about job vacancies and labor availabilities, the costs of mobility, and so on.

This point may seem obvious today, but its origins are fairly modern – the 1960s were not so long ago in the history of economic theory. The contributions of Friedman, Phelps and Alchian were reactions to the place of unemployment in Keynesian models, which posited a stable trade-off between unemployment and inflation – the 'Phillips curve'. But their broader impact was to establish that unemployment is an equilibrium phenomenon that occurs for the reasons stated above. This view has framed virtually all subsequent research on unemployment, and its formalization is a continuing research endeavour.

Formalization began with the search theories of McCall (1970), Mortensen (1970), and Gronau (1971); see Rogerson et al. (2005) for a modern survey. The subsequent 'islands' metaphor of Lucas and Prescott (1974) established a formal analysis of equilibrium unemployment. See search models of unemployment.

Data on unemployment are collected by government statistical agencies, based on household surveys that follow more or less uniform standards and definitions in developed countries. For example, in the United States unemployment statistics are collected monthly as part of the Current Population Survey, administered by the Bureau of Labor Statistics, which is a rotating sample of roughly 60,000 households that records (among other things) respondents' self-reported labour market activities during the previous week. Jobless persons who have engaged in some effort to find employment in the past four weeks, or who are awaiting recall to a previous job, are recorded as unemployed. Jobless persons who do not report search activities are recorded as 'out of the labour force'.

In a dynamic economy people may be unemployed for many reasons. Young, new entrants to the labour force seek employment, and, as in marriage, it is typically not a good strategy to

take the first opportunity that comes along. Other unemployed individuals may have left their previous job to look for something better, or they may have been permanently laid off from a previous job because of changing market conditions. Still others may be on temporary layoff, anticipating recall by their previous employer. These examples demonstrate that unemployment (and other labour force 'states') is inherently dynamic: the *stock* of unemployed is ever–changing, and is determined by labour market *flows*. New individuals are constantly joining the ranks of the unemployed via quits, layoffs, or entry to the labour force – the *inflow* to unemployment – while other unemployed job seekers locate and accept new jobs, or choose to stop looking – the *outflow* from unemployment. Changes in either flow affect the level of unemployment.

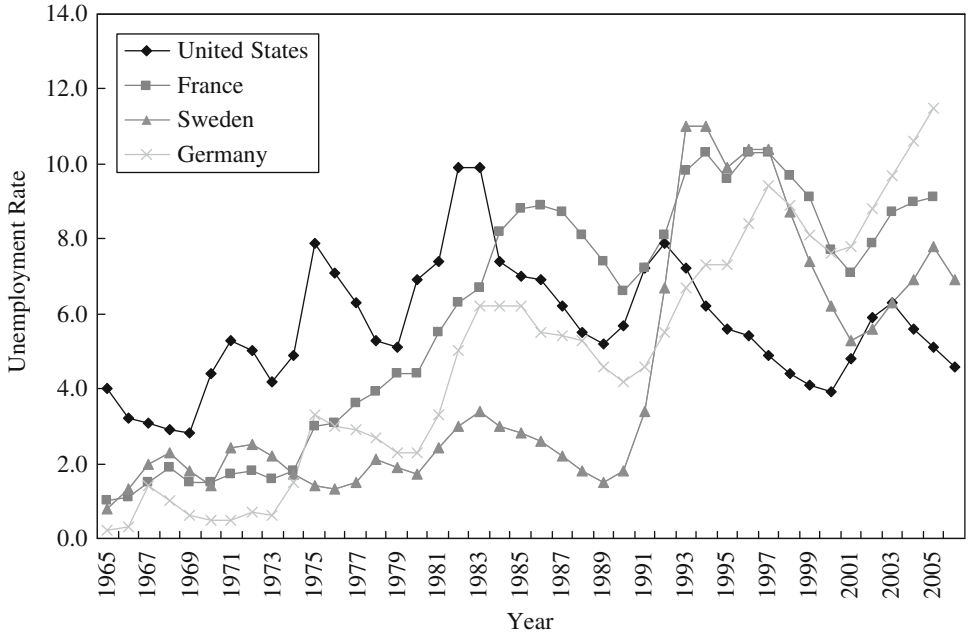
No brief overview can do justice to the vast literature on unemployment, nor can it evaluate the myriad social polices – such as unemployment insurance, public employment agencies or 'active' labour market programmes – that are meant to reduce unemployment or soften its impact on individuals. (Layard et al. 1991, is a slightly dated survey of key issues.) So my aims are more modest. I will summarize key facts about unemployment in the United States, and the factors that have affected the evolution of unemployment, while drawing parallels with other developed economies. Following the arguments in Juhn, Murphy and Topel (JMT) (1991, 2002) and Murphy and Topel (1987, 1997), I provide evidence of a long-term decline in the relative demands for less-skilled workers, so the rewards to employment have declined for marginal workers. This would raise the 'natural rate' of unemployment, but the story is complicated by the fact that some of the unemployed eventually leave the labour force. Over the long run, these changes in economic fundamentals increase joblessness – the total of unemployment and non-participation. This means that current unemployment data have a much different interpretation than in the past. For example, the US unemployment rates of 1974, 1997 and 2006 were about equal, at 4.9% of the labour force. Yet non-participation among prime-aged men rose from 5.2 per cent in 1974 to 8.2 per cent in 1987 and 9.4 per cent in 2006. The reason is that many among the least-skilled had given up searching for work – they were no longer counted as unemployed, but changing labour market forces had left them jobless.

Labour Market Flows and the Evolution of Unemployment

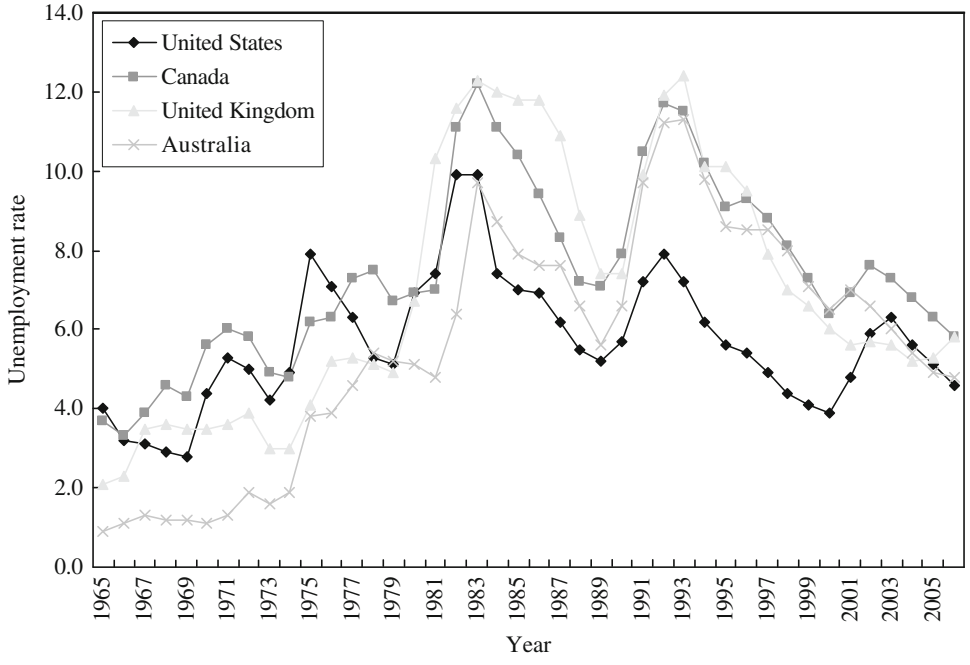
Figures 1 and 2 show the evolution of the male unemployment rate in the United States and several other developed economies since 1965, using comparable definitions. (I present evidence on men because women's labour force participation is typically more varied.)

Focusing on long-term changes, the key message is that unemployment drifted upward in most industrial countries, but especially in Western Europe. The United States is an exception – in comparison with Western Europe, the United States had relatively high unemployment in the 1960s, but substantially lower relative unemployment after about 1990.

Figure 1 also highlights the recent return to 'low' unemployment in the United States. By 2000 the US unemployment rate had reached its lowest level in 30 years, and unemployment rates in 1999–2000 were close to the extremely low rates seen during the late 1960s. This is the culmination of a long downward trend in unemployment: the peak unemployment rates in the recessions of 1982–1983, 1991–1992 and 2001–2002 were progressively lower over time, reversing the trend of rising peaks between the 1970–1971, 1974–1975 and 1982–1983 recessions. (The recession of 1980 did not fit this pattern but as, seen in the figure, did not represent much of a peak in terms of unemployment rates.) It appears that US unemployment has come full circle: unemployment rose for 15 years (from 1968 to 1983), and then fell over the next 17 years (from 1983 to 2000), with intervening cyclical swings. One might conclude from these data that the labour market conditions of the late 1960s and late 1990s were comparable. But in fact the decline in unemployment masks more



Unemployment, Fig. 1 Unemployment rates in the United States and selected countries, 1965–2005 (*Source*: produced by author)



Unemployment, Fig. 2 Unemployment rates in the United States and selected countries, 1965–2005 (*Source*: produced by author)

fundamental changes in labour market flows, driven largely by changes in labour demand that have affected less skilled workers.

The level of unemployment is determined by labour market flows in and out joblessness. One reason for the divergence of US and European unemployment rates is the importance of very long unemployment spells in Europe. According to data collected by the OECD, the *average* duration of unemployment spells in France, Germany or Sweden is over 1 year, compared with only 4 months in the United States. (For Sweden, *I* count individuals enrolled in active labour market programmes, which are required of persons who have not found employment within a fixed number of months.) For OECD Europe as a whole, about 45 per cent of all unemployment spells last more than one year, compared with only 12 per cent in the United States. This means that transitions *out of* unemployment in Europe occur more slowly, which (other transitions equal) raises the unemployment rate.

Suppose there are only two labour market 'states,' employment (E) and unemployment (U). Denote by λ_{EU} the instantaneous transition (hazard) rate from E to U – the inflow to unemployment – and let λ_{EU} be the corresponding hazard for transitions from U to E – the outflow from unemployment. If these transition rates are constant over time and across individuals, then the probability that an individual is unemployed at any date is simply:

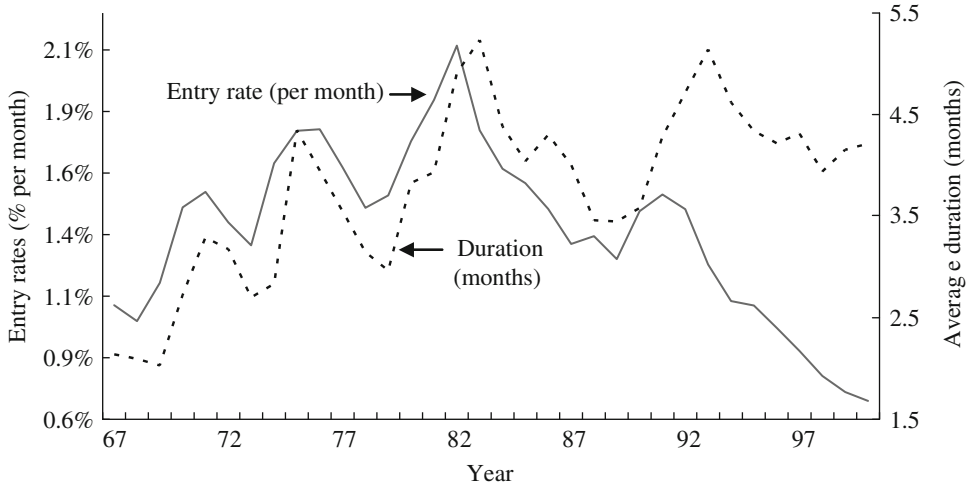
$$\tilde{u} = \frac{\lambda_{EU}}{\lambda_{EU} + \lambda_{UE}} \tag{1}$$

Under these assumptions, Eq. 1 is also the *unemployment rate* in a large population of identical individuals, with corresponding employment rate $e = 1 - \tilde{u}$. Equation 1 accords with the intuition about labour market flows stated earlier: the unemployment rate will be higher the greater the rate of inflow to unemployment, λ_{EU}, or the smaller the rate of outflow from unemployment, λ_{EU}. In this simple setup the expected duration of an unemployment spell is $D = \lambda_{UE}^{-1}$, so policies, institutions or events that increase the duration of spells will increase measured unemployment. In an accounting sense this is why unemployment in Europe is so high – the unemployed remain so for a very long time. *Why* are unemployment durations so long in Europe and relatively short in the United States? I return to this issue below.

Equation 1 demonstrates the key elements of labour market flows, but it isn't very satisfactory as an empirical tool, for (at least) two reasons. First, as noted above, some jobless individuals are not actively seeking employment, at least by the definitions of labour market surveys, and are categorized as 'out of the labour force' (O). Yet many of these 'non-participants' do take jobs, and they may join the ranks of the unemployed by initiating job search. We can accommodate these facts by adding 'O' as a third labour market state, which also adds more transition possibilities (λ_{OE},$\lambda_{EO}\lambda_{UO}$ and so on). Second, labour market flows are obviously not constant – they vary over time and generate corresponding fluctuations in employment, unemployment and labour force participation. So let transition rates be time-varying (for example, $\lambda_{EU}(t)$ is the hazard rate from E to U at time t). Define $e(t)$, $U(t)$ and $o(t)$ as the fractions of the relevant *population* that are employed, unemployed or out of the labour force at date t. (The unemployment *rate* is the fraction of the labour force that is unemployed, or $\tilde{u}(t) = \frac{u(t)}{1-o(t)}$.) Then the law of motion for $u(t)$ is

$$\frac{du(t)}{dt} = e(t)\lambda_{EU}(t) + o(t)\lambda_{OU}(t) - u(t)[\lambda_{UE}(t) + \lambda_{UO}(t)] \tag{2}$$

As above, changes in unemployment are driven by labour-market flows. Other things the same, the fraction of the population that is unemployed increases when transitions *to* unemployment rise. These newly unemployed individuals may have been employed (E) or they may be previous non-participants (O) who have begun to search for work. Similarly, unemployment will fall if transition rates *from* unemployment rise. One usually thinks of this in terms of greater 'job finding'($\lambda_{UE}(t)$), but (2) makes clear that transitions to non-participation – say because of deteriorating labour market opportunities that



Unemployment, Fig. 3 Entry rates and durations for unemployment, 1967–2000 (*Source*: produced by author)

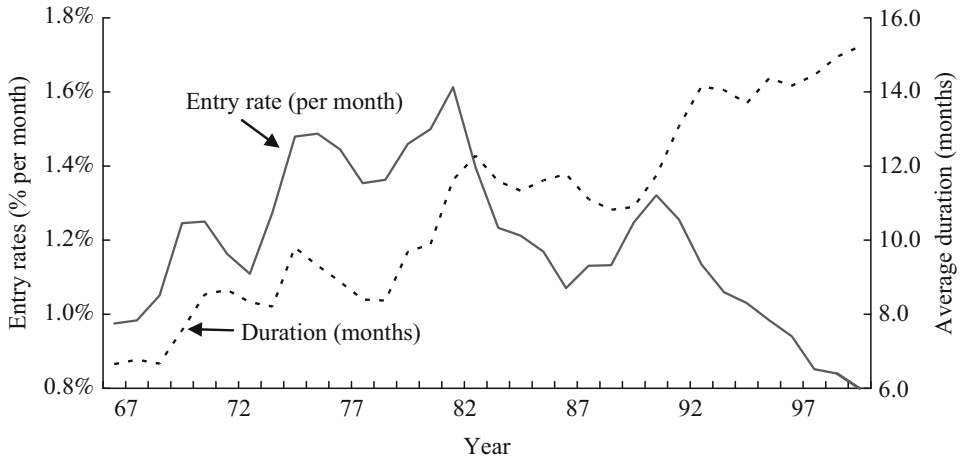
reduce the return to continued search – will also reduce unemployment.

JMT (Juhn et al. 2002) build on (2) to calculate average entry rates and durations of spells for both unemployment (U) and non-employment ($N = U + O$) among American men from 1967 through 2000. Figure 3 shows their results for unemployment. Through the late 1980s entry rates and durations of unemployment spells showed a common pattern, rising in recessions and falling in recoveries, with some evidence of a secular increase in both components. But this tight correspondence was broken in the 1990s – increased incidence of spells played a minor role in the recession of 1991–1992, while durations soared. The ensuing decline in unemployment during the 1990s expansion was driven almost entirely by a reduction in the incidence of unemployment spells, while durations of unemployment remained high – in fact, flows into unemployment fell below their levels in low-unemployment 1960s, while durations were roughly twice as long. With fewer but longer spells, the population distribution of unemployment became much more concentrated than before.

The dichotomy between incidence and duration is more extreme for nonemployment, N, which measures total joblessness in the population without regard to whether individuals are searching for a job. The incidence (entry rate) of jobless spells roughly corresponds with the incidence of unemployment – compare Fig. 3 – but durations of non-employment show a steady increase, more than doubling (to 15 months) since the 1960s. The sharp increase in average durations in the 1990s is especially noteworthy, reflecting the increased proportion of American men who have simply withdrawn from the labour force. Why did this occur? In an accounting sense it is because a large fraction of labour-force 'withdrawals' were temporary in earlier decades, but by the 1990s many men had become 'full-time' non-participants. They had left the labour force and made no efforts to find work, so that average transition rates from non-employment to employment plummeted (Fig. 4).

Who Are the Jobless?

To get a handle on why this occurred, it is worth examining the characteristics of those without jobs. The most basic fact is that unemployment and overall joblessness are much more common among the least skilled. Measuring skill by years of completed schooling, unemployment rates are higher among those with fewer years of schooling, and they also increase more during recessions. JMT (2002) examine a broader definition of skill, based on an individual's position in the overall wage distribution. (JMT impute wages for

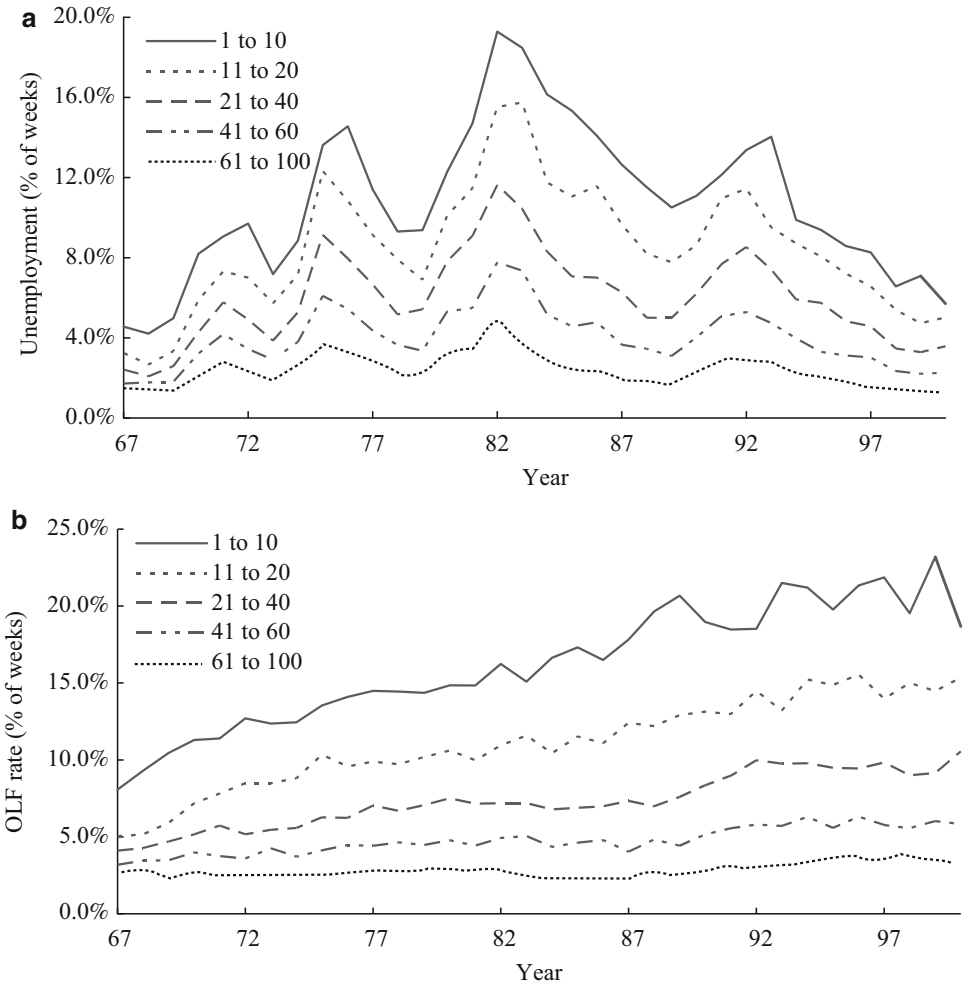


Unemployment, Fig. 4 Entry rates and durations for non-employment, 1967–2000 (*Source*: produced by author)

year-round non-workers from the wage distribution of those who work very few weeks. See JMT 2002, for a description of their methods.)

Figure 5a, b show percentages of the population who are unemployed and out of the labour force (OLF) by percentage intervals of the wage distribution between 1967 and 2000. As with education, both unemployment and non-participation are more common among the less skilled, and in each recession their unemployment rates rise most sharply. Up through the early 1990s, both components of joblessness showed a secular increase that was concentrated among low-wage individuals. For unemployment this trend was reversed in the 1990s but non-participation continued to rise, especially among the least skilled. By 2000 about 20 per cent of men whose skills would put them in the bottom decile of the wage distribution were out of the labour force, which is more than double the fraction of non-participants in the 1960s. Adding the unemployed, by the end of the 1990s nearly 30 per cent of these men were jobless. By comparison, men whose skills put them above the 60th percentile of the wage distribution showed virtually no long-term increase in either unemployment or nonparticipation. In other words, to understand rising joblessness in the United States, we must focus on changes in economic fundamentals that have affected mainly the lower end of the skill distribution.

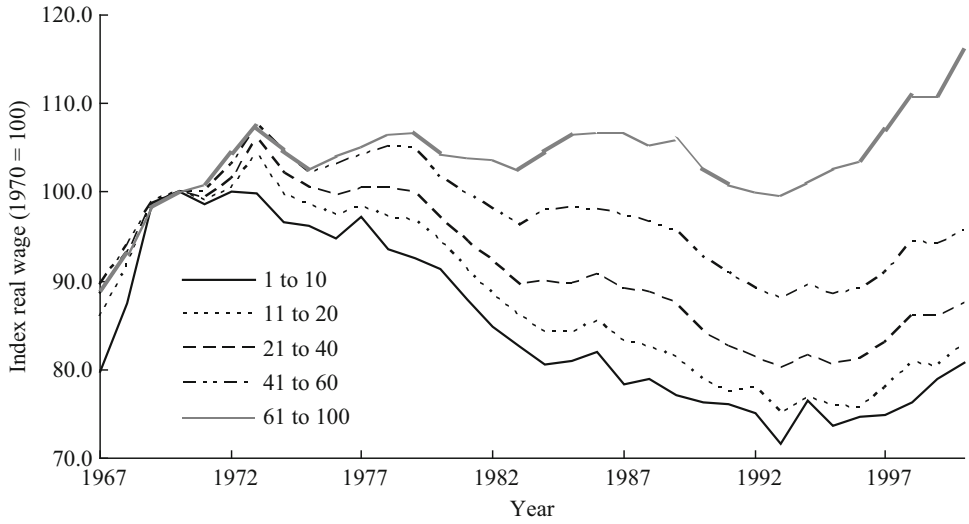
The most obvious explanation is that shifts in relative labour demands have reduced labour market opportunities available to the least skilled, so they end up working less. Evidence consistent with this is the well-documented decline in relative wages among low-skilled workers, shown in Fig. 6. Between 1970 and 1993, real wages of men in the first decile of the wage distribution declined by over 25 per cent, with smaller though still important declines for all workers below the median of the wage distribution. The post-1993 growth in real wages, which affected the entire skill distribution, corresponds to a convergence of relative fractions of time spent unemployed (Fig. 5), but non-participation remained quite high among low-wage workers. The continue rise in non-participation while real wage grew in the 1990s suggests a shift in labour supply. Autor and Duggan (2003, 2006) point to Disability Insurance (DI), which became relatively more attractive to low-skill workers who faced a long-term deterioration of labour market opportunities. They document that participation continued to fall because DI subsidized non-work, which was most attractive to low-skilled workers. In earlier times these individuals would have spent transitory periods of unemployment or non-participation, but they would have remained attached to the labour force over the long run. By 2000, many less skilled individuals – faced with declining working opportunities – had simply withdrawn.



Unemployment, Fig. 5 (**a**) Unemployment rates by wage percentile groups, 1967–2000 (*Source*: produced by author). (**b**) Out of labour force rates by wage percentile groups, 1967–2000 (*Source*: produced by author)

Conclusion: US and European Unemployment Revisited

The broad message of the above evidence is that the 'natural rate' of unemployment (or non-employment) is not a constant towards which the labour market converges; rather it varies with labour market fundamentals – as originally suggested by Phelps (1974). But why have unemployment rates evolved so differently in the United States and Europe? A compelling interpretation is that differences in labour market institutions generate differences in measured unemployment despite similarities in labour market fundamentals. In the United States, the maximum duration of unemployment insurance (UI) is dtypically six months, and the fraction of earnings replaced by UI is about half. Aside from DI, income support for the long-term jobless is comparatively low. These features may accelerate job-finding among those with marketable skills but weaken labour force attachment among the least skilled, whose opportunities have deteriorated. Then long jobless durations among the least skilled show up in labour force withdrawal. In Europe, UI coverage is much more liberal in terms of both the level and duration of benefits, so those who would leave the labour force in the



Unemployment, Fig. 6 Indexed real wages by percentile group, 1967–2000 (*Source*: produced by author)

United States are counted as long-term unemployed. The forces at work and the employment prospects of affected workers are not as different as standard unemployment statistics may suggest.

See Also

- ▶ Labour Market Search
- ▶ Labour Supply
- ▶ Search Models of Unemployment
- ▶ Unemployment and Hours of Work, Cross Country Differences
- ▶ Unemployment Insurance
- ▶ Unemployment Measurement

Bibliography

Alchian, A. 1969. Information costs, pricing, and resource utilization. *Western Economic Journal* 7: 109–128.

Autor, D.H., and Mark G. Duggan. 2003. The rise in the disability rolls and the decline in unemployment. *Quarterly Journal of Economics* 118: 157–205.

Autor, H., and Mark G. Duggan. 2006. The growth in the Social Security disability rolls: A fiscal crisis unfolding. *Journal of Economic Perspectives* 20 (3): 71–96.

Friedman, M. 1968. The role of monetary policy. *American Economic Review* 58: 1–17.

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

Juhn, C., K.M. Murphy, and R.H. Topel. 1991. Why has the natural rate of unemployment increased over time? *Brookings Papers on Economic Activity* 1991 (1): 75–126.

Juhn, C., K.M. Murphy, and R.H. Topel. 2002. Current unemployment, historically contemplated. *Brookings Papers on Economic Activity* 2002 (1): 79–136.

Layard, R., S. Nickell, and R. Jackman. 1991. *Unemployment: Macroeconomic performance and the labour market*. Oxford: Oxford University Press.

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

Lucas, R., and E. Prescott. 1974. Equilibrium search and unemployment. *Journal of Economic Theory* 4: 103–124.

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

Mortensen, D. 1970. A theory of wage and employment dynamics. In *Microeconomic foundations of employment and inflation theory*, ed. E.S. Phelps et al. New York: W.W. Norton.

Murphy, K.M., and R.H. Topel. 1987. The evolution of unemployment in the United States. In *NBER macroeconomics annual 1987*, ed. S. Fischer. Cambridge, MA: MIT Press.

Murphy, K.M., and R.H. Topel. 1997. Unemployment and nonemployment. *American Economic Review* 187: 295–300.

Phelps, E.S. 1974. Economic policy and unemployment in the 1960s. The Public Interest 34(Winter), 30–46.

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

Topel, R.H. 1993. What have we learned from empirical studies of unemployment and turnover? *American Economic Review* 83: 110–115.

Unemployment and Hours of Work, Cross Country Differences

Richard Rogerson

Abstract

Since the 1960s labour market outcomes among the world's richest economies have changed dramatically, especially in terms of unemployment rates and time devoted to market work. This article summarizes the evidence regarding these changes and discusses some of the explanations that have been proposed for why these labour market outcomes have evolved so differently across economies.

Keywords

Barriers to entry; Common shock; Cross-country differences in unemployment and hours worked; Employment protection; Home production; Hours worked; Income support; Labour market regulation; Layoffs; Leisure; Neoclassical growth theory; Product market regulation; Productivity growth; Skill-biased technical change; Technical change; Time use; Unemployment; Unemployment insurance; Wage dispersion; Wage rigidity; Wage setting institutions

JEL Classifications

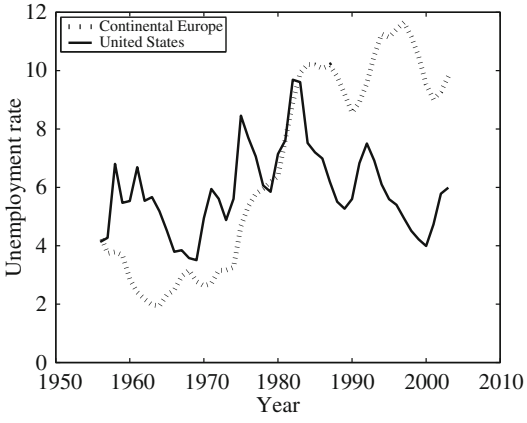
D4; D10

Understanding the forces that determine resource allocations in decentralized economies is one of the fundamental objectives of economics. Because countries often differ greatly in terms of economic policies and institutions, examining how resource allocations differ across countries provides a promising opportunity to learn about these forces. Conversely, when we see large differences in allocations across countries, we are presented with an important opportunity to learn about what factors can generate these large differences. One prominent case in point is the large differences in labour market outcomes – specifically in terms of unemployment rates and hours of work – that exist across rich industrialized economies. A large literature has emerged that documents the nature of these differences and seeks to determine which factors can account for them. This article provides a brief overview of this literature.

Cross-Country Differences in Unemployment

The literature on cross-country differences in unemployment was motivated by a simple real world development: a large and persistent rise in unemployment in several continental European countries relative to the United States, starting in the mid-1970s. Figure 1 displays this fact by showing the evolution of unemployment rates since 1956 for the United States and the average of four economies from Continental Europe: Belgium, France, Germany and Italy.

As the figure reveals, prior to 1970 the unemployment rate in these European countries averaged around three per cent, while since 1990 it has averaged more than ten per cent. This dramatic increase is concentrated in the period 1975–85. In contrast, while the United States also experienced



Unemployment and Hours of Work, Cross Country Differences, Fig. 1 Unemployment in the United States and Continental Europe (*Source:* OECD Labor Market Database)

an increase in average unemployment during the 1975–85 period, this increase is only transitory; the US unemployment rate has averaged close to five per cent in both the early and later time periods. Faced with this dramatic change in relative unemployment rates across countries, economists were naturally led to ask why. Based on the time series plots in Fig. 1, an answer to this question had to address two important issues:

1. Why did the increase occur in the 1975–85 period?
2. Why did it occur in some countries and not in others?

At a general level, one can imagine two classes of explanations. One class postulates that something changed in one set of economies during the period 1975–85 but not in the others. A second class postulates that something changed in all economies during the 1975–85 period, but that economies responded in different ways to this change due to differences in their institutions or policies.

Krugman (1994) was the first to suggest that an explanation of the second type seemed promising. His story emphasized that the process of skill-biased technological change became more prominent beginning in the 1970s, thereby creating an economic force tending to increase the dispersion in individual wages. Although all economies were subjected to this underlying change in technological progress, this change was propagated differently across economies. In the United States, wage setting institutions were 'flexible' and the result was increased wage dispersion and little change in unemployment. In the economies of Continental Europe, wage setting institutions were 'rigid' and did not allow wages to become more dispersed, so the result was instead an increase in unemployment. While subsequent work (see for example, Card et al. 1999) did not provide support for this story, at least in its simplest form, the contribution was important because it suggested an important class of explanations.

The issue of rigorously distinguishing between the two classes of explanations was subsequently taken up in a paper by Blanchard and Wolfers (2000). Based on statistical analysis, these authors argued that the 'common shock' story was most promising. The force of this conclusion is tempered somewhat by two features of the analysis. First, the result that the 'different shocks' explanation does not provide a good account of the data is very much dependent on what shocks are explicitly incorporated in the analysis. In particular, Blanchard and Wolfers did not incorporate the fact that taxes changed considerably over this time period, a point we will return to below. Second, their analysis did not attempt to identify what the important common shock(s) were, and did not construct an explicit model to quantify how various institutions affected the propagation of these shocks. However, making use of advances in general equilibrium modelling of unemployment (such as the Diamond–Mortensen–Pissarides matching model or the Lucas–Prescott island model), subsequent work has sought to remedy this limitation by quantitatively evaluating particular candidates from the 'common shock' class of explanations in the context of fully specified models.

An early example in this literature was Bertola and Ichino (1995). They argued that the common shock was a permanent increase in the transient nature of production opportunities. The key differences in economic institutions were wage setting institutions that compressed wages and employment protection policies that made layoffs prohibitively costly. Several alternative analyses have since followed. Ljungqvist and Sargent (1998) argue that the common shock was a permanent increase in the amount of 'turbulence' for workers, and that the key institutional difference is generosity of income support for displaced workers. Mortensen and Pissarides (1999) and Marimon and Zilibotti (1999) both construct models in which the common shock was skill-biased technological change. While Mortensen and Pissarides stress differences in unemployment insurance (UI) benefits and employment protection as the key institutional differences, Marimon and Zilibotti simply stress differences in UI benefits. Closely related, Hornstein et al. (2007) argue that the common shock is an increase in the rate of capital embodied technological change and that the key institutional differences are taxes, income support programmes, and employment protection.

While explanations of the 'common shock' variety have become popular in the literature, some researchers have argued against them. Using purely statistical methods, Nickell et al. (2006) challenge the Blanchard and Wolfers finding. In terms of model based analyses, Daveri and Tabellini (2000) argue that differences in the changes in tax rates between Continental Europe and the United States were central to understanding the different evolutions in unemployment rates. They also argue that the impact of higher taxes is very much influenced by wage setting institutions, thereby explaining why some other European countries that also experienced large increases in tax rates did not experience sharp increases in unemployment. This last point – that the effects of individual policies and institutions are not additive – has recently been emphasized by both Blanchard and Giavazzi (2002) and Pries and Rogerson (2005). Another model-based analysis is contained in Pissarides (2007). He argues that a significant part of the relative increase in European unemployment is associated with the slowing of productivity growth that came as European productivity converged to US levels.

The literature has made important headway in evaluating specific combinations of driving forces and propagation mechanisms, but there is still much more work to be done. First, much of the work to date has contrasted the behaviour of the United States with an average European country. Given the substantial heterogeneity within Europe, in terms of both policies and outcomes, it is desirable to push these exercises to consider outcomes on a country-by-country basis. Second, as noted above, the literature has focused almost exclusively on accounting for the rise in unemployment in a handful of European economies since 1970. There are also many other interesting episodes in the data. For example, Spain, the United Kingdom and the Netherlands all experienced dramatic increases in unemployment similar to those documented earlier, but each of these countries has subsequently experienced a sharp decrease in unemployment. Understanding the sources of these dynamics should also prove to be very valuable.

While the above discussion has focused on the efforts to understand the sharply different evolutions of unemployment across a small set of countries since the 1970s, the broader research issue is to understand the quantitative consequences of various policy and institutional features on aggregate unemployment. Alvarez and Veracierto (1999) is one example of work that fits with this more general objective. As motivation for this general question one need only look at the distribution of unemployment rates across countries at any point in time. For example, Table 1 shows the distribution of unemployment rates in 2005.

As the reader can see, the dispersion of unemployment rates across countries is large. Understanding the forces that shape this distribution of outcomes remains an open and challenging research issue.

Other Measures of Labour Market Outcomes

If one thinks more carefully about characterizing resource allocations across countries, and specifically as this pertains to the labour market, it becomes clear that differences in unemployment rates may not be the best summary measure of differences in labour market allocations. The benchmark conceptual framework for modern thinking about aggregate resource allocation is

Unemployment and Hours of Work, Cross Country Differences, Table 1 Unemployment rates (2005)

$u < 4.5$	$4.5 < u < 6$	$6 < u < 8$	$u > 8$
NZ (3.7)	Norway (4.6)	Canada (6.8)	Belgium (8.1)
Ireland (4.3)	UK (4.6)	Portugal (7.6)	Finland (8.4)
Japan (4.4)	Denmark (4.8)	Italy (7.7)	Spain (9.2)
Switzerland (4.5)	Australia (5.1) US (5.1)	Sweden (7.7)	France (9.8) Germany (11.2)
	Netherlands (5.2) Austria (5.2)		

Source: OECD Labor Market Database

the one-sector neoclassical growth model. This model stresses two margins that economists believe to be of first-order importance in thinking about aggregate allocations: the fraction of available time that is devoted to market work, and the fraction of output that is invested rather than consumed. Viewed from this perspective, the unemployment rate is the natural summary statistic to focus on only if it is a good measure of time devoted to market work. Historically, the framework of traditional Keynesian models assumed that this was indeed the case: the simple versions of these models assumed that labour supply is represented as some constant volume of available hours that was unaffected by any aspects of the economic environment. The only reason that observed hours of work would differ from this given value was unemployment. In such a conceptual framework, unemployment and total hours of work provide exactly the same information. But is this conceptual framework adequate to examine labour allocations in modern industrialized economies? Developments such as the rise of female labour force participation, the trend towards early retirement, the changing workweek, and the expansion of part-time work suggest that to view labour supply as a fixed volume of work determined only by the size of the population is to neglect some important economic forces.

To pursue this issue further it is instructive to take a closer look at the data. We look at data for 2005, but the basic messages of the analysis are not affected by the choice of year. Table 2 reports hours of work across countries, organizing the countries into four groups based on their hours worked. For each country, aggregate hours of work are computed as the product of two series from the OECD Labor Market Database: total civilian employment and annual hours of work per person in employment. It is important to note that the measure of annual hours of work per person in employment in this data-set attempts to take into account not only the length of the standard workweek but also the number of statutory holidays, sick days and vacation days. To compare aggregate hours of work across countries one has to make some normalization based on population. One could imagine different normalizations, such as the entire population, the adult population (those aged 15 and over), or the working-age population (those aged 15–64). The resulting patterns are not much affected by this choice, and the numbers reported in Table 2 are based on dividing total hours by the size of the working age population.

The correlation between unemployment rates and hours of work in the 2005 cross section is −0.58, indicating a fairly sizable negative correlation. This negative correlation is reflected in the fact that Germany and France are both among the highest unemployment countries as well as the lowest hours-worked countries, while New Zealand is the lowest unemployment country and the highest hours-worked country. However, there are also several counter-examples to this pattern. The Netherlands and Norway, for example, have hours worked and unemployment rates that are both substantially below the average, while Canada has unemployment and hours of work that are both substantially above average. We conclude from this that even at a qualitative

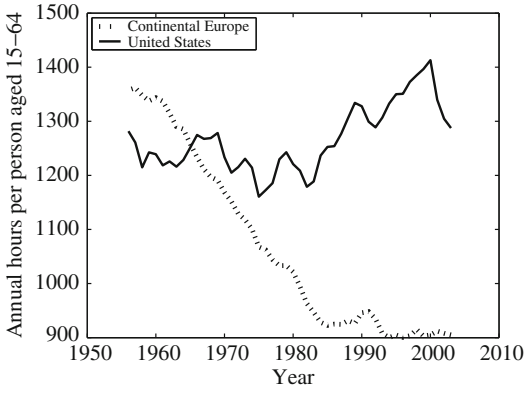
Unemployment and Hours of Work, Cross Country Differences, Table 2 Annual hours worked per person aged 15–64 (2005)

$h < 1{,}000$	$1{,}000 < h < 1{,}150$	$1{,}150 < h < 1{,}300$	$h > 1{,}300$
Belgium (941)	Norway (1,044)	Finland (1,167)	Australia (1,323)
Germany (954)	Italy (1,046)	Denmark (1,191)	Japan (1,333)
France (961)	Ireland (1,122)	Sweden (1,193)	US (1,339)
Netherlands (979)	Austria (1,134)	Portugal (1,213)	NZ (1,386)
	Spain (1,145)	UK (1,240)	
		Canada (1,284) Switzerland (1,286)	

Source: OECD Labor Market Database

level differences in hours of work and differences in unemployment are sometimes quite distinct.

But more importantly, even when differences in hours of work and differences in unemployment describe similar situations qualitatively, the quantitative differences are dramatically different. For example, consider the following question. If the unemployment rate in a country such as France were reduced to the same level as the United States by placing some currently unemployed French people into employment, and having them work the same amount as those French people who are currently working, by how much would the gap in hours worked between France and the United States drop? The answer is that the gap would drop from its current value of 378 to 343, a decrease of less than ten per cent. From a pure accounting perspective, differences in unemployment rates are an almost insignificant component of differences in aggregate hours of work. Put somewhat differently, even if we completely understood the factors that give rise to observed differences in unemployment rates, we would understand practically none of the differences in hours of work.



Unemployment and Hours of Work, Cross Country Differences, Fig. 2 Hours of work in the United States and Continental Europe (*Sources:* OECD Labor Market Database; GGDC Database)

Cross-Country Differences in Hours of Work

The previous section suggests that, if one examines cross-country labour market outcomes from the perspective of differences in resource allocations across economies, it will be useful to look at differences in hours of work rather than unemployment rates. An interesting starting point is to look at the evolution of hours worked for the two sets of economies depicted earlier in Fig. 1. Figure 2 presents this information, where the reader is reminded that in this figure Continental Europe reflects the simple average for the economies of Belgium, France, Germany and Italy.

Several features are worth noting. Hours of work in Continental Europe decrease at a fairly constant rate from 1956 through to the mid-1980s, at which point they flatten out. The magnitude of this decrease in hours worked in Europe is enormous – a drop of over 35 per cent. In contrast, hours worked in the United States are virtually the same in 2005 as they were in 1956, though there are some low-frequency movements in the series during this time period. In contrast with the unemployment rate evolutions, it is of particular interest to note that there is nothing distinctive about the period 1975–85 from the perspective of the decline of hours in Europe.

Given the dramatic change in relative hours worked across these two sets of economies, it is not surprising that a literature has emerged that seeks to understand this change. Here too, one can imagine two different classes of explanations, one based on changes in some feature of the economic environment in Continental Europe relative to the United States, and the other based on a common change in the economic environments that has been propagated differently in the two sets of economies. There are two key differences relative to the literature on unemployment rates: timing and magnitude. Because we see changes beginning in the mid-1950s, and continuing at a fairly constant rate through to the mid-1980s, we presumably need to identify changes that exhibit this general time series pattern. The second difference is that we are talking about changes that are roughly an order of magnitude larger in terms of implications for hours of work.

Interestingly, whereas the unemployment literature has for the most part pursued the 'common shocks' explanation, the hours of work literature has instead focused on differences in policy changes across countries. This view has been heavily influenced by the contribution of Prescott (2004). He argues that, once cross-country differences in taxes are incorporated into the standard growth model, hours of work in the United States and several European economies in both the early 1970s and early 1990s are very close to those predicted by the model that assumes no other differences between the different economies. This general finding is further supported by Davis and Henrekson (2005), Ohanian et al. (2006), and Rogerson (2007).

One argument against the tax explanation, as noted by Alesina et al. (2005), is that it requires an aggregate labour supply elasticity that is larger than the individual labour supply elasticities typically found in estimation exercises using micro data. Recent work suggests that this critique is misplaced. Chang and Kim (2006) and Rogerson and Wallenius (2007) both argue that, when non-convexities are relevant for individual level labour supply decisions, the tight connection between micro and macro elasticities is broken. In particular, both papers argue that reasonable calibrations imply that small micro elasticities are consistent with large macro elasticities. Additional discussion can be found in Prescott (2006) and Rogerson (2006).

While an explanation based on taxes has been the one most developed to date, there are competing explanations. In work that is closely related but distinct from the Prescott analysis, Ljungqvist and Sargent (2007) argue that it is the nature of benefit programmes in Europe rather than the high tax rates per se that are responsible for the large differences in employment rates between the United States and several European economies. Interestingly, Ljungqvist and Sargent (2007) argue that large aggregate elasticities are inconsistent with the observed cross-country differences once one properly models benefit programmes. On a very different note, Blanchard (2005) has argued that the dominant factor is differences in preferences across countries. In particular, he suggests that the income effect on leisure is larger in European countries than in the United States, and as European productivity has increased to near US levels since the 1960s, Europeans have responded with a larger increase in leisure than that which occurred earlier in the United States. An interesting connection between this explanation and the one based on tax rates is that it is the income effect that is central to generating the tax effects in the analysis of Prescott.

While the discussion has thus far focused on understanding the reasons for the very different evolutions of hours worked between the United States and Continental Europe, the data in Table 2 reveal a host of other differences that are also of interest. Rogerson (2006) describes some additional patterns of interest that are found when one disaggregates the data by age, gender and sector as well as along the employment and hours per worker margins. At a general level, the issue is to understand both qualitatively and quantitatively how various policies or institutions influence hours worked in an economy. Prominent examples of the policies and institutions of interest include such things as labour market regulations (for example, minimum wages, employment protection), product market regulation (for example, entry barriers), tax and transfer policies (for example, unemployment insurance, social security, disability), wage setting practices (for example, importance of unions, centralized versus decentralized wage negotiations). In addition to the papers mentioned earlier, other examples of work that address some of these issues are Bertrand and Kramarz (2002) and Messina (2006) for entry barriers, and Bentolila and Bertola (1990) and Hopenhayn and Rogerson (1993) for firing taxes.

Economists have also recently begun to examine differences in time allocations across countries. If it is the case that individuals in one country spend much more time in market work than individuals in another country, where does this show up in terms of other uses of time? Specifically, to what extent do these differences reflect differences in leisure versus differences in time devoted to home production? In a

provocative paper, Freeman and Schettkat (2001) analysed time use data for American and German couples in the 1990s and found that total working time was similar across the two economies, but that there was a systematic difference in the composition of this total working time: couples in the United States devoted more time to market work than German couples, whereas German couples devoted more time to home production than American couples. This finding, coupled with the fact that additional time use surveys have been initiated both in Europe and the United States, has spawned a growing literature on the general issue of time allocations across countries, including Freeman and Schettkat (2005) and Hamermesh et al. (2007). Olovsson (2004), Ragan (2005) and Rogerson (2007) all argue that modelling home production is important in understanding the differences in hours of market work across countries.

Conclusions

Labour market outcomes differ dramatically across industrialized countries along several dimensions, including hours of work and unemployment. These differences have changed dramatically over time. Understanding the source of these differences and their evolution over time is a key challenge for economists, and will likely have important consequences for policymakers as well as yield important insights regarding the role of various factors in shaping labour market outcomes. This article has provided a brief introduction to this line of research. While much progress has been made to date, much more work remains to be done.

See Also

- ▶ Hours Worked (Long-Run Trends)
- ▶ Labour Market Institutions
- ▶ Labour Supply
- ▶ Search Models of Unemployment
- ▶ Unemployment

Bibliography

Alesina, A., E. Glaeser, and B. Sacerdote. 2005. Work and leisure in the US and Europe: Why so different? In *NBER macroeconomics annual*, vol. 20, ed. M. Gertler and K. Rogoff. Cambridge, MA: MIT Press.

Alvarez, F., and M. Veracierto. 1999. Labor market policies in an equilibrium search model. In *NBER macroeconomics annual*, vol. 14, ed. B.S. Bernanke and J.J. Rotemberg. Cambridge, MA: MIT Press.

Bentolila, S., and G. Bertola. 1990. Firing costs and labor demand: How bad is eurosclerosis? *Review of Economic Studies* 57: 381–402.

Bertola, G., and A. Ichino. 1995. Wage inequality and unemployment: United States vs. Europe. In *NBER Macroeconomics annual*, vol. 10, ed. B.S. Bernanke and J.J. Rotemberg. Cambridge, MA: MIT Press.

Bertrand, M., and F. Kramarz. 2002. Does entry regulation hinder job creation? Evidence from the French retail industry. *Quarterly Journal of Economics* 117: 1369–1413.

Blanchard, O. 2005. The economic future of Europe. *Journal of Economic Perspectives* 18(4): 3–26.

Blanchard, O., and F. Giavazzi. 2002. Macroeconomic effects of regulation and deregulation in goods and labor markets. *Quarterly Journal of Economics* 117: 879–907.

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

Card, D., F. Kramarz, and T. Lemieux. 1999. Changes in relative structure of wages and employment: A comparison of the United States, Canada and France. *Canadian Journal of Economics* 32: 843–877.

Chang, Y., and S. Kim. 2006. From individual to aggregate labor supply: A quantitative analysis based on a heterogeneous agent macroeconomy. *International Economic Review* 47: 1–27.

Daveri, F., and G. Tabellini. 2000. Unemployment, growth and taxation in industrial countries. *Economic Policy* 15: 47–104.

Davis, S., and M. Henrekson. 2005. Tax effects on work activity, industry mix and shadow economy size: Evidence from rich country comparisons. In *Labour supply and incentives to work in Europe*, ed. R. Gomez-Salvador, A. Lamo, B. Petrongolo, M. Ward, and E. Wasmer. Northampton: Edward Elgar.

Freeman, R., and R. Schettkat. 2001. Marketization of production and the US–Europe employment gap. *Oxford Bulletin of Economics and Statistics* 63: 647–670.

Freeman, R., and R. Schettkat. 2005. Marketization of household production and the EU–US gap in work. *Economic Policy* 20: 6–50.

Hamermesh, D., M. Burda, and P. Weil. 2007. The distribution of total work in the EU and the US. In *Are Europeans lazy or Americans crazy?* ed. T. Boeri. Oxford: Oxford University Press.

Hopenhayn, H., and R. Rogerson. 1993. Job turnover and policy evaluation: A general equilibrium analysis. *Journal of Political Economy* 101: 915–938.
Hornstein, A., P. Krusell, and G. Violante. 2007. Technology–policy interactions in frictional labor markets. *Review of Economic Studies* 74: 1089–1124.
Krugman, P. 1994. Past and prospective causes of high unemployment. *Economic Review of the Federal Reserve Bank of Kansas City* 79(4th quarter): 23–43.
Ljungqvist, L., and T. Sargent. 1998. The European unemployment dilemma. *Journal of Political Economy* 106: 514–550.
Ljungqvist, L., and T. Sargent. 2007. *Do taxes explain European unemployment? Indivisible labor, human capital, lotteries and savings*. Working paper, New York University.
Marimon, R., and F. Zilibotti. 1999. Unemployment versus mismatch of talents: Reconsidering unemployment benefits. *Economic Journal* 109: 266–291.
Messina, J. 2006. The role of product market regulations in the process of structural change. *European Economic Review* 50: 1863–1890.
Mortensen, D., and C. Pissarides. 1999. Unemployment responses to skill-biased technology shocks. *Economic Journal* 109: 242–265.
Nickell, S., L. Nunziata, and W. Ochel. 2006. Unemployment in the OECD since the 1960s. What do we know? *Economic Journal* 115: 1–27.
Ohanian, L., A. Raffo, and R. Rogerson. 2006. *Long-term changes in labor supply and taxes: Evidence from OECD countries, 1956–2004*. Working paper no. 12786. Cambridge, MA: NBER.
Olovsson, C. 2004. *Why do Europeans work so little?* Mimeo: Stockholm School of Economics.
Pissarides, C. 2007. Unemployment and hours of work: The North Atlantic divide revisited. *International Economic Review* 48: 1–36.
Prescott, E. 2004. Why do Americans work so much more than Europeans? *Quarterly Review of the Federal Reserve Bank of Minneapolis* 28(1): 2–13.
Prescott, E. 2006. The transformation of macroeconomic policy and research, 2004 Nobel Prize address. *Journal of Political Economy* 114: 203–236.
Pries, M., and R. Rogerson. 2005. Hiring policies, labor market institutions and labor market flows. *Journal of Political Economy* 113: 811–839.
Ragan, K. 2005. *Taxes, transfers and time use: Fiscal policy in a household production model*. Mimeo, University of Chicago.
Rogerson, R. 2006. Understanding differences in hours worked. *Review of Economic Dynamics* 9: 365–409.
Rogerson, R. 2007. *Structural transformation and the deterioration of European labor market outcomes*. Working paper no. 12889. Cambridge, MA: NBER.
Rogerson, R., and J. Wallenius. 2007. *Micro and macro elasticities in a life cycle model with taxes*. Working paper no. 13017. Cambridge, MA: NBER.

Unemployment Insurance

Patricia M. Anderson

Abstract

Unemployment insurance (UI) is a social insurance programme in which compensation is paid to unemployed workers. Much of the research on UI has focused on the inherent disincentives. For example, higher benefits have been found to increase unemployment durations, with little clear positive impact on the quality of new jobs. Additionally, financing UI through payroll taxes that are not completely experience-rated provides an incentive for firms to lay off workers. Thus, while UI is an important safety net for unemployed workers, it may also increase unemployment overall.

Keywords

Adjustment costs; Consumption smoothing; Labour demand; Labour supply; Layoffs; Reservation wage; Search models of unemployment; Tax incidence; Unemployment; Unemployment insurance

JEL Classification

J6

Unemployment insurance (UI) is a social insurance programme whereby compensation is paid to unemployed workers. The federal–state UI programme in the United States dates from the Social Security Act of 1935, while many European countries began national programmes even earlier. For example, the National Insurance Act of 1911 established UI in the United Kingdom, while Italy and Germany established programmes in 1919 and 1927, respectively.

Institutional Aspects of UI

While specific institutional details differ across countries, a typical UI programme is limited to

workers who are unemployed through no fault of their own (that is, who have neither quit nor been fired for cause), who are actively looking for work, and who have a demonstrated attachment to the labour force, with the benefits being paid for a limited period of time (see for example, Atkinson and Micklewright 1991). Typically, the weekly benefit amount (WBA) is based on previous earnings, using a replacement rate schedule that is subject to a minimum and maximum WBA.

Financing of UI programmes also differs across countries. In some cases it is funded out of general revenues, while in most countries it is funded by a flat tax levied on employers and sometimes on employees (see Storey and Niesner 1997 for a complete overview of UI programmes in the G-7 nations). Empirical evidence from the United States, though, shows that even when the tax is levied solely on the employer, the incidence is largely on the employee (Anderson and Meyer 1997, 2000). In the United States, the employer tax is also experience-rated. That is, a firm's tax rate depends on the use of the UI system by its previous employees, with new firms typically charged a rate based on industry experience for the first few years. While each state has its own institutions, a typical system can be characterized by thinking of each firm as having an account with the state UI authority. Taxes are paid into this account, and benefits are paid out of it when the firm lays off employees. A schedule then relates the balance in this account to a tax rate, subject to minimum and maximum rates. A high account balance will merit a lower tax rate, while a lower (possibly even negative) balance will merit a higher tax rate. This characterization is a gross simplification, but captures the basic components of an experience-rated system.

Economic Incentives of UI

With experience rating, a firm which lays off a worker today can expect to pay some fraction of that worker's benefits through higher tax payments in the future (for early theoretical work, see Feldstein 1976; Baily 1976; Brechling 1977a, b). For the most common systems used in the United States, one can calculate this marginal tax cost (MTC) of a layoff (see for example, Topel 1983; this derivation follows Anderson and Meyer 1993). Let θ be the growth of employment (that is, $N_{t+1} = \theta N_t$), γ be the growth in the taxable wage base (that is, $W_{t+1} = \gamma W_t$), and approximate the tax schedule as a linear relationship with slope η. The tax bill can then be expressed in terms of benefits paid, N, Wand the parameters γ and η. For interest rate i, when benefits paid increase by a dollar, the present value of the change in this tax bill is

$$\sum_{t=1}^{\infty} \frac{\theta^2\gamma^2\eta}{1-\theta^2\gamma^2\eta}\left(\frac{1-\theta^2\gamma^2\eta}{1+i}\right)^t.$$

This sum converges to

$$\frac{\theta^2\gamma^2\eta}{i+\theta^2\gamma^2\eta},$$

which is referred to as the marginal tax cost (MTC) of a layoff. For firms at the maximum (or minimum) tax rate, $\eta = 0$ and this MTC will be zero, implying that benefits to this firm's workers are completely subsidized. Alternatively, the steeper the slope of the tax schedule the closer this marginal tax cost is to 1, and the more perfectly experience-rated is the system.

The subsidy inherent from incomplete experience rating provides an incentive to lay off workers. In fact, Topel (1983) estimates that incomplete experience rating in the United States could be responsible for about one-third of temporary-layoff unemployment spells. More broadly, the MTC can be thought of as a simple adjustment cost, implying that tighter experience rating would not only reduce layoffs in a downturn, but also reduce hiring in a boom, resulting in decreased employment fluctuations (see for example, Anderson 1993; Card and Levine 1994; Anderson and Meyer 1994).

More work exists on the employee disincentives of UI than the firm disincentives. Two simple models imply that higher benefit levels will result in longer unemployment durations. First, as shown in Moffitt and Nicholson (1982),

incorporating UI into the budget constraint of a labour supply model results in income and substitution effects which both imply fewer weeks worked. Similarly, a simple job-search model implies that higher UI benefits lower the cost of unemployment, thus increasing the reservation wage and lowering the probability of accepting a new job. The result is again an increase in the duration of unemployment.

The simple job-search model is extended in Mortensen (1977) to incorporate realistic UI programme features such as minimum work requirements for initial qualification, and limited duration of benefits. In this model, for the unemployed who are not currently qualified to receive UI, a new job that could lead to future UI qualification is more valuable the higher the benefit level. In this case, a negative relationship between duration and benefit level would be expected. Allowing benefits to be of limited duration has additional implications as well. First, the level of benefits should have no effect on the reservation wage after they run out. However, search intensity may increase around exhaustion, implying that potential duration of benefits may have a direct effect on unemployment duration.

The economic effects of UI are not all negative. For example, a job-search model also implies that increased duration should result in higher-quality jobs being found. Additionally, UI benefits can allow individuals to smooth consumption during unemployment spells. Finally, this consumption smoothing benefit implies UI can also play an important role as a macroeconomic stabilizer by helping maintain aggregate spending. More broadly, there is a growing literature on optimal UI which takes into account the need to balance costs and benefits (see for example, Baily 1978; Chetty 2006, for partial equilibrium and Hopenhayn and Nicolini 1997; and Acemoglu and Shimer 1999 for general equilibrium analyses).

Empirical Evidence on the Effects of UI

Most empirical work on UI focuses on the costs, with studies generally confirming a positive relationship between duration and benefit levels. For example, studies of UI recipients in the United Kingdom have found elasticities of around 0.3, while those in the United States have found elasticities in the range of 0.4–1. Studies in other OECD countries have typically found relatively low elasticities, although they tend to be measured without much precision (see Atkinson and Micklewright 1991 for a review of these studies). Additionally, empirical studies that allow for the exit rate out of unemployment to 'spike' around the time of benefit exhaustion have found just such an effect both overall (for example, Meyer 1990) and for new jobs and recalls separately (for example, Katz and Meyer 1990). Additionally, studies in the United States have tended to find that a 1-week increase in potential duration results in between a 0.1 and 0.2 week increase in unemployment, with Canadian studies finding slightly larger effects (Atkinson and Micklewright 1991).

There are fewer empirical studies of the benefits of UI. A notable exception is Gruber (1997), which finds a large consumption smoothing effect of higher benefits. Finally, while a job-search model implies that higher-quality jobs should be found, empirical evidence is largely mixed on this effect (see Decker 1997 for a review of US studies of benefit effects). In particular, in the United States, several re-employment bonus experiments took place in which UI claimants were offered a cash bonus if they found a new job within a specific amount of time (see Meyer 1995, for a review). The early experiments found significant reductions in unemployment durations, but no real decline in post-unemployment earnings as would be expected if benefits were significantly subsidizing search. Overall, then, while UI is an important safety net for unemployed workers, it may also increase unemployment.

See Also

- ▶ Adjustment Costs
- ▶ Layoffs
- ▶ Social Insurance
- ▶ Unemployment

Bibliography

Acemoglu, D., and R. Shimer. 1999. Efficient unemployment insurance. *Journal of Political Economy* 107: 893–928.

Anderson, P.M. 1993. Linear adjustment costs and seasonal labor demand: Evidence from retail trade firms. *Quarterly Journal of Economics* 108: 1015–1042.

Anderson, P.M., and B.D. Meyer. 1993. Unemployment insurance in the United States: Layoff incentives and cross subsidies. *Journal of Labor Economics* 11(1, Part II): S70–S95.

Anderson, P.M., and B.D. Meyer. 1994. *The effects of unemployment insurance taxes and benefits on layoffs using firm and individual data*, Working paper no. 4960. Cambridge, MA: NBER.

Anderson, P.M., and B.D. Meyer. 1997. The effects of firm specific taxes and government mandates with an application to the U.S. unemployment insurance program. *Journal of Public Economics* 65: 119–145.

Anderson, P.M., and B.D. Meyer. 2000. The effects of the unemployment insurance payroll tax on wages, employment, claims and denials. *Journal of Public Economics* 78: 81–106.

Atkinson, A.B.., and J. Micklewright. 1991. Unemployment compensation and labor market transitions: A critical review. *Journal of Economic Literature* 29: 1679–1727.

Baily, M. 1976. On the theory of layoffs and unemployment. *Econometrica* 45: 1043–1063.

Baily, M.N. 1978. Some aspects of optimal unemployment insurance. *Journal of Public Economics* 10: 379–402.

Brechling, F. 1977a. The incentive effects of the U.S. unemployment insurance tax. In *Research in labor economics*, ed. R. Ehrenberg. Greenwich: JAI Press.

Brechling, F. 1977b. Unemployment insurance and labor turnover: Summary of theoretical findings. *Industrial and Labor Relations Review* 30: 483–494.

Card, D., and P.B. Levine. 1994. Unemployment insurance taxes and the cyclical and seasonal properties of unemployment. *Journal of Public Economics* 53: 1–29.

Chetty, R. 2006. A general formula for the optimal level of social insurance. *Journal of Public Economics* 90: 1879–1901.

Decker, P.T. 1997. Work incentives and disincentives. In *Unemployment insurance in the United States: Analysis of policy issues*, ed. C.J. O'Leary and S.A. Wandner. Kalamazoo: W.E. Upjohn Institute for Employment Research.

Feldstein, M.S. 1976. Temporary layoffs in the theory of unemployment. *Journal of Political Economy* 84: 937–957.

Gruber, J. 1997. The consumption smoothing benefits of unemployment insurance. *American Economic Review* 87: 192–205.

Hopenhayn, H.A., and J.P. Nicolini. 1997. Optimal unemployment insurance. *Journal of Political Economy* 105: 412–438.

Katz, L.F., and B.D. Meyer. 1990. Unemployment insurance, recall expectations, and unemployment outcomes. *Quarterly Journal of Economics* 105: 973–1002.

Meyer, B.D. 1990. Unemployment insurance and unemployment spells. *Econometrica* 58: 757–782.

Meyer, B.D. 1995. Lessons from the U.S. unemployment insurance experiments. *Journal of Economic Literature* 33: 91–131.

Moffitt, R., and W. Nicholson. 1982. The effect of unemployment insurance on unemployment: The case of federal supplemental benefits. *Review of Economics and Statistics* 64: 1–11.

Mortensen, D.T. 1977. Unemployment insurance and job search decisions. *Industrial and Labor Relations Review* 30: 505–517.

Storey, J.R., and J.A. Niesner. 1997. Unemployment compensation in the group of seven nations. In *Unemployment insurance in the United States: Analysis of policy issues*, ed. C.J. O'Leary and S.A. Wandner. Kalamazoo: W.E. Upjohn Institute for Employment Research.

Topel, R.H. 1983. On layoffs and unemployment insurance. *American Economic Review* 73: 541–559.

Unemployment Measurement

Katharine Bradbury

Abstract

Measures of unemployment tally people without a job who are looking for one. For measurement purposes, the critical question is what constitutes 'looking'. This article summarizes how unemployment is measured in the United States and Europe, and describes recent research investigating the permeability of the dividing line between the unemployed and 'marginally attached' subgroups of those out of the labour market. A continuum between unemployed and entirely inactive individuals indicates that additional measures beyond unemployment may be useful in judging the state of the labour market.

Keywords

Labour force participation rate; Labour market search; Non-employment; Unemployment; Unemployment measurement

JEL Classifications
C8

Measures of unemployment attempt to count individuals who do not have a job but are looking for one. While the concept is reasonably straightforward, various measurement approaches are used to distinguish those out of work who are looking from those who are not, generally based on the specific methods these individuals use to search for employment, how intensively they search, and how long it has been since they 'actively' searched.

Official Measure in the United States

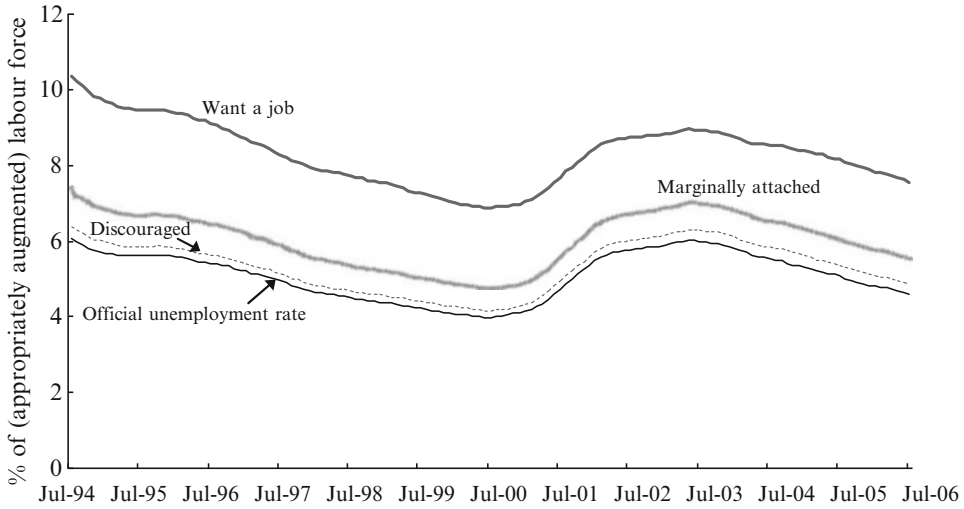
In the United States, unemployment is gauged by comparing the number of unemployed individuals with the size of the labour force, as determined by a monthly survey of households. The civilian labour force is defined as individuals aged 16 and older who are either employed or unemployed, but not on active duty in the armed forces. (See U.S. Bureau of Labor Statistics 2006; Jacobs 2006, pp. 4–8.)

Individuals are considered unemployed if (*a*) they lack a job, (*b*) are available to work, and (*c*) have actively sought employment in the four weeks preceding the survey or are awaiting callback to an existing job (even if they did not actively seek employment). Active job search includes contacting employers or employment agencies, sending out résumés, or placing or answering advertisements; simply reading want ads is not considered active job search. Individuals are employed if they worked at least one hour as paid employees during the reference week, or worked in their own business or farm, or worked unpaid for 15 hours or more in a family business, or had a job but were temporarily away from it due to illness, vacation, labour management disputes, parental leave, job training, or other personal or family related reasons.

Anyone in the civilian non-institutional population (ages 16 and older and neither in the military nor in an institution such as a prison or mental hospital) who is not employed and not unemployed is considered to be out of the labour force. The U.S. Bureau of Labor Statistics (BLS) collects information on those out of the labour force to assess the degree to which they may be 'marginally attached' to the labour force, by asking about their interest in and availability for work.

Based on these questions, the BLS defines a set of 'alternative measures of labor underutilization' that either subtract from the official unemployment rate (for example, by counting only the long-term unemployed) or add to it. The nature of the questions and the alternative measures were revised as of January 1994. Figure 1 plots two of these 'alternative measures' and a third concept, along with the official unemployment rate, over the 1994–2006 period. The line marked 'discouraged' adds to the official unemployed all discouraged workers, defined as those who have given a job market-related reason for not currently looking for a job, including people who think that no work is available, or who could not find work, or who lack schooling or training, or who say potential employers think they are too young or old, or who believe they have been subject to other types of discrimination. The next measure adds all other marginally attached workers, defined as persons who currently are neither working nor looking for work but indicate that they want, and are available for, a job and have looked for work at some time in the preceding 12 months.

Beyond what the BLS defines as 'marginally attached' is a group who has not looked for work in the last 12 months but answers 'yes' to the question 'do you currently want a job?' This line is labelled 'want a job' in the figure. The alternative measures add from a few tenths of a percentage point (discouraged workers) to a percentage point (marginally attached) to several percentage points (want a job) to the official unemployment rate. While too small to be seen clearly in the figure, the number of individuals in these marginal categories varied cyclically over the 1994–2005 period in a manner similar to the number unemployed.



Unemployment Measurement, Fig. 1 US measures of labour underutilization, 1994–2006 (*Notes:* Not seasonally adjusted, 12-month centred moving averages. Discouraged workers are a subset of the marginally attached; the marginally attached are a subset of those who say they want a job. Each measure adds the noted group to 'officially' unemployed individuals and expresses that sum as a percent of the labour force plus the noted group. *Sources:* US Bureau of Statistics and author's calculations)

Unemployment Measures Elsewhere in the Industrialized World

In most of the industrialized world, the concept of unemployment is the same and the definition for measurement purposes is quite similar. The dimensions along which measures differ include the type of job search activities that distinguish the unemployed from the marginally attached, the definition of 'currently available' for work, and the treatment of individuals on layoff or waiting to start a new job. In addition, age cut-offs may vary. (For more detailed discussion of inter-country definitional differences, see Sorrentino 2000.)

The measurement definition chosen by the Statistical Office of the European Communities (Eurostat) in September 2000 is used to calculate 'standardized' unemployment rates for member nations. Eurostat conducts the European Union Labor Force Survey (EU LFS) on a quarterly basis and also releases 'harmonized' monthly estimates based on data from member states. The EU LFS divides the population of working age (defined as ages 15 and older) into three groups: employed, unemployed, and inactive. The economically active population (like the labour force in the United States) comprises employed and unemployed persons.

According to Eurostat, the unemployed are those aged 15–74 (or 16–74 in a few nations), who (*a*) were without work during the reference week, (*b*) were available to start work within two weeks, and (*c*) had either actively sought work in the past four weeks or had already found a job to start within the next three months. The specific steps that qualify as actively seeking work include any of the following: being in contact with a public employment office or a private agency to find work, applying to employers directly, asking among friends, relatives, unions, and so on, to find work, placing or answering job advertisements, studying job advertisements, taking a recruitment test or examination or being interviewed, or undertaking various activities to set up a business.

Thus, the EU includes those who only study job advertisements as unemployed (if they pass the other screens) – this is true in Canada as well – while the United States does not consider reading ads as active job search. In addition, persons waiting to start a new job are considered unemployed in Europe, but they are not considered unemployed in the United States unless they

have actively searched for work within the previous four weeks. Persons on temporary layoff are considered unemployed in the United States even if they do not seek work, but in Europe they may be counted as employed, unemployed, or inactive, depending on search activity and the strength of their attachment to their job (based on pay and/or a definite recall date within three months).

The Dividing Line Between Being Unemployed and Out of the Labour Force

Because the distinction is necessarily arbitrary for measurement purposes, a research literature has investigated the dividing line between the unemployed and those out of the labour force. One strand of this literature focuses on transition probabilities among labour market states, investigating the degree to which those out of the labour force (or a marginally attached subset of them) are as likely to move into employment as those labelled unemployed. The issue is that measured unemployment might miss some signals of labour market tightness or slack if those out of the labour force behave similarly to the unemployed. Clark and Summers reported that 'many of those classified as not in the labor force are functionally indistinguishable from the unemployed' (1979, p. 31). Flinn and Heckman (1983), by contrast, examined young workers' transition probabilities into employment and rejected the hypothesis that the distinction between unemployment and being out of the labour force is behaviourally meaningless.

Several recent papers – including Jones and Riddell (1999, 2006) focusing on Canada, Garrido and Toharia (2004) for Spain, Brandolini et al. (2006) for Italy, and Schweitzer (2003) for the UK – describe selected groups of individuals who are officially out of the labour force but might be considered close to the unemployed, as they are attached to the labour force in various ways. These authors test hypotheses regarding behavioral differences – most notably transition probabilities to employment over various time horizons – between the unemployed and subgroups of those out of the labour force. For example, Jones and Riddell consider those who say they want a job; Brandolini, Cipollone and Viviano examine those who searched for employment between five and eight weeks before the survey; Garrido and Toharia examine 'passive' job searchers; Schweitzer subdivides those available for or wanting a job according to their primary non-labour market activity. Jones and Riddell also subdivide the unemployed into several categories along similar dimensions, and Schweitzer distinguishes the long-term and short-term unemployed; they examine rates of transition to employment of these unemployed subgroups as well.

These researchers find that most of the marginally attached categories lie between the unemployed and the remainder of the 'inactive' group; that is, their transitions into employment are higher than the completely inactive, but still generally lower than those of the unemployed. Jones and Riddell and Schweitzer also find heterogeneity within the ranks of the unemployed and note that some marginally attached categories have higher transition rates into employment than selected subcategories of the unemployed.

Measuring Labour Market Slack

These measurement issues are of more than academic interest because unemployment is the most widely used indicator of the degree of tightness or slack in the labour market and, by extension, in the overall economy; as a consequence, it is used by policymakers as a key signal of potential inflationary pressures. The research discussed above points to a continuum of labour market attachment among the non-employed, from those classified as unemployed through various marginally attached groups to people who expressly do not want a job. Some of the research authors argue that unemployment should be defined and measured more inclusively than it is currently. More generally, the arbitrariness of the dividing line between the states of being unemployed and out of the labour force, together with heterogeneity among subgroups within the inactive population, suggest that policymakers might gain useful information by looking at a range of measures – along with the

official unemployment rate – in judging the state of the labour market.

Because the relationship between the measured unemployment rate and 'true' economic slack and hence inflation may vary, depending on the specific definitions used in measuring unemployment, potential labour market entrants, the age and gender composition of the population, and labour market institutions, researchers have developed and investigated a variety of alternative indicators of labour market slack. One set of alternative measures sidesteps the difficulty of choosing a dividing line between the unemployed and inactive population by concentrating on the distinction between employment and non-employment. For example, the European Council revised its labour market targets in 2000, replacing the goal of reducing unemployment with the goal of increasing employment rates (employment–population ratios) (European Parliament 2000). Similarly, Juhn et al. (1991, 2002), and Murphy and Topel (1997) analyse non-employment and argue that 'the unemployment rate has become progressively less informative about the state of the labor market' (1997, p. 295). Others consider the labour force participation rate an indicator of interest along with the unemployment rate (for example, Anderson et al. 2005; Bradbury 2005). Complementary approaches consider a variety of direct measures of labour market tightness, either individually (for example, Shimer's 2005 job-finding rate among the unemployed) or in combination (for example, a composite measure of US labour market tightness compiled by Barnes et al. 2007).

See Also

- ▶ Labour Supply
- ▶ Natural Rate of Unemployment
- ▶ Unemployment

Bibliography

Anderson, K., L. Barrow, and K.F. Butcher. 2005. Implications of changes in men's and women's labor force participation for real compensation growth and inflation. *Topics in Economic Analysis & Policy* 5-(1) (Article 7).

Barnes, M., R. Chahrour, G. Olivei, and G. Tang. 2007. A principal components approach to estimating labour market pressure and its implicaitons for inflation. Federal Reserve Bank of Boston Public Policy Brief Series, No. 07-2.

Bradbury, K. 2005. Additional slack in the economy: The poor recovery in labor force participation during this business cycle. Public Policy Brief No. 05–2. Boston: Federal Reserve Bank of Boston.

Brandolini, A., P. Cipollone, and E. Viviano. 2006. Does the ILO definition capture all unemployment? *Journal of the European Economic Association* 4: 153–179.

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

European Parliament. 2000. Lisbon European Council 23 and 24 March 2000 presidency conclusions. http://www.europarl.europa.eu/summits/ lis1_en.htm. Accessed 20 Dec 2006.

Flinn, C.J., and J.J. Heckman. 1983. Are unemployment and out of the labor force behaviorally distinct labor force states? *Journal of Labor Economics* 1: 28–42.

Garrido, L., and L. Toharia. 2004. What does it take to be (counted as) unemployed: The case of Spain. *Labour Economics* 11: 507–523.

Jacobs, E.E., ed. 2006. *Handbook of U.S. labor statistics*. 9th ed. Lanham: Bernan Press.

Jones, S.R.G., and W.C. Riddell. 1999. The measurement of unemployment: An empirical approach. *Econometrica* 67: 147–161.

Jones, S.R.G., and W.C. Riddell. 2006. Unemployment and nonemployment: Heterogeneities in labor market states. *Review of Economics and Statistics* 88: 314–323.

Juhn, C., K.M. Murphy, and R. Topel. 1991. Why has the natural rate of unemployment increased over time? *Brookings Papers on Economic Activity* 1991(2): 75–142.

Juhn, C., K.M. Murphy, and R. Topel. 2002. Current unemployment, historically contemplated. *Brookings Papers on Economic Activity* 2002(1): 79–116.

Murphy, K.M., and R. Topel. 1997. Unemployment and nonemployment. *American Economic Review* 87: 295–300.

Schweitzer, M. 2003. Ready, willing, and able? Measuring labour availability in the UK. Working Paper No. 03-03. Cleveland: Federal Reserve Bank of Cleveland.

Shimer, R. 2005. Reassessing the ins and outs of unemployment. Mimeo, Department of Economics, University of Chicago.

Sorrentino, C. 2000. International unemployment rates: How comparable are they? *Monthly Labor Review* 123: 3–20.

U.S. Bureau of Labor Statistics. 2006. How the government measures unemployment. http://ww.bls.gov/cps/cps_htgm.htm. Accessed 20 Dec 2006.

Unequal Exchange

Ednaldo Araquem Da Silva

Marxists have long attempted to explain the uneven development of 'productive forces' (labour productivity) and the resulting income differences in the world capitalist economy primarily by means of the 'surplus drain' hypothesis (see Emmanuel 1972; Andersson 1976). Adopting Prebisch's division of the world capitalist economy into the 'centre' and 'periphery', Marxists have argued that surplus transfer has restrained the economic development of the periphery and exacerbated its income gap vis–àvis the centre.

Before Emmanuel's work, the surplus transfer argument consisted of a loose intertwining of Prebisch's thesis over the secular deterioration of the terms of trade in the periphery, Marx's writings on 'the colonial question', and Lenin's theory of imperialism. Although presented inelegantly in terms of Marx's tableaux, Emmanuel introduced a coherent surplus drain theory utilizing Marx's transformation of values into production prices.

Emmanuel (1972) formulated his theory of surplus transfer through unequal exchange by comparing values with Marxian prices of production (see Okishio 1963, pp. 296–8). Subsequently, Braun (1973) introduced unequal exchange utilizing Sraffa's framework (see Evans's 1984, critical survey), Bacha (1978) introduced a neoclassical counterpart, and Shaikh (1979) suggested an alternative preserving Marx's theory of value.

Departing from recent reformulations, it is helpful to explain Emmanuel's unequal exchange theory within its original Marxist framework. The value (t) of a product is the sum of constant capital (c), variable capital (v), and surplus value (s), whereas its corresponding Marxian production price (p) includes the average profit rate (r):

$$t = c + v + s \tag{1}$$

$$p = (1 + r)(c + v) \tag{2}$$

In a world capitalist system consisting of the centre (A) and periphery (B) as trading partners, unequal exchange is defined as the difference (g) between Marxian production prices and values (see Marelli 1980, p. 517). In fact, unequal exchange compares two terms of trade under different assumptions about the wage rate in each country:

$$g_i = p_i - t_i \quad i = \mathrm{A}, \mathrm{B} \tag{3}$$

A positive g denotes a *surplus gain* for exporters, while a negative g denotes a *surplus loss*.

Emmanuel's theory rests on the assumptions of a single world-wide profit rate resulting from international capital mobility, and the existence of a wage gap resulting from the immobility of labour from the periphery to the centre. The wage rate is an independent variable. Based on these assumptions, Emmanuel showed that unequal exchange depends on a country's rate of surplus value and on its organic composition of capital in relation to world average. Subtracting (1) from (2), we obtain:

$$g_i = r(c_i + v_i) - s_i \tag{4}$$

now consider these definitions:

(a) $s_i = e_i v_i$	rate of surplus value,
(b) $r = e/(1 + k)$	average profit rate,
(c) $c_i = \mathrm{K}_i v_i$	organic composition of capital.

After substituting the definitions for the rate of surplus value, the average profit rate, and the organic composition of capital into equation (4), we obtain a formula to measure unequal exchange:

$$g_i = v_i \left\{ e \frac{1 + k_i}{1 + k} - e_i \right\}. \tag{5}$$

Unequal exchange will disappear when the profit rate of the centre or the periphery approaches the world average profit rate, i.e. $r_i = r$. This is satisfied when these conditions hold:

$$\text{(i) } e_i = e \text{ and (ii) } k_i = k.$$

Emmanuel's distinction between the *broad* and *strict* definitions of unequal exchange can be easily understood by referring to equation (5). Even when the wage rates and thus the rates of surplus are equalized between the centre and the periphery, unequal exchange in the 'broad sense' occurs resulting from differences in the organic composition of capital. This type of unequal exchange can also exist *within* a country because of the differences in the organic composition of capital among sectors.

If condition (i) is satisfied and the rates of surplus value in the centre and periphery are equalized, the unequal exchange equation (5) becomes:

$$g_i = v_i e\left\{\frac{1+k_i}{1+k} - 1\right\}. \qquad (5')$$

As a result, there will be a surplus gain through trade when the individual organic composition of capital exceeds the world average. Likewise, if condition (ii) is satisfied and the organic compositions of capital are equal in both the centre and the periphery, the unequal exchange equation (5) becomes:

$$g_i = v_i(e - e_i). \qquad (5'')$$

In this case, corresponding to Emmanuel's unequal exchange in the 'strict sense', there will be a surplus gain through trade when the world average rate of surplus value exceeds the individual rate.

The periphery tends to transfer surplus through trade because its rate of surplus value is higher than the world average, resulting from an international wage gap favouring workers in the centre. Therefore, even if the organic compositions of capital are equalized, unequal exchange results from the existence of a wage gap between the centre and the periphery, expressed as the rate of surplus value being lower in the centre than in the periphery (the rate of surplus value can be expressed as one over the value of labour power or 'wage share' minus one, $e = (1/w)–1$). According to Emmanuel, unequal exchange in the 'strict sense' characterizes the trade relations between the centre and periphery.

Emmanuel's (1972, p. 61) basic conclusions is that 'the inequality of wages as such, all other things being equal, is alone the cause of the inequality of exchange'. As a corollary, Emmanuel (1972, p. 131) argued that 'by transferring, through non-equivalent [exchange], a large part of its surplus to the rich countries, [the periphery] deprives itself of the means of accumulation and growth'. Thus, an important implication of Emmanuel's theory is that a widening wage gap leads to a deterioration of the periphery's terms of trade, and a subsequent reduction in its rate of economic growth.

Emmanuel's work generated an interesting international debate. One contentious issue is the relationship of Emmanuel's theory to Marx's theory of value, leading to reformulations of Emmanuel's theory within the context of the Marx–Sraffa debate (Gibson 1980; Mainwaring 1980; Dandekar 1980; Evans 1984; Sau 1984). Another view holds that Emmanuel's theory does not sufficiently explain uneven development because it omits the 'blocking of the productive forces' by entrenched and reactionary social classes in the periphery (Bettelheim, in the Appendix to Emmanuel 1972). Bettelheim also argues that the rate of surplus value is higher in the centre resulting from its higher labour productivity, thus giving rise to unequal exchange reversal.

At the same time, Amin (1977) has emphasized non-specialized trade between the centre and the periphery, claiming the 'end of a debate', while the debate survived a virulent 'exchange of errors' among Marxists in India (see Dandekar 1980; Sau 1984). De Janvry and Kramer (1979) criticize unequal exchange as a theory of underdevelopment because capital mobility tends to eliminate wage differences by exhausting the 'reserve army' in the periphery, an argument which is challenged by Gibson (1980). Andersson (1976) surveys some pre-Emmanuel views, adding a formalization similar to Braun (1973), while Liossatos (1979) and Marelli (1980) have recast Emmanuel's theory in a modern, Morishima-like Marxian framework.

Although Emmanuel's primary objective involves 'model building', it is important to recognize that his references to standard trade theory are dated, largely confined to the literature of the 1950s, perhaps indicating that his work suffered

from a long gestation period. Therefore, one should be cautious about treating Emmanuel's work as a critique of standard trade theory. Outside of Ricardian and Marxian circles, the reception of Emmanuel's work has been tepid if not neglectful.

Looking ahead, Harris (1975) suggests that a convincing theory of economic development should include a theory of value and distribution and a theory of accumulation on a world scale. Emmanuel's theory of unequal exchange, especially in subsequently more rigorous formulations (Andersson 1976; Liossatos 1979; Marelli 1980; Gibson 1980; Evans 1984; Sau 1984) has an assured place in this curriculum. In this way, Emmanuel's theory of unequal exchange is definitely linked to the original theory of Prebisch, Singer, Lewis and Baran, on trade and development.

See Also

► Periphery

Bibliography

Amin, S. 1977. *Imperialism and unequal development*. New York: Monthly Review Press.

Andersson, J. 1976. *Studies in the theory of unequal exchange between nations*. Abo: Abo Akademi.

Bacha, E. 1978. An interpretation of unequal exchange from Prebisch–Singer to Emmanuel. *Journal of Development Economics* 5(4): 319–330.

Braun, O. 1973. *International trade and imperialism*. Atlantic Highlands: Humanities Press, 1984.

Dandekar, V. 1980. Unequal exchange of errors. *Economic and Political Weekly* 15(13): 645–48. Continued in 16(6): 205–212.

De Janvry, A., and F. Kramer. 1979. The limits of unequal exchange. *Review of Radical Political Economics* 11(4): 3–15.

Emmanuel, A. 1972. *Unequal exchange: A study of the imperialism of trade* (with additional comments by Charles Bettelheim). New York: Monthly Review Press.

Evans, D. 1984. A critical assessment of some neo-Marxian trade theories. *Journal of Development Studies* 20(2): 202–226.

Gibson, B. 1980. Unequal exchange: Theoretical issues and empirical findings. *Review of Radical Political Economics* 12(3): 15–35.

Harris, D. 1975. The theory of economic growth: A critique and reformulation. *American Economic Review* 65(2): 329–337.

Liossatos, P. 1979. Unequal exchange and regional disparities. *Papers of the Regional Science Association* 45: 87–103.

Mainwaring, L. 1980. International trade and the transfer of labour values. *Journal of Development Studies* 17(1): 22–31.

Marelli, E. 1980. An intersectoral analysis of regional disparities in terms of transfers of surplus value. *Revista internazionale di scienze economiche e commerciali* 27(6): 507–526.

Okishio, N. 1963. A mathematical note on Marxian theorems. *Weltwirtschaftliches Archiv* 91(2): 287–298.

Sau, R. 1984. *Underdeveloped capitalism and the general law of value*. Atlantic Highlands: Humanities Press.

Shaikh, A. 1979. Foreign trade and the law of value: Part I. *Science and Society* 43(3): 281–302. Part II was published in 44(1).

Uneven Development

Donald J. Harris

Abstract

A striking characteristic of capitalist development is the phenomenon of uneven development, defined as persistent differences in levels and rates of economic development between different sectors of the economy. However, much existing economic theory predicts that many observed features of differentiation would tend to wash out as a result of competitive market forces. This article seeks to bridge this gap. It proposes a strategy for the analysis of uneven development that advances toward a historically and empirically relevant theory. The analysis draws in part on elements of the emerging paradigm of neo-Schumpeterian evolutionary theory and on some documented empirical regularities.

Keywords

Aggregation (theory); Bounded rationality; Competition; Concentration; Creative accumulation; Creative destruction; Differentiation among firms; Diffusion of technology;

Endogenous growth; Evolutionary economics; Firm, theory of; First-mover advantages; Growth centres; Harrod–Domar growth model; Industry evolution; Innovation; Invention; Irreversible investment; Learning; Life-cycle of industry; Market failure; Saturation effect; Schumpeterian competition; Shift effect; Technical change; Underdevelopment; Uneven development

JEL Classifications

B25; B41; D2; D4; E12; F12; L1; L2; L6; N10; N90; O1; O3; O4; O5; R11; R12

In examining the general character of the process of capitalist development as it has appeared historically across many different countries over a long period of time, it emerges that one of its most striking characteristics is the phenomenon of uneven development. Specifically, the process is marked by persistent differences in levels and rates of economic development between different sectors of the economy.

This differentiation appears at many levels and in terms of a multiplicity of quantitative and qualitative indices (Kuznets 1966; Maddison 1982; Mueller 1990; Pritchett 1997; Salter 1966). Relevant measures that sharply identify the phenomenon include the level of labour productivity in different sectors, the level of wages, occupational and skill composition of the labour force, the degree of mechanization and vintage of production techniques, rates of profit, rates of growth, and the size structure of firms. The phenomenon appears regardless of the level of aggregation or disaggregation of the economy, except for the extreme case of complete aggregation – in which case, structural properties of the economy are made to disappear. For example, it appears at the level of comparing the broad aggregates of manufacturing industry and agriculture, at the level of individual industries within the manufacturing sector, and at the level of individual firms in an industry. It appears on a regional level within national economies as well as on a global scale between different national economies. In this latter context, one form taken is the continued differentiation between underdeveloped and advanced economies, usually identified as the problem of underdevelopment.

These disparities appear from observation of the economy as a whole at any given moment and over long periods of time. While the relative position of particular sectors may change from one period to another, there is, nevertheless, always a definite pattern of such differentiation. We may say, therefore, and certainly it is an implication of these observations, that these disparities are continually reproduced by the process of development. Uneven development, in this sense, is an intrinsic or inherent property of the economic process. Far from being merely transitory, it appears to be a pervasive and permanent condition.

Now, it is an equally striking fact that, when we examine the theoretical literature on economic growth, we find the completely opposite picture. In particular, the dominant conception of the growth process that has characterized the post-Second World War literature is constructed in terms of uniform rates of expansion in output, productivity and employment in all sectors of the economy. In this sense, it is largely a literature of steady-state growth, whether presented in multi-sectoral or aggregative models (Burmeister and Dobell 1970; Harris 1978). Some notable and relevant exceptions, including Haavelmo (1954), Leon (1967), Nelson and Winter (1982), Pasinetti (1981), Salter (1965), explicitly examine aspects of the problem of persistent differentiation posed here. The recent flurry of work in endogenous growth theory seeks to incorporate some relevant elements of the problem into the neoclassical conception of the growth process (Aghion and Howitt 1998). However, much of existing economic theory predicts that, given enough time, many of the features of differentiation which we observe empirically would tend to wash out as a result of the operation of competitive market forces (Harris 1988). Such differentiation should therefore be viewed only as a transitory feature of the economic process.

Thus, on the one side, we find a historical picture of uneven development as a persistent phenomenon, and on the other, a theory that

essentially negates and denies this fact. It is possible to go some of the way towards bridging this gap. Accordingly, I consider here a strategy for analysis of uneven development that breaks through the narrow limits of the existing steady-state theory and advances towards a historically and empirically relevant theory.

The Analytics of Uneven Development

It is necessary to start by recognizing the intrinsic character of the individual firm as an expansionary unit of capital with a complex organization. Various efforts have been made to develop a theory of the firm on this basis. (See, for instance, Penrose 1959; Baumol 1967; Marris 1967; Winter 2006). In this conception, growth is the strategic objective on the part of the firm. This urge to expand is not a matter of choice. Rather, it is a necessity enforced upon the firm by its market position and by its existence within a world of firms where each must grow in order to survive. It is reinforced also by sociological factors. It is this character of the firm that constitutes the driving force behind the process of expansion of the economy.

In the aggregate, the global economy is conceived to consist of an ordered system of firms (an interlocking network of individual circuits of capital) and its sectors (classified variously as industries, regions, national economies) likewise to be clusters of the firms that are the component units of this system. In this system, it is the firms that compete, not industries or regions, national economies or 'North' versus 'South'. The state sets the rules and jointly determines the external conditions (externalities) within which the firms operate.

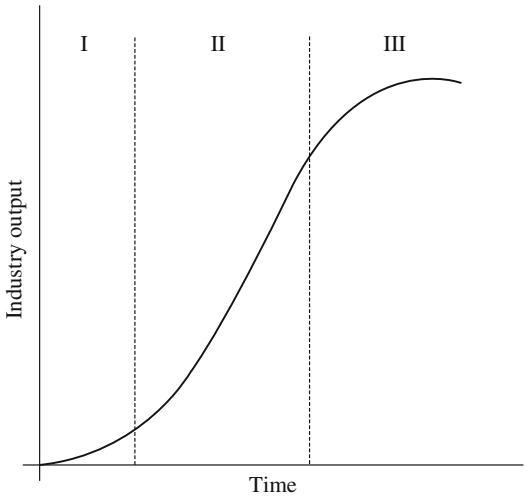
This is a crucial starting point because it establishes the idea of growth as the outcome of a process driven by active agents, not by exogenous factors. In particular, in the context of the capitalist economy, growth is the outcome of the selfdirected and self-organizing activity of firms, each of which seeks to expand and improve its competitive position in relation to the rest. Once this principle is recognized it becomes possible to move towards an understanding of the problem of uneven development.

The imperative of growth impels the firm constantly to seek new investment opportunities wherever they are to be found. Such opportunities may lie within a wide range – in existing product lines, in new products and processes, in new geographical spaces and natural resource frontiers, or in the take-over of existing firms. However, at the core of this movement, viewed historically over the long term, are the invention, innovation and diffusion of new technologies that give rise to new products and services (Freeman 1982; Landes 1969, 1999; Marx 1906, ch. 15; Mokyr 1990, 2002).

The emergence of growth centres or leading sectors is a reflection of this underlying process. It is a consequence of the effort on the part of many firms to create or to rush into those spheres where a margin of profitability exists that allows them to capture new profit and growth opportunities. It may be conceived to take the form of a 'swarm' (Schumpeter 1934, p. 223) or 'contagion' (Baumol 1967, p. 101), marked by both entry and exit of firms. Such spheres are opened up, typically through complementary 'macroinventions' and 'microinventions' (Mokyr 1990, p. 13) and in a sporadic and discontinuous pattern, as a consequence of the ongoing investment and innovative activity of firms and the competitive interactions among them. It is this constant flux, consisting of the emergence of new growth centres, their rapid expansion relative to existing sectors, and the relative decline of others, that shows up in the economy as a whole as uneven development.

The Process of Industry Evolution

The form of this process, as it appears at the level of particular industries and products, has been identified in terms of certain empirical regularities, though there are also significant variations across industries and products. Studies show that, with some exceptions, the growth of many new industries and products follows a life cycle pattern (Gold 1964; Gort and Klepper 1982; Klepper 1997; Klepper and Graddy 1990; Mullor-Sebastian 1983; Wells 1972). It may be represented schematically by an S-shaped curve of the time-path of output as in Fig. 1. (For



Uneven Development, Fig. 1 Life cycle of an industry

simplicity, no distinction is made here between products and processes, an industry is assumed to produce a single product, and short-term turbulence in the path of output is ignored.) Accordingly, we may distinguish three phases of expansion: I, the initial phase, where total output is a minute share of aggregate output and grows at a low rate; II, a phase of rapid growth in which output expands rapidly and its share of aggregate output grows; III, the sector reaches a threshold beyond which its growth rate tends to level off and perhaps to decline.

To characterize the associated pattern of technological innovation, Kuznets (1979) identifies a sequence of four distinct phases constituting the life cycle of 'major' innovations: (1) the *preconception* phase in which necessary scientific and technological preconditions are laid; (2) the phase of *initial application* involving the first successful commercial application of the innovation; (3) the *diffusion* phase marked by spread in adoption and use of the innovation along with continued improvements in quality and cost; (4) the phase of *slowdown* and *obsolescence* in which further potential of the innovation is more or less exhausted and some contraction may occur. This taxonomy is not all-embracing, and there are others that emphasize other features, but it is suggestive in pointing to a certain internal logic of the innovation process.

The process of industry evolution is also typically associated with a changing firm-structure of the industry. In many industries, there is a proliferation of small firms in phase I. As the diffusion of the product occurs and growth speeds up, there is a 'shaking out' process by which many of the smaller firms disappear (exit) and the available market is concentrated in the remaining firms. When the industry reaches 'maturity', in phase III, there is likely to be a high degree of concentration. This association between industry life cycle and changing firm-structure (commonly called 'co-evolution') suggests that the dynamic of expansion through innovation is simultaneously a process of the concentration of capital.

This sequence of a single product-cycle, schematically described here, is but a small segment of the time sequence characterizing the historical evolution of the economy. Given that firms are growing, making profits, and seeking to continue to grow, it must be supposed that at least some of them, having entered into phase III, would seek to launch into new investment opportunities. They will therefore actively seek new products that will initiate a corresponding new sequence. Alternatively, the new sequence could come from entry of new start-up firms.

It follows that we can map out the dynamic evolution of the economy as *a sequential process* that is discontinuous, punctuated and stochastic, with varying and overlapping time-scales of the different product-cycles, where the overall growth is accountable for on the basis of (1) the individual growth of particular new products coming on stream, (2) the growth of pre-existing products, each of which is growing at a different rate depending on the particular phase reached in its life cycle, and (3) over time the irregular accretion of new products as the innovation process continues.

In this context, the relative position of any firm-cluster (region or national economy) at any time on a relevant index of development may be seen as a matter of the particular products it has managed to capture as a result of the previous pattern of accumulation, the ongoing activity of firms operating within it and the particular timing of their entry into the life cycle of new products.

The causes that produce and sustain the observed patterns of differentiation must then be found within the internal dynamics of this process, leaving aside such historically contingent factors as wars, colonial control, 'foreign' intervention, that may also be considered relevant and important. What role is to be assigned to demand as a factor in this process? At the level of individual consumer products or industries, a common conception is that demand acts as an autonomous factor with a definite influence on the life-cycle pattern of evolution of the product. That influence is exerted in the early phase of introduction of a new product because of an element of resistance due to 'habit' formed in a customary pattern of consumption. It is exerted also in the maturity phase because of the operation of 'saturation effects' in consumption. But there are reasons to doubt the strength and effectiveness of such factors, as well as their supposed autonomy.

First, in an economy undergoing regular and rapid change, it is not evident what role there is for habit except for the habit of change itself. The experience of, and adaptation to, change may create a high degree of receptivity to change. What then becomes decisive in the evolution of demand (for consumer goods) is the growth of income, and the changing relative price and quality of products. Income and price elasticities of demand are an imperfect, proximate expression of this dynamic effect.

Second, in so far as these latter factors are crucial to the formation of demand, it may be argued that there is a certain self-fulfilling aspect of the expansionary process at the level of industry demand. In particular, investment generates the demand that provides the market for the new products which investment itself creates. This occurs in two ways. First structural interdependence in the economy at the level of both production and expenditure patterns allows for the possibility of a certain mutual provisioning of markets when expansion takes place on a broad front. Second, as a new product unfolds through the stages of the innovation process, it undergoes both improvements in quality and a decline in price relative to other products. This development provides a substantive basis for making inroads into the market for existing closely related products and hence promotes demand through a shift from 'old' to 'new' products. It is perhaps this shift effect which is mistakenly identified as a saturation effect by adopting a one-sided and static view of a dynamic and interdependent process.

Each and every individual firm must of course secure a place in the market for its product. Its success in this regard is dependent on its own efforts and capabilities.

Competition, Firm Capabilities, Entry/Exit Conditions, and the Social Environment

Analytical treatment (including formal modelling) of the process of industry evolution has flourished since the 1980s in tandem with an outpouring of empirical studies covering different industries, countries and time periods. Much of this work is done within the frame of an emerging paradigm in the Schumpeterian tradition of evolutionary dynamics (Futia 1980; Iwai 1984a, b; Nelson and Winter 1982; Dosi 1984) and there are other theoretical approaches (Loury 1979; Dasgupta and Stiglitz 1980; Durlauf 1993). For a review of the current state of the art and challenges for research, focusing on the evolutionary approach, see Malerba (2006). Relevant for the present purposes are the significant insights provided so far by this work into the mechanisms and causal factors that govern the process of industry evolution and account for the persistence of differentiation among firms.

The neo-Schumpeterian approach develops an explicit formulation of 'Schumpeterian competition' in which firms innovate to win super-normal profits, profits are reinvested to provide further growth through innovation and market expansion, and there are winners and losers due to the operation of selection mechanisms and learning mechanisms. Decisions are typically based on bounded rationality. It is shown that such competitive behaviour under specified conditions gives rise to persistent differentiation among firms in terms of size, productivity, costs of production, product characteristics, profitability and growth and may

breed long-term sustainable market concentration among surviving firms, with or without entry. Economies of scale and scope are not a necessary part of the story; a key factor is increasing returns to knowledge and learning. Though there exists a strong tendency to concentration, it is not inevitable, and depends on industry characteristics that vary across industries. There also exist dual tendencies of 'creative destruction' and 'creative accumulation'.

A distinctive feature of this approach is the conception of the firm itself as an organizational unit. The firm is conceived as the embodiment of a set of strategic assets (competences or capabilities), tangible and intangible, consisting of knowledge, skills, and routines, gained through path-dependent experience and learning, that are specific to each firm and non-tradeable. These assets evolve over time (through 'competence accumulation') with the ongoing process of evolution of the industry and through interaction with the changing environment. Consequently, diversity among firms is not only a characteristic of the system of firms, it is also reproduced by the evolutionary dynamics of the competitive process.

Some key factors determining the evolutionary path of industry structure in terms of firm composition are the following. (1) First- (second-, third-) mover advantages arising from a combination of unique internal attributes of the mover, product characteristics, network effects among users, and random chance events. (2) Non-pecuniary network externalities associated with cues and information gained from interacting with the 'local' social environment of firms, users of the product, and institutions involved in knowledge creation and information dissemination (on the national level, the 'national system of innovation'). (3) Spillover effects among firms and across industries, which may be both positive and negative. (4) Increasing returns to knowledge and learning within the firm. (5) A firm may become 'lockedin' to its own trajectory of technology development and reap increasing returns therefrom, but eventually suffer a disadvantage from generating irreversibilities and inertia causing inability to adjust to change ('success breeds failure'). (6) The very same factors that confer advantages upon early entrants and incumbent firms may create barriers to entry for 'latecomers', depending on the stage of industry evolution and timing of entry.

Some relatively neglected factors that need to be integrated into a more comprehensive analysis include: (1) the role of market demand, as related to the mutual interaction between producers and users (consumers, other firms, and the state); (2) the role of the financial system (Schumpeter had assigned a crucial role to the granting of credit 'as an order on the economic system to accommodate itself to the purposes of the entrepreneur' (1934, p. 107)); (3) workplace and labour market interactions, lightly touched upon by Mansfield (1968, ch. 5) and vividly described in historical detail by Braverman (1974); (4) the system of governance by the state, that sets and enforces the rules and norms, including property rights, governing conduct by firms.

Within this extended framework of analysis, it is possible to explain not only how some firms (or firm-clusters) come to capture the position of leaders (and may eventually lose it to others), but equally how some are left behind, others drop out altogether (exit), and still others remain on the 'periphery' (so to speak) lacking the internal and external capabilities to enter. In this regard, the explanatory power of this analysis is readily applicable to commonly discussed empirical and historical phenomena such as 'deindustrialization', 'catching up' (convergence), and 'falling behind' (divergence).

What emerges from this analysis also is an understanding of the critical role of public policy and programmes to foster economic development. Because of the pervasiveness of externalities and various forms of coordination problems, market failures are intrinsic to the process, calling thereby for collective intervention to achieve efficiency and socially optimal results.

The Aggregation Problem

All the preceding analysis concerns the pattern of industrial growth viewed at the level of an

individual industry and the firms (or firm-clusters interpreted as, say, regions or countries) that compose it. There is nothing in this analysis to indicate how the pattern of growth of different industries translates into aggregate expansion at the level of the economy as a whole, or how the various industrial patterns fit together to form a complete whole. This is a substantive problem requiring further analytical treatment on its own terms. Its significance derives from recognition that the economy as a whole is not just the sum of its parts. Hence, the motion of the economy cannot simply be deduced from the movement of its parts. The usual methodological device of the 'representative firm' necessarily fails in the present context.

A related aspect of the problem is associated with the manifold and complex ways in which growth in one sector (however defined) mutually conditions and is conditioned by growth in other sectors. Such mutual interaction is a necessary consequence of economic interdependence in both production and exchange. (Hence, models of international trade that claim to show uneven development arising uniquely from exchange of products give a one-sided representation of the problem.) The existence of such interaction implies that there is a certain cumulative effect intrinsic in the growth process. Understanding the exact mechanisms through which this effect operates is one of the central analytical problems for the analysis of uneven development.

There is no guarantee that in the aggregate there is always sufficient demand for all products. It is here that the analysis comes full circle, back to the problem of overall effective demand that motivated the early post-war growth theory initiated by Harrod (1948) and Domar (1957). This problem was a central focus of the analysis in the Keynesian and Post Keynesian tradition, less so in the case of the neoclassical tradition (as detailed in Harris 1985). It appears now that it cannot be escaped in making the transition to the analysis of uneven development.

The analytical framework presented here lays the groundwork for addressing this larger set of problems.

See Also

- ▶ Development Economics
- ▶ Economic Growth, Empirical Regularities in
- ▶ Endogenous Growth Theory
- ▶ Schumpeterian Growth and Growth Policy Design

Bibliography

Aghion, P., and P. Howitt. 1998. *Endogenous growth theory*. Cambridge, MA: MIT Press.

Baumol, W.J. 1967. *Business behaviour, value and growth*, revised ed. New York: Harcourt, Brace and World.

Braverman, H. 1974. *Labor and monopoly capital, the degradation of work in the twentieth century*. New York: Monthly Review Press.

Burmeister, E., and A.R. Dobell. 1970. *Mathematical theories of economic growth*. London: Macmillan.

Dasgupta, P., and J. Stiglitz. 1980. Industrial structure and the nature of innovative activity. *Economic Journal* 90: 266–293.

Domar, E.D. 1957. *Essays in the theory of economic growth*. New York: Oxford University Press.

Dosi, G. 1984. *Technical change and industrial transformation*. London: Macmillan.

Durlauf, S.N. 1993. Nonergodic economic growth. *Review of Economic Studies* 60: 349–366.

Freeman, C. 1982. *The economics of industrial innovation*, 2nd ed. Cambridge, MA: MIT Press.

Futia, C.A. 1980. Schumpeterian competition. *Quarterly Journal of Economics* 94: 675–695.

Gold, B. 1964. Industry growth patterns: Theory and empirical results. *Journal of Industrial Economics* 13: 53–73.

Gort, M., and S. Klepper. 1982. Time paths in the diffusion of product innovations. *Economic Journal* 92: 630–653.

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

Harris, D.J. 1978. *Capital accumulation and income distribution*. Stanford: Stanford University Press.

Harris, D.J. 1985. The theory of economic growth: From steady states to uneven development. In *Contemporary issues in macroeconomics and distribution*, ed. G. Feiwel. London: Macmillan.

Harris, D.J. 1988. On the classical theory of competition. *Cambridge Journal of Economics* 12: 139–167.

Harrod, R.F. 1948. *Towards a dynamic economics*. London: Macmillan.

Iwai, K. 1984a. Schumpeterian dynamics, part I: An evolutionary model of selection and imitation. *Journal of Economic Behavior and Organization* 5: 159–190.

Iwai, K. 1984b. Schumpeterian dynamics, part II: Technological progress, firm growth and 'economic selection'.

Journal of Economic Behavior and Organization 5: 321–351.
Klepper, S. 1997. Industry life cycles. *Industrial and Corporate Change* 6: 145–181.
Klepper, S., and E. Graddy. 1990. The evolution of new industries and the determinants of market structure. *RAND Journal of Economics* 21: 27–44.
Kuznets, S. 1966. *Modern economic growth, rate, structure, and spread*. New Haven: Yale University Press.
Kuznets, S. 1979. Technological innovations and economic growth. In *Growth, population, and income distribution, selected essays*. New York: Norton.
Landes, D.S. 1969. *The unbound Prometheus*. Cambridge: Cambridge University Press.
Landes, D.S. 1999. *The wealth and poverty of nations*. New York: Norton.
Leon, P. 1967. *Structural change and growth in capitalism*. Baltimore: Johns Hopkins Press.
Loury, G.C. 1979. Market structure and innovation. *Quarterly Journal of Economics* 93: 395–410.
Maddison, A. 1982. *Phases of capitalist development*. Oxford: Oxford University Press.
Malerba, F. 2006. Innovation and the evolution of industries. *Journal of Evolutionary Economics* 16: 3–23.
Mansfield, E. 1968. *The economics of technological change*. New York: Norton.
Marris, R. 1967. *The economic theory of 'managerial' capitalism*. London: Macmillan.
Marx, K. 1906. *Capital*, vol. I. Chicago: Charles H. Kerr.
Mokyr, J. 1990. *The lever of riches, technological creativity and economic progress*. Oxford: Oxford University Press.
Mokyr, J. 2002. *The gifts of Athena: Historical origins of the knowledge economy.* Princeton: Princeton University Press.
Mueller, D.C. 1990. *The dynamics of company profits: An international comparison*. Cambridge: Cambridge University Press.
Mullor-Sebastian, A. 1983. The product life cycle theory: Empirical evidence. *Journal of International Business Studies* 14: 95–105.
Nelson, R.R., and S.G. Winter. 1982. *An evolutionary theory of economic change*. Cambridge, MA: Harvard University Press.
Pasinetti, L.L. 1981. *Structural change and economic growth*. Cambridge: Cambridge University Press.
Penrose, E.T. 1959. *The theory of the growth of the firm*. Oxford: Blackwell.
Pritchett, L. 1997. Divergence, big time. *Journal of Economic Perspectives* 11(3): 3–17.
Salter, W.E.G. 1965. Productivity growth and accumulation as historical processes. In *Problems in economic development*, ed. E.A.G. Robinson. London: Macmillan.
Salter, W.E.G. 1966. *Productivity and technical change*. Cambridge: Cambridge University Press.
Schumpeter, J.A. 1934. *The theory of economic development*. Cambridge, MA: Harvard University Press.
Wells Jr., L.T. 1972. *The product life cycle and international trade*. Boston: Harvard University Press.
Winter, S.G. 2006. Toward a neo-Schumpeterian theory of the firm. *Industrial and Corporate Change* 15: 125–141.

Unforeseen Contingencies

Barton L. Lipman

Abstract

While unforeseen contingencies – possible events that agents do not think of when planning or contracting – are often said to greatly affect the nature of contracting, we lack useful formal models. Most of the existing models boil down to assuming that agents give zero probability to some events that might actually occur, an approach which is not particularly useful for studying the effects of unforeseen contingencies on contracting.

Keywords

Control rights; Expected utility; Incomplete contracts; Long-term and short-term contracts; Probability; Rationality, bounded; Short-term contracts; Uncertainty; Unforeseen contingencies

JEL Classifications

D8

Many writers have suggested that the nature of contracting, firm structure, and even political constitutions cannot be well understood without taking account of the role of *unforeseen contingencies*. As I explain in more detail below, many definitions are possible, but I will define unforeseen contingencies to be possibilities that the agent does not 'think about' or recognize as possibilities at the time he makes a decision. In virtually any reasonably complex situation, real people do not consider all of the many possible situations that may arise. Because of this,

contracts, for example, typically assign broad categories of rights and obligations rather than calling for very specific actions as a function of what might occur. Similarly, firms are designed to figure out what to do rather than simply being programmed to implement some given set of actions. Finally, laws, especially sweeping ones such as constitutions, are intentionally left vague to allow adaptation to circumstances as they arise.

Unfortunately, while it is easy to find eloquent statements in the economics literature regarding the importance of unforeseen contingencies for understanding the nature of economic and political institutions – see, for example, Hayek (1960); Williamson (1975); or Hart (1995) – there is no agreed formal model. I sketch a few known approaches below, but none of them provides a model that can be used to study these issues.

To make this point concretely, I focus on a particular example of an aspect of contracting that we would want a model of unforeseen contingencies to help us understand, namely, the choice between long-term and short-term contracts. It seems obvious that one of the main advantages of a series of short-term contracts is that it is easier to anticipate the relevant contingencies for the near future than for the distant future. Hence in environments with many unforeseen contingencies relative to the value of long-term contracting, we should expect to see more short-term contracting. I will argue below that none of the models of unforeseen contingencies in the literature can be used to illustrate this simple idea.

Before discussing the approaches taken, it is important to clarify what I mean by unforeseen contingencies. First, as I use the term, unforeseen contingencies are not events that the agent has considered but assigned zero probability. This notion is something standard models deal with perfectly well. More importantly, the existence of such events seems to have little to do with the features of economic and political institutions we believe to be related to unforeseen contingencies. To be concrete, consider the trade-off discussed above between long-term contracts and short-term contracts. If the only sense in which some contingencies in the distant future are not foreseen is that they are given zero probability, then the agents will perceive zero costs to excluding them. Hence they will see no 'foreseeability' advantage to short-term contracts, so a model of unforeseen contingencies based on such a definition cannot say anything interesting about the trade-off.

It is also important to note that the use of the term 'unforeseen' in law is often closer to the zero probability definition than the definition I use here. In particular, legal usage often seems to suggest that a contingency is 'unforeseen' by an agent if it occurred even though the agent gave it 'low' probability *ex ante*. For example, a 1997 US tax law allows a person who sells his home to exclude some of the capital gains from taxation under certain conditions if 'unforeseen circumstances' precipitated the move. The Internal Revenue Service (2006, p. 16) recently issued regulations listing events that would 'count' as such unforeseen circumstances, including divorce, job loss, or multiple births from a single pregnancy. Surely, most homeowners would not be startled to learn that couples sometimes divorce, that job losses can occur, or that a pregnancy could yield triplets, so these circumstances are not 'unforeseen' in the sense used here. Instead, such events are 'unforeseen' in the sense that they were very unlikely *ex ante*, too unlikely to influence the home purchase decision.

While this use of 'unforeseen' is evidently valuable for some purposes, it does not seem to be appropriate to the issues of interest here, though it is closer than the zero probability definition. To see this, consider again the trade-off between long-term and short-term contracts. If 'unforeseen' contingencies are recognized but given low probability, they can still be incorporated into the contract and, in the absence of costs to doing so, will be. Hence, in the absence of contracting costs, again, short-term contracts will have no foreseeability advantages if this is what we mean by foreseeability. On the other hand, if there are costs of writing 'long' contracts, it may be optimal to exclude contingencies with low probability. Hence a sequence of short-term contracts (which delays some of the writing costs) may be better. On the other hand, it is not clear that the advantage of short-term contracts is a gain

in foreseeability so much as it is a delay in writing. See Al Najjar et al. (2006) for a particularly interesting related model.

Turning to models, the idea of how unforeseen contingencies are represented is common to most of the models in the area, though with many variations. (A different approach, which I do not discuss, involves an explicit logic rather than focusing on a state space. See Halpern and Rêgo 2005 for a good example of this approach and an overview of much of this literature.) In standard models of uncertainty without unforeseen contingencies, there is a set of states of the world, say Ω, which represents the uncertainty. A state $\omega \in \Omega$ should be thought of as a specification of every possible circumstance conceivably relevant to the agent's situation. For example, for a firm, a state might specify input prices, demand conditions, technological possibilities, what is going on with its rivals, and so on. Part of what is meant by the phrase 'every relevant circumstance' is that, if we know the state, then we know the exact consequence (profits or utility) the agent receives as a function of whatever course of action he might choose.

To give a concrete example, suppose there are two relevant sources of uncertainty: whether it rains and whether there is a revolution in country X. This gives us four possible states of the world:

$$\Omega = \{(\text{rain, revolution}), (\text{rain, no revolution}), (\text{no rain, revolution}), (\text{no rain, no revolution})\}.$$

Consider an agent who has never considered the possibility of revolution. This agent sees only two possibilities: rain or no rain. That is, the agent has a *subjective state space S*, describing the possibilities as he perceives them, given by

$$S = \{(\text{rain}), (\text{no rain})\}.$$

We can think of the 'state' (rain) as the event {(rain, revolution), (rain, no revolution)} and think of the 'state' (no rain) as analogous. Thus the state space as seen by the agent is actually a partition of the true state space. The variation within an event of this partition is variation that the agent has simply not thought of.

This basic idea appears in numerous forms in the literature. This partition description appears in Ghirardato (2001) and Dekel et al. (2001), among others. A more complex form appears in Li (2006a) and in Heifetz et al. (2006a), both of which allow the possibilities recognized by the agent to vary with the true state of the world. For example, it might be that when the true state is (rain, revolution), the agent recognizes the possibility of the revolution, while if it is (rain, no revolution), he does not.

While this idea for representing knowledge is almost uniformly used in the literature, there is greater variation in the way decision-making is represented. Continuing with the rain and revolution example, suppose the agent has a certain amount of money and can either use it to buy an umbrella or invest it in country X or simply save it. Suppose Table 1 gives the true, objective payoff of the agent as a function of his choice and the real state.

Turning to the agent's perceptions, continue to assume that he sees the possibilities only as rain versus no rain. Intuitively, the consequences of buying the umbrella or saving money are unambiguous since they only depend on this. Putting it differently, these acts are measurable with respect to the agent's awareness, so there seems to be no problem. On the other hand, how does the agent perceive the payoffs to investment?

A number of papers in the literature use models that treat the payoffs to investing in 'states' (rain) and (no rain) as exogenously given (see, for example, Heifetz et al. 2006b; Li 2006b). In a somewhat more restrictive version of the same idea, Modica et al. (1998) assume that if the agent does

Unforeseen Contingencies, Table 1 Objective payoffs

Objective state	Payoff if buy	Payoff if save	Payoff if invest
(rain, revolution)	5	0	−100
(rain, no revolution)	5	0	10
(no rain, revolution)	6	8	−100
(no rain, no revolution)	6	8	10

not foresee some possibility, this means he implicitly assumes some particular resolution of this uncertainty. For example, perhaps the agent implicitly assumes there will not be a revolution, so he sees the payoff to investing in each 'state' as ten. In other words, these approaches come down to treating the agent's view of the available actions as given by Table 2 for some numbers x and x'.

With more than one agent, these models generally do allow the agents to perceive different possibilities and to assign payoffs differently. For example, if two agents have to jointly agree on what to do with their money in the example above, we could assume that one of them perceives only rain versus no rain, while the other perceives only revolution versus no revolution.

While such a model can be useful for some purposes, it does not appear useful for studying the kinds of issues mentioned in my opening paragraph. To see this, note that the default approach of Modica et al. (1998) is identical to a model where the states (rain, revolution) and (no rain, revolution) have zero probability. While other values of x and x' are not as directly interpreted, again, the model is identical to one where the agent believes a revolution is impossible (and may have 'incorrect' beliefs about his payoffs). As argued above, if 'unforeseen' is taken to mean 'zero probability', then short-term contracts do not have foreseeability advantages over long-term contracts. Hence, at least for the purposes of studying the trade-off between short-term and long-term contracts, this model of decision-making does not appear to be useful.

Part of what this approach omits is recognition by an agent that his conception of the world is incomplete. Intuitively, what we need to understand the postulated trade-off between short-term and long-term contracts is a recognition by the agent that his conception of what the world will be like in 2016 will be clearer in 2015 than in 2006.

To state this more concretely in the context of the example, the agent might perceive only the possibility of rain versus no rain, but understand that this omits many currently unforeseen possibilities. How might we represent such a situation? One approach is to separate the agent's uncertainty about what events may occur in the world from his uncertainty about what his payoff will be given a particular action. In the story above, we said that the agent's payoff to investing depends on whether it rains and whether there is revolution. Presumably, the agent does not care about rain or revolution per se but instead cares about what utility or payoff he receives. That is, the 'states' as the agent perceives them may be more usefully thought of statements about what payoff the agent gets from each action. In the example, then, any vector of three numbers (giving the payoffs to the three actions in some order) could be a 'state'. Table 3 shows one possible representation along this line.

The subjective states (rain, 1) and (rain, 2) represent the agent's uncertainty about the payoffs to investing when the only objective uncertainty he can think of (rain versus no rain) is resolved in favour of rain.

This idea also appears in various forms in the literature. Fishburn (1970) includes an early statement of the idea and more involved treatments appear in Ghirardato (2001), Kreps (1979, 1992), Dekel et al. (2001), Halpern and Rêgo (2006) (embodied in their 'virtual moves'), and Epstein et al. (2007), among others. Some of these models (for example, Kreps or Dekel, Lipman, and Rustichini) use expected utility over these

Unforeseen Contingencies, Table 2 Perceived payoffs, version 1

Subjective state	Perceived payoff if buy	Perceived payoff if save	Perceived payoff if invest
(rain)	5	0	x
(no rain)	6	8	x′

Unforeseen Contingencies, Table 3 Perceived payoffs, version 2

Subjective state	Perceived payoff if buy	Perceived payoff if save	Perceived payoff if invest
(rain, 1)	5	0	10
(rain, 2)	5	0	−20
(no rain, 1)	6	8	20
(no rain, 2)	6	8	−10

'states', while others (for example, Ghirardato or Epstein, Marinacci, and Seo) use models of agents who are uncertainty averse with respect to what the 'right' payoff is.

At first glance, this approach appears to be capable of generating a model we could use for the purposes of interest. In fact, this model looks very similar to the 'observable but unverifiable' uncertainty story used by Grossman and Hart (1986), Hart and Moore (1988), and Hart (1995) to model incomplete contracts. For brevity, I henceforth refer to this as the *GHM approach.* This approach assumes that some of the variables relevant to the contracting parties are observed by them but cannot be 'shown' to a court. As a result, the parties cannot, according to this approach, contract on these variables because a dispute about their realizations cannot be settled by the court. (These papers often also use the assumption that certain actions are indescribable, but this aspect of the GHM approach is not relevant here.) Hence contracts can only allocate control rights – that is, assign the rights to make various decisions *ex post*. Intuitively, if some variables cannot be contracted on, one has to rely on the parties to choose appropriately in the relevant contingencies.

Similarly, it seems natural to assume that these subjective states cannot be contracted over. If a 'state' is simply a specification of a utility function for each agent as a function of the actions, how can such a state be verified? Thus the GHM approach appears to fit naturally with this approach to modelling unforeseen contingencies.

To be more concrete, consider again the trade-off between short-term and long-term contracts. It seems natural to assume that the set of subjective states for a period far in the future is larger than the set for a period closer to the present. In this sense, any contract regarding a distant period cannot be as 'fine tuned' as a contract for a close period. This will naturally give a trade-off between the value of contracting far in advance and the value of contracting once the parties know more, and so can contract better.

To be still more explicit, suppose the table above gives the subjective state space. Suppose that if the agents write a long-term contract today, it can only specify an outcome as a function of whether it rains or not. On the other hand, assume that they expect that if they wait and write a contract at a later date, they will learn whether the 'correct' state space is {(rain, 1), (no rain, 1)} or {(rain, 2), (no rain, 2)}. Thus a contract written later will not share risk as well, but can specify the outcomes given rain and no rain more efficiently.

Unfortunately, the work of Maskin and Tirole (1999) calls this conclusion into doubt. They show that observable but unverifiable variables (and indescribable actions) do not justify a departure from standard contract theory. More specifically, even with observable but unverifiable variables, a mechanism can be designed that will induce the parties to reveal to the court what the values of the unverifiable variables are. Thus the fact that they cannot *prove* these facts to the court is not a problem since the court knows they will tell the truth.

Given the similarity to GHM, this suggests that the above model of unforeseen contingencies may not yield results different from standard contract theory either. In terms of the example above, the agents might be able to set up a mechanism that effectively enables them to write a contract specifying different outcomes in the subjective states (rain, 1) and (rain, 2). If so, there cannot be any gain in waiting.

In principle, there are ways one could introduce more realism to the Maskin–Tirole framework and overturn their conclusion. For example, they suggest that bounded rationality may imply that the agents do not understand and so cannot use the complex mechanisms needed to enforce truth telling. On the other hand, it seems surprising that considerations other than unforeseen contingencies would be needed to generate something different from standard contract theory. Maskin and Tirole do not allow the kind of aversion to payoff uncertainty present in Epstein, Marinacci, and Seo, so this is another direction that may be fruitful.

In short, unforeseen contingencies appear to be important to understanding economic and political institutions but, as yet, economic theory lacks a formal model of the phenomenon that can be used to study these issues.

See Also

- ▶ Incomplete Contracts
- ▶ Long Run and Short Run

Acknowledgments The author would like to thank Eddie Dekel, Jing Li, Aldo Rustichini, and Marie- Odile Yanelle for discussions and comments.

Bibliography

Al Najjar, N., L. Anderlini, and L. Felli. 2006. Undescribable events. *Review of Economic Studies* 73: 849–868.

Dekel, E., B. Lipman, and A. Rustichini. 2001. Representing preferences with a unique subjective state space. *Econometrica* 69: 891–934.

Epstein, L., M. Marinacci, and K. Seo. 2007. Coarse contingencies. Working paper. University of Rochester.

Fishburn, P. 1970. *Utility theory for decision making*, Publications in Operations Research, No. 18. New York: Wiley.

Ghirardato, P. 2001. Coping with ignorance: Unforeseen contingencies and non-additive uncertainty. *Economic Theory* 17: 247–276.

Grossman, S., and O. Hart. 1986. The costs and benefits of ownership: A theory of vertical and lateral integration. *Journal of Political Economy* 94: 691–719.

Halpern, J., and L. Rêgo. 2005. Interactive awareness revisited. In *Proceedings of tenth conference on theoretical aspects of rationality and knowledge*, ed. R. van der Meyden. Singapore: National University of Singapore.

Halpern, J., and L. Rêgo. 2006. Extensive games with possibly unaware players. In *Proceedings of the fifth international joint conference on autonomous agents and multiagent systems*. New York: ACM Press.

Hart, O. 1995. *Firms, contracts, and financial structure*. Oxford: Clarendon Press.

Hart, O., and J. Moore. 1988. Incomplete contracts and renegotiation. *Econometrica* 56: 755–786.

Hayek, F. 1960. *The constitution of liberty*. Chicago: University of Chicago Press.

Heifetz, A., M. Meier, and B. Schipper. 2006a. Interactive unawareness. *Journal of Economic Theory* 130: 78–94.

Heifetz, A., M. Meier, and B. Schipper. 2006b. Unawareness, beliefs, and games. Working paper. University of California–Davis.

Internal Revenue Service. 2006. *Selling your home*, Publication 523. Washington, DC: Internal Revenue Service, US Treasury Department.

Kreps, D. 1979. A representation theorem for 'preference for flexibility'. *Econometrica* 47: 565–576.

Kreps, D. 1992. Static choice and unforeseen contingencies. In *Economic analysis of markets and games: Essays in honor of Frank Hahn*, ed. P. Dasgupta, D. Gale, O. Hart, and E. Maskin. Cambridge: MIT Press.

Li, J. 2006a. Information structures with unawareness. Working paper, University of Pennsylvania.

Li, J. 2006b. Dynamic games with unawareness. Working paper, University of Pennsylvania.

Maskin, E., and J. Tirole. 1999. Unforeseen contingencies and incomplete contracts. *Review of Economic Studies* 66: 83–114.

Modica, S., A. Rustichini, and J.-M. Tallon. 1998. Unawareness and bankruptcy: Aa general equilibrium model. *Economic Theory* 12: 259–292.

Williamson, O. 1975. *Markets and hierarchies: Analysis and antitrust implications*. New York: Free Press.

Uniqueness of Equilibrium

Michael Allingham

In general equilibrium theory, equilibrium prices may be interpreted as those prices which coordinate the buying and selling plans of all the various agents in the economy; equivalently, they may be interpreted as the values of the commodities. Such values will only be well defined if there is only one system of coordinating prices, that is, if the equilibrium is unique. If this does not obtain then at least the set of equilibrium price systems should not be too large, that is, there should be only a finite number of equilibria.

The question of uniqueness was first posed by Walras (1874–7), but received its first systematic treatment by Wald (1936). In the present discussion we commence with a formal definition of uniqueness. We then note that there may be multiple, and even infinitely many, equilibria, but show that the latter possibility is unlikely. In the light of this we examine various conditions which are sufficient to ensure that equilibrium is unique. Finally, we note some problems which may arise in the presence of multiple equilibria.

We may represent an economy with n commodities by the excess demand function f: $S \to R^n$, where $S = R^n_+ - 0$. The interpretation of this is that $f(p)$ is the vector of aggregate excess demands (positive) or excess supplies (negative)

expressed at the price system p. Under some reasonable assumptions on the underlying parameters of the economy, that is the individual preferences and endowments, this excess demand function has the following properties:

Homogeneity: $f(tp) = f(p)$ for all positive t.
Walras' Law: $p \cdot f(p) = 0$ for all p.
Desirability: $fi(p)$ is infinite if $pi = 0$.
Differentiability: f is continuously differentiable.

The price system p is an equilibrium price system if $f(p) = 0$. Because of desirability it is clear that if p is an equilibrium then p is strictly positive. We shall denote by E the set of equilibrium prices. Now if p is in E then so is t_p for any positive t (because of Homogeneity), so we take the equilibrium p to be unique if q is in E implies that $q = t_p$ for some positive t. Equivalent formulations specify that the equilibrium p is unique if it is the only equilibrium in the unit simplex in R^n, or if it is the only equilibrium with, say, $p_n = 1$. Of course, the question of uniqueness of equilibrium only arises if there is at least one equilibrium: however, under the above four conditions on the excess demand function this existence is assured (Debreu 1959).

The first point to note is that equilibrium may well not be unique: indeed, there may be infinitely many equilibria. This follows from the fact that the above four conditions are, at the most, the only restrictions which economic theory places on the excess demand function (Debreu 1970). It is therefore straightforward to construct examples of economies with many equilibria, and even of economies in which all positive prices are equilibrium prices.

In the light of this point we first consider the likelihood of encountering an infinite number of equilibria. Let $F(p)$ be the Jacobian of excess supply, that is of $-f$, at p with the last row and last column deleted. We lose no information in working with F rather than with the full Jacobian: simply because we can set $p_n = 1$ without loss of generality (Homogeneity) and because if $f_i(p) = 0$ for all i other than n then $f_n(p) = 0$ (Walras' Law). The economy is said to be regular if $F(p)$ is of full rank at all p in E. The importance of this is that almost all economies are regular, in that the set of economies which are not regular, or critical economies, is a closed null subset of the set of all economies, as may be shown using Sard's theorem (Debreu 1970).

With this in mind we may now observe that the number of equilibria in a regular economy is finite. This may be shown, using the Poincaré–Hopf index theorem, by defining the index $i(p) = 1$ if the determinant $\det F(p) > 0$ and $i(p) = -1$ if $\det F(p) < 0$ and noting that the sum of $i(p)$ over all p in E is 1 (Dierker 1972). This result has two immediate corollaries: the first is that the number of equilibria in a regular economy is odd; the second is that if $\det F(p)$ is positive for all p in E then equilibrium is unique. Taking the above two results together we note that in almost all economies the number of equilibria is finite.

The economic interpretation of $\det F(p)$ being positive in the two-dimensional case is that excess demand is 'downward-sloping' (or excess supply 'upward-sloping'). It is intuitively clear that this ensures uniqueness. In the general case, however, the economic interpretation of this property is not so clear; we therefore examine some more interpretable properties which ensure uniqueness.

An economy with excess demand function f has the revealed preference property if $p \cdot f(q) > 0$ wherever p is in E and q is not in E. It is well known that if g is an individual's excess demand function then $q \cdot g(p) \leq 0$ (that is $g(p)$ is available to the individual at price q) implies that $p \cdot f(q) > 0$ (that is $g(q)$ is not available at price p). If all individuals are identical this property will hold in aggregate, where, if p is in E, $q \cdot f(p) = 0$ immediately, so that $p \cdot f(q) > 0$ if q is not in E. Thus if all agents are identical the economy has the revealed preference property. In fact the essential reason why the property holds if all agents are identical is that there is then no trade at equilibrium. It can readily be seen that the property holds if there is no trade at equilibrium for whatever reason. This becomes relevant if we consider today's endowments as being the result of yesterday's trading, with no intervening consumption or production.

Now assume that f has the revealed preference property and let p and q be in E but suppose that r, a proper linear combination of p and q, is not in E.

Then if $p \cdot f(r) > 0$ and if $q \cdot f(r) > 0$ so that $r \cdot f(r) > 0$ which contradicts Walras' Law. This shows that E is convex. Since in almost all economies E is finite, and the only finite convex set is a singleton, it follows that in almost all economies with the revealed preference property equilibrium is unique.

An economy with excess demand function f has the gross substitute property if $p_i > q_i$ and $p_j = q_j$ for each $j \neq i$ imply that $f_j(p) > f_j(q)$ for each $j \neq i$. If this property obtains then Walras' Law implies that excess demand must be 'downward-sloping'. The interpretation of this is that all commodities are substitutes for each other (in the gross sense, that is including income effects as well as substitution effects). In fact, the gross substitutes property implies the revealed preference property; instead of showing this implication we will demonstrate directly that the gross substitutes property ensures uniqueness.

Let p be in E and for any $q \neq p$ define $m = \max_i q_i/p_i = q_k/p_k$ say, and let $r = mp$. Then $r_i \geqq q_i$ for each i with equality for $i = k$ and inequality for some $i \neq k$, so by repeated use of the gross substitutes property we have $f_k(r) > f_k(q)$. But by Homogeneity $f(r) = f(p) = 0$, so that $f_k(q) < 0$ and q is not in E. Thus equilibrium is unique.

If $p_i > q_i$ and $p_j = q_j$ for each $j \neq i$ imply that $f_j(p) \geqq f_j(q)$ for each $j \neq i$ then the economy has the weak gross substitutes property. Arguments analogous to that above show that in this case E is convex, so that in almost all economies equilibrium is unique. Alternatively, if the economy is, in addition, connected in some specific sense, then equilibrium is definitely unique.

Finally, we should note that these properties of revealed preference and gross substitutes do not depend on differentiability. If we accept differentiability there are other properties which ensure uniqueness. One such is diagonal dominance, which is that F has a positive diagonal and that there are some units in which commodities can be measured such that each of their excess demands are more sensitive to a change in their own price than they are to a change in all other non-numeraire prices combined.

It is clear from the above discussion that uniqueness is a strong property. If it does not obtain equilibrium prices will still coordinate individual agents' plans, but they will not, of course, define values uniquely.

One more specific problem which arises under multiple equilibria concerns stability. Assume that we have a process for changing prices such that no change is made in equilibrium and define an equilibrium price p to be stable under this process if prices converge to p whatever their initial values. Then if there are two equilibria, say p and q, neither can be stable: the path starting at p will remain at p so that q is not stable, and conversely. This problem may be avoided by considering only system stability, that is by defining the set E of equilibrium prices to be stable if all paths converge to E. It may also be avoided by considering only local stability, that is by defining p to be stable if prices converge to p given initial values sufficiently close to p. It is clear that even local stability requires equilibria to be separated, and thus finite. As we have seen, this applies in a regular economy; in this case the index theorem then implies that if there are $2k + 1$ equilibria (we know the number to be odd) then $k + 1$ will typically be locally stable and k unstable.

A further specific problem which arises under multiple equilibria concerns comparative statics. Assume that we want to compare the set of equilibria E of the economy with the set of equilibria E' of some new economy obtained from the original economy by some specified parameter change. If there are multiple equilibria we may be able to say very little: for example if p is in E and both p' and q' are in E' and $p' < p < q'$ all comparative statics results are ambiguous. However, in regular economies, where not only are equilibria separated but also the elements of E and of E' correspond to one another in a natural one-to-one way, this problem may be avoided by considering only local comparative statics, interpreted analogously to local stability.

See Also

▶ General Equilibrium
▶ Regular Economies

Bibliography

Debreu, G. 1959. *Theory of value*. New York: Wiley.
Debreu, G. 1970. Economies with a finite set of equilibria. *Econometrica* 38(3): 387–392.
Dierker, E. 1972. Two remarks on the number of equilibria of an economy. *Econometrica* 40(5): 867–881.
Wald, A. 1936. Über einige Gleichungssysteme der mathematischen Ökonomie. *Zeitschrift für Nationalökonomie* 7: 637–670; trans. as: 1951. On some systems of equations of mathematical economics, *Econometrica* 19: 368–403.
Walras, L. 1874–7. *Eléments d'économie politique pure*. Definitive edn, Lausanne: Corbaz, 1926. Translated by W. Jaffé as *Elements of pure economics*. London: George Allen & Unwin, 1954.

Unit Roots

Peter C. B. Phillips

Abstract

Models with autoregressive unit roots play a major role in modern time series analysis and are especially important in macroeconomics, where questions of shock persistence arise, and in finance, where martingale concepts figure prominently in the study of efficient markets. The literature on unit roots is vast and applications of unit root testing span the social, environmental and natural sciences. The present article overviews the theory and concepts that underpin this large field of research and traces the originating ideas and econometric methods that have become central to empirical practice.

Keywords

ARCH models; ARMA time-series processes; Bayesian inference; Bayesian statistics; Bayesian time series analysis; Bias reduction; Break point analysis; Classical statistics; Cointegration; Confidence intervals; Efficient markets hypothesis; Forecasting; Functional central limit theorem; GARCH models; Integrated conditional heteroskedasticity models; Lagrange multipliers; Long-run variance; Martingales; Model selection; Nonstationarity; Polynomials; Present value; Probability; Rational expectations business cycle models; Real business cycles; Spurious regressions; Statistical inference; Stochastic trends; Term structure of interest rates; Unit root distributions; Unit roots; Wiener process

JEL Classifications

C22

Economic and financial time series have frequently been successfully modelled by autoregressive moving-average (ARMA) schemes of the type

$$a(L)y_t = b(L)\varepsilon_t, \tag{1}$$

where ε_t is an orthogonal sequence (that is, $E(\varepsilon_t) = 0, E(\varepsilon_t\varepsilon_s) = 0$ for all $t \neq s$), L is the backshift operator for which $Ly_t = y_{t-1}$ and $a(L)$, $b(L)$ are finite-order lag polynomials

$$a(L) = \sum_{i=0}^{p} a_i L^i, \quad b(L) = \sum_{j=0}^{q} b_j L^j,$$

whose leading coefficients are $a_0 = b_0 = 1$. Parsimonious schemes (often with $p + q \leq 3$) are usually selected in practice either by informal 'model identification' processes such as those described in the text by Box and Jenkins (1976) or more formal order-selection criteria which penalize choices of large p and/or q. Model (1) is assumed to be irreducible, so that $a(L)$ and $b(L)$ have no common factors. The model (1) and the time series y_t are said to have an *autoregressive unit root* if $a(L)$ factors as $(1 - L)a_1(L)$ and a *moving-average unit root* if $b(L)$ factors as $(1 - L)b_1(L)$.

Since the early 1980s, much attention has been focused on models with autoregressive unit roots. In part, this interest is motivated by theoretical considerations such as the importance of martingale models of efficient markets in finance and the dynamic consumption behaviour of representative economic agents in macroeconomics; and, in part,

the attention is driven by empirical applications, which have confirmed the importance of random walk phenomena in practical work in economics, in finance, in marketing and business, in social sciences like political studies and communications, and in certain natural sciences. In mathematics and theoretical probability and statistics, unit roots have also attracted attention because they offer new and important applications of functional limit laws and weak convergence to stochastic integrals. The unit root field has therefore drawn in participants from an excitingly wide range of disciplines.

If (1) has an autoregressive unit root, then we may write the model in difference form as

$$\Delta y_t = u_t = a_1(L)^{-1} b(L)\varepsilon_t, \tag{2}$$

where the polynomial $a_1(L)$ has all its zeros outside the unit circle. This formulation suggests more general nonparametric models where, for instance, u_t may be formulated in linear process (or Wold representation) form as

$$_{ut} = c(L)\varepsilon_t = \sum_{j=0}^{\infty} c_j \varepsilon_{t-j}, \quad \text{with} \quad \sum_{j=0}^{\infty} c_j^2 < \infty, \tag{3}$$

or as a general stationary process with spectrum $f_u(\lambda)$. If we solve (2) with an initial state y_0 at $t = 0$, we have the important partial sum representation

$$y_t = \sum_{j=1}^{t} u_j + y_0 = S_t + y_0, \tag{4}$$

showing that S_t and hence y_t are 'accumulated' or 'integrated' processes proceeding from a certain initialization y_0. A time series y_t that satisfies (2) or (4) is therefore said to be integrated of order one (or a unit root process or an I(1) process) provided $f_u(0) > 0$. The latter condition rules out the possibility of a moving-average unit root in the model for u_t that would cancel the effect of the autoregressive unit root (for example, if $b(L) = (1 - L)b_1(L)$ then model (2) is $\Delta y_t = \Delta a_1(L)^{-1} b_1(L)\varepsilon_t$ or, after cancellation, just $y_t = a_1(L)^{-1} b_1(L)\varepsilon_t$, which is not I(1)). Note that this possibility is also explicitly ruled out in the ARMA case by the requirement that $a(L)$ and $b(L)$ have no common factors. Alternatively, we may require that $u_t \neq \Delta v_t$ for some weakly stationary time series v_t, as in Leeb and Potscher (2001) who provide a systematic discussion of I(1) behaviour. The partial sum process S_t in (4) is often described as a *stochastic trend*.

The representation (4) is especially important because it shows that the effect of the random shocks u_j on y_t does not die out as the time distance between j and t grows large. The shocks u_j then have a *persistent* effect on y_t in this model, in contrast to stationary systems. Whether actual economic time series have this characteristic or not is, of course, an empirical issue. The question can be addressed through statistical tests for the presence of a unit root in the series, a subject which has grown to be of major importance since the mid-1980s and which will be discussed later in this article. From the perspective of economic modelling the issue of persistence is also important because, if macroeconomic variables like real GNP have a unit root, then shocks to real GNP have permanent effects, whereas in traditional business cycle theory the effect of shocks on real GNP is usually considered to be only temporary. In more recent real business cycle theory, variables like real GNP are modelled in such a way that over the long run their paths are determined by supply side shocks that can be ascribed to technological and demographic forces from outside the model. Such economic models are more compatible with the statistical model (4) or close approximations to it in which the roots are local to unity in a sense that is described later in this essay.

Permanent and transitory effects in (4) can be distinguished by decomposing the process u_t in (3) as follows

$$\begin{aligned} u_t &= \{C(1) + (L-1)\tilde{C}(L)\}\varepsilon_t \\ &= C(1)\varepsilon_t + \tilde{\varepsilon}_{t-1} - \tilde{\varepsilon}_t, \end{aligned} \tag{5}$$

where $\tilde{\varepsilon}_t = \tilde{C}(L)\varepsilon_t, \tilde{C}(L) = \sum_0^{\infty} \tilde{c}_j L^j|$ and $\tilde{c}_j = \sum_{j+1}^{\infty} c_s$ The decomposition (5) is valid algebraically if

$$\sum_{j=0}^{\infty} j^{1/2}|c_j| < \infty, \tag{6}$$

as shown in Phillips and Solo (1992), where validity conditions are systematically explored. Equation (5) is sometimes called the Beveridge–Nelson (1981) or BN decomposition of u_t, although both specialized and more general versions of it were known and used beforehand. The properties of the decomposition were formally investigated and used for the development of laws of large numbers and central limit theory and invariance principles in the paper by Phillips and Solo (1992). When the decomposition is applied to (4) it yields the representation

$$\begin{aligned} y_t &= C(1)\sum_1^t \varepsilon_j + \widetilde{\varepsilon}_0 - \widetilde{\varepsilon}_t + y_0 \\ &= C(1)\sum_1^t \varepsilon_j + \xi_t + y_0, \quad \text{say,} \end{aligned} \tag{7}$$

where $\xi_t = \widetilde{\varepsilon}_0 - \widetilde{\varepsilon}_t$. The right side of (7) decomposes y_t into three components: the first is a martingale component, $Y_t = C(1)\sum_1^t \varepsilon_j$, where the effects of the shocks ε_j are permanent; the second is a stationary component, where the effects of shocks are transitory, viz. $\xi_t = \widetilde{\varepsilon}_0 - \widetilde{\varepsilon}_t$, since the process $\widetilde{\varepsilon}_t$ is stationary with valid Wold representation $\widetilde{\varepsilon}_t = \tilde{C}(L)\varepsilon_t$ under (6) when ε_t is stationary with variance σ^2; and the third being the initial condition y_0. The relative strength of the martingale component is measured by the magnitude of the (infinite dimensional) coefficient $C(1) = \sum_{j=0}^{\infty} c_j$, which plays a large role in the measurement of long-run effects in applications. Accordingly, the decomposition (7) is sometimes called the martingale decomposition (cf., Hall and Heyde 1980) where it was used in various forms in the probability literature prior to its use in economics.

The leading martingale term $Y_t = C(1)\sum_{s=1}^t \varepsilon_s$ in (7) is a partial sum process or stochastic trend and, under weak conditions on ε_t (see Phillips and Solo 1992, for details) this term satisfies a functional central limit theorem whereby the scaled process

$$n^{-1/2}Y_{[nr]} \Rightarrow B(r), \tag{8}$$

a Brownian motion with variance $\omega^2 = C(1)^2\sigma^2 = 2\pi f_u(0)$, a parameter which is called the long-run variance of u_t, and where $[\,\cdot\,]$ signifies the integer part of its argument. Correspondingly,

$$n^{-1/2}Y_{[nr]} \Rightarrow B(r), \tag{9}$$

provided $y_0 = o_p(\sqrt{n})$ A related result of great significance is based on the limit

$$n^{-1}\sum_{t=1}^{[nr]} Y_{t-1}\varepsilon_t C(1) \Rightarrow \int_0^r B\,dB \tag{10}$$

of the sample covariance of Y_{t-1} and its forward increment, $C(1)\varepsilon_t$. $C(1)\varepsilon_t$. The limit process $M(r) = \int_0^r B\,dB$ is represented here as an Ito (stochastic) integral and is a continuous time martingale. The result may be proved directly (Solo 1984; Phillips 1987a; Chan and Wei 1988) or by means of martingale convergence methods (Ibragimov and Phillips, 2004) which take advantage of the fact that $\sum_{t=1}^k Y_{t-1}\varepsilon_t$ is a martingale. The limit theory given (9) and (10) was extended in Phillips (1987b, 1988a) and Chan and Wei (1987) to cases where the model (2) has an autoregressive root in the vicinity of unity ($\rho = 1 + \frac{c}{n}$, for some fixed c) rather than precisely at unity, in which case the limiting process is a linear diffusion (or Ornstein–Uhlenbeck process) with parameter c. This limit theory has proved particularly useful in the analysis of asymptotic local power functions of unit root tests (Phillips 1987b) and the construction of confidence intervals (Stock 1991). Phillips and Magdalinos (2007) considered moderate deviations from unity of the form $\rho = 1 + \frac{c}{k}$, where $k \to \infty$ but $\frac{k}{n} \to 0$, so that the roots are local but further away from unity, showing that central limit laws rather than functional laws apply in this case (see also Giraitis and Phillips 2006). This theory is applicable to mildly explosive processes (where $c > 0$) and therefore assists in bridging the gap between the limit theory for the stationary, unit root and explosive cases.

Both (8) and (10) have important multivariate generalizations that play a critical role in the study

of spurious regressions (Phillips 1986) and cointegration limit theory (Phillips and Durlauf 1986; Engle and Granger 1987; Johansen 1988; Phillips 1988a; Park and Phillips 1988, 1989). In particular, if $y_t = (y'_{at}, y'_{bt})'$, $u_t = (u'_{at}, u'_{bt})$ and $\varepsilon_t = (\varepsilon'_{at}, \varepsilon'_{bt})'$ are vector processes and $E(\varepsilon_t \varepsilon'_t) = \Sigma$, then: (i) the decomposition (5) continues to hold under (6), where $|c_j|$ is interpreted as a matrix norm; (ii) the functional law (8) holds and the limit process is vector Brownian motion $B = (B'_a, B'_b)'$ with covariance matrix $\Omega = C(1)\Sigma C(1)'$; and (iii) sample covariances converge weakly to stochastic processes with drift, as in

$$n^{-1}\sum_{t=1}^{[nr]} Y_{at-1}u'_{bt} \Rightarrow \int_0^r B_a dB'_b + \lambda_{ab} r, \qquad (11)$$

where $\lambda_{ab} = \sum_{k=1}^{\infty} E(u_{a0}u'_{bk})$ is a one sided long-run covariance matrix. The limit process on the right side of (11) is a semimartingale (incorporating a deterministic drift function $\lambda_{ab}r$) rather than a martingale when $\lambda_{ab} \neq 0$.

The decomposition (7) plays an additional role in the study of cointegration (Engle and Granger 1987). When the coefficient matrix $C(1)$ is singular and β spans the null space of $C(1)'$, then $\beta^0 C(1) = 0$ and (7) leads directly to the relationship

$$\beta' Y_t = 0, \quad \text{a.s.},$$

which may be interpreted as a long run equilibrium (cointegrating) relationship between the stochastic trends (Y_t) of y_t. Correspondingly, we have the empirical cointegrating relationship

$$\beta' y_t = v_t,$$

among the observed series y_t with a residual $v_t = \beta'(\xi_t + y_0)$ that is stationary. The columns of β span what is called the cointegration space.

The above discussion presumes that the initialization y_0 has no impact on the limit theory, which will be so if y_0 is small relative to the sample size, specifically, if $y_0 = o_p(\sqrt{n})$. However, if $y_0 = O_p(\sqrt{n})$, for example if $y_0 = y_{0\theta_n}$ is indexed to depend on past shocks u_{-j} (satisfying a process of the form (3)) to some point in the distant past θ_n which is measured in terms of the sample size n, then the results can differ substantially. Thus, if $\theta_n = [\kappa n]$, for some fixed parameter $\kappa > 0$, then $y_{0\theta_n} = \sum_1^{[\kappa n]} u_{-j}$, and $n^{-1/2} y_{0\theta_n} \Rightarrow B_0(\kappa)$, for some Brownian motion $B_0(\kappa)$ with covariance matrix Ω_{00} given by the long-run variance matrix of u_{-j}. Under such an initialization, (9) and (11) are replaced by

$$\begin{aligned} n^{-1/2} y_{[nr]} &\Rightarrow B(r) + B_0(\kappa) : \\ &= B(r, \kappa), \quad \text{say} \end{aligned} \qquad (12)$$

and

$$n^{-1}\sum_{t=1}^{[nr]} y_{at-1}u'_{bt} \Rightarrow \int_0^r B_a(s,\kappa) dB_b(s)' + \lambda_{ab} r,$$

so that initializations play a role in the limit theory. This role becomes dominant when κ becomes very large, as is apparent from (12). The effect of initial conditions on unit root limit theory was examined in simulations by Evans and Savin (1981, 1984), by continuous record asymptotics by Phillips (1987a), in the context of power analysis by Müller and Elliott (2003), for models with moderate deviations from unity by Andrews and Guggenberger (2006), and for cases of large κ by Phillips (2006).

Model (4) is of special interest to economists working in finance because its output, y_t, behaves as if it has no fixed mean and this is a characteristic of many financial time series. If the components u_j are independent and identically distributed (i.i.d.) then y_t is a random walk. More generally, if u_j is a martingale difference sequence (mds) (that is orthogonal to its own past history so that $E_{j-1}(u_j) = E(u_j | u_{j-1}, u_{j-2}, \ldots, u_1) = 0$ then y_t is a martingale. Martingales are the essential mathematical elements in the development of a theory of fair games and they now play a key role in the mathematical theory of finance, exchange rate determination and securities markets. Duffie (1988) provides a modern treatment of finance that makes extensive use of this theory.

In empirical finance much attention has recently been given to models where the

conditional variance $E\left(u_j^2 | u_{j-1}, u_{j-2}, \ldots, u_1\right) = \sigma_j^2$ is permitted to be time varying. Such models have been found to fit financial data well and many different parametric schemes for σ_j^2 have been devised, of which the ARCH (autoregressive conditional heteroskedasticity) and GARCH (generalized ARCH) models are the most common in practical work. These models come within the general class of models like (1) with mds errors. Some models of this kind also allow for the possibility of a unit root in the determining mechanism of the conditional variance σ_j^2 and these are called integrated conditional heteroskedasticity models. The IGARCH (integrated GARCH) model of Engle and Bollerslev (1986) is an example, where for certain parameters $\omega \geq 0$, $\beta \geq 0$, and $\alpha > 0$, we have the specification

$$\sigma_j^2 = \omega + \beta\sigma_{j-1}^2 + \alpha u_{j-1}^2, \tag{13}$$

with $\alpha + \beta = 1$ and $u_j = \sigma_j z_j$, where the z_j are i.i.d. innovations with $E(z_j) = 0$ and $E\left(z_j^2\right) = 1$. Under these conditions, the specification (13) has the alternative form

$$\sigma_j^2 = \omega + \sigma_{j-1}^2 + \alpha\sigma_{j-1}^2\left(z_{j-1}^2 - 1\right), \tag{14}$$

from which it is apparent that σ_j^2 has an autoregressive unit root. Indeed, since

$$E\left(\sigma_j^2 | \sigma_{j-1}^2\right) = \omega + \sigma_{j-1}^2,$$

σ_j^2 is a martingale when $\omega = 0$. It is also apparentfrom (14) that shocks as manifested in the deviation $z_{j-1}^2 - 1$ are persistent in σ_j^2. Thus, σ_j^2 shares some of the characteristics of an I(1) integrated process. But in other ways, σ_j^2 is very different. For instance, when $\omega = 0$ then $\sigma_j^2 \to 0$ almost surely as $j \to \infty$ and, when $\omega > 0$, σ_j^2 is asymptotically equivalent to a strictly stationary and ergodic process. These and other features of models like (13) for conditional variance processes with a unit root are studied in Nelson (1990).

In macroeconomic theory also, models such as (2) play a central role in modern treatments. In a highly influential paper, R. Hall (1978) showed that under some general conditions consumption is well modelled as a martingale, so that consumption in the current period is the best predictor of future consumption, thereby providing a macroeconomic version of the efficient markets hypothesis. Much attention has been given to this idea in subsequent empirical work.

One generic class of economic model where unit roots play a special role is the 'present value model' of Campbell and Shiller (1988). This model is based on agents' forecasting behaviour and takes the form of a relationship between one variable Y_t and the discounted, present value of rational expectations of future realizations of another variable $X_{t+i} (i = 0,1,2,\ldots)$. More specifically, for some stationary sequence c_t (possibly a constant) we have

$$Y_t = \theta(1-\delta)\sum_{i=0}^{\infty}\delta^i E_t(X_{t+i}) + c_t. \tag{15}$$

When X_t is a martingale, $E_t(X_{t+i}) = X_t$ and (15) becomes

$$Y_t = \theta X_t + c_t, \tag{16}$$

so that Y_t and X_t are cointegrated in the sense of Engle and Granger (1987). More generally, when X_t is I(1) we have

$$Y_t = \theta X_t + \overline{c}_t, \tag{17}$$

where $\overline{c}_t = c_t + \theta\sum_{k=1}^{\infty}\delta^k E_t(\Delta X_{t+k})$, so that Y_t and X_t are also cointegrated in this general case. Models of this type arise naturally in the study of the term structure of interest rates, stock prices and dividends and linear-quadratic intertemporal optimization problems.

An important feature of these models is that they result in parametric linear cointegrating relations such as (16) and (17). This linearity in the relationship between Y_t and X_t accords with the linear nature of the partial sum process that determines X_t itself, as seen in (4), and has been

extensively studied since the mid-1980s. However, in more general models, economic variables may be determined in terms of certain nonlinear functions of fundamentals. When these fundamentals are unit root processes like (4), then the resulting model has the form of a nonlinear cointegrating relationship. Such models are relevant, for instance, in studying market interventions by monetary and fiscal authorities (Park and Phillips 2000; Hu and Phillips 2004) and some of the asymptotic theory for analysing parametric models of this type and for statistical inference in such models is given in Park and Phillips (1999, 2001), de Jong (2004), Berkes and Horváth (2006) and Pötscher (2004). More complex models of this type are nonparametric and different methods of inference are typically required with very different limit theories and typically slower convergence rates (Karlsen et al. 2007; Wang and Phillips 2006). Testing for the presence of such nonlinearities can therefore be important in empirical practice (Hong and Phillips 2005; Kasparis 2004).

Statistical tests for the presence of a unit root fall into the general categories of classical and Bayesian, corresponding to the mode of inference that is employed. Classical procedures have been intensively studied and now occupy a vast literature. Most empirical work to date has used classical methods but some attention has been given to Bayesian alternatives and direct model selection methods. These approaches will be outlined in what follows.

Although some tests are known to have certain limited (asymptotic) point optimality properties, there is no known procedure which uniformly dominates others, even asymptotically. Ploberger (2004) provides an analysis of the class of asymptotically admissable tests in problems that include the simplest unit root test, showing that the conventional likelihood ratio (LR) test (or Dickey and Fuller 1979; Dickey and Fuller 1981, t test) is not within this class, so that the LR test, while it may have certain point optimal properties, is either inadmissible or must be modified so that it belongs to the class. This fundamental difficulty, together with the nonstandard nature of the limit theory and the more complex nature of the asymptotic likelihood in unit root cases partly explains why there is such a proliferation of test procedures and simulation studies analysing performance characteristics in the literature.

Classical tests for a unit root may be classified into parametric, semiparametric and nonparametric categories. Parametric tests usually rely on augmented regressions of the type

$$\Delta y_t = a y_{t-1} + \sum_{i=1}^{k-1} \phi_i \Delta y_{t-i} + e_t, \qquad (18)$$

where the lagged variables are included to model the stationary error u_t in (2). Under the null hypothesis of a unit root, we have $a = 0$ in (18) whereas when y_t is stationary we have $a < 0$. Thus, a simple test for the presence of a unit root against a stationary alternative in (18) is based on a one-sided t-ratio test of $\mathcal{H}_0 : a = 0$ against $\mathcal{H}_1 : a < 0$. This test is popularly known as the ADF (or augmented Dickey–Fuller) test (Said and Dickey 1984) and follows the work of Dickey and Fuller (1979, 1981) for testing Gaussian random walks. It has been extensively used in empirical econometric work since the Nelson and Plosser (1982) study, where it was applied to 14 historical time series for the USA leading to the conclusion that unit roots could not be rejected for 13 of these series (all but the unemployment rate). In that study, the alternative hypothesis was that the series were stationary about a deterministic trend (that is, trend stationary) and therefore model (18) was further augmented to include a linear trend, viz.

$$\Delta y_t = \mu + \beta t + a y_{t-1} + \sum_{i=1}^{k-1} \phi_i \Delta y_{t-i} + e_t, \quad (19)$$

When y_t is trend stationary we have $a < 0$ and $\beta \neq 0$ in (19), so the null hypothesis of a difference stationary process is $a = 0$ and $\beta = 0$. This null hypothesis allows for the presence of a non-zero drift in the process when the parameter $\mu \neq 0$. In this case a joint test of the null hypothesis $\mathcal{H}_0 : a = 0, \beta = 0$ can be mounted using a regression F-test. ADF tests of $a = 0$ can also be mounted directly using the coefficient estimate from (18) or

(19), rather than its t ratio (Xiao and Phillips 1998).

What distinguishes both these and other unit root tests is that critical values for the tests are not the same as those for conventional regression F- and t-tests, even in large samples. Under the null, the limit theory for these tests is nonstandard and involves functionals of a Wiener process. Typically, the critical values for five or one per cent level tests are much further out than those of the standard normal or chi-squared distributions. Specific forms for the limits of the ADF t-test (ADF_t) and coefficient (ADF_a) test are

$$ADF_t \Rightarrow \frac{\int_0^1 WdW}{\left(\int_0^1 W^2\right)^{1/2}}, \quad ADF_a \Rightarrow \frac{\int_0^1 WdW}{\int_0^1 W^2}, \tag{20}$$

where W is a standard Wiener process or Brownian motion with variance unity. The limit distributions represented by the functionals (20) are known as unit root distributions. The limit theory was first explored for models with Gaussian errors, although not in the Wiener process form and not using functional limit laws, by Dickey (1976), Fuller (1976) and Dickey and Fuller (1979, 1981), who also provided tabulations. For this reason, the distributions are sometimes known as Dickey–Fuller distributions. Later work by Said and Dickey (1984) showed that, if the lag number k in (18) is allowed to increase as the sample size increases with a condition on the divergence rate that $k = 0(n^{1/3})$, then the ADF test is asymptotically valid in models of the form (2) where u_t is not necessarily autoregressive.

Several other parametric procedures have been suggested, including Von Neumann ratio statistics (Sargan and Bhargava 1983; Bhargava 1986; Stock 1994a), instrumental variable methods (Hall 1989; Phillips and Hansen 1990) and variable addition methods (Park 1990). The latter also allow a null hypothesis of trend stationarity to be tested directly, rather than as an alternative to difference stationarity. Another approach that provides a test of a null of trend stationarity is based on the unobserved components representation

$$y_t = \mu + \beta t + r_t + u_t, \quad r_t = r_{t-1} + v_t. \tag{21}$$

which decomposes a time series y_t into a deterministic trend, an integrated process or random walk (r_t) and a stationary residual (u_t). The presence of the integrated process component in y_t can then be tested by testing whether the variance $\left(\sigma_v^2\right)$ of the innovation v_t is zero. The null hypothesis is then $\mathcal{H}_0 : \sigma_v^2 = 0$, which corresponds to a null of trend stationarity. This hypothesis can be tested in a very simple way using the Lagrange multiplier (LM) principle, as shown in Kwiatkowski et al. (1992), leading to a commonly used test known as the KPSS test. If $\hat{e}_t$ denotes the residual from a regression of y_t on a deterministic trend (a simple linear trend in the case of (21) above) and $\hat{\omega}_e^2$ is a HAC (heteroskedastic and autocorrelation consistent) estimate constructed from $\hat{e}_t$, then the KPSS statistic has the simple form

$$LM = \frac{n^{-2}\sum_{t=1}^{n} S_t^2}{\hat{\omega}_e^2},$$

where S_t is the partial sum process of the residuals $\sum_{j=1}^{t} \hat{e}_j$. Under the null hypothesis of stationarity, this LM statistic converges to $\int_0^1 V_X^2$, where V_X is a generalized Brownian bridge process whose construction depends on the form (X) of the deterministic trend function. Power analysis indicates that test power depends importantly on the choice of bandwidth parameter in HAC estimation and some recent contributions to this subject are Sul et al. (2006) and Müller (2005) and Harris et al. (2007). Other general approaches to testing I(0) versus I(1) have been considered in Stock (1994a, 1999).

By combining r_t and u_t in (21) the components model may also be written as

$$y_t = \mu + \beta t + x_t, \quad \Delta x_t = a x_{t-1} + \eta_t. \tag{22}$$

In this format it is easy to construct an LM test of the null hypothesis that y_t has a stochastic trend component by testing whether $a = 0$ in (22). When $a = 0$, (22) reduces to

$$\Delta y_t = \beta + \eta_t, \quad \text{or} \quad y_t = \beta t + \sum_1^t \eta_i + y_0, \tag{23}$$

and so the parameter μ is irrelevant (or surplus) under the null. However, the parameter β retains the same meaning as the deterministic trend term coefficient under both the null and the alternative hypothesis. This approach has formed the basis of several tests for a unit root that have been developed (see Bhargava 1986; Schmidt and Phillips 1992) and the parameter economy of this model gives these tests some advantage in terms of power over procedures like the ADF in the neighbourhood of the null.

This power advantage may be further exploited by considering point optimal alternatives in the construction of the test and in the process of differencing (or detrending) that leads to (23), as pursued by Elliott et al. (1995). In particular, note that (23) involves detrending under the null hypothesis of a unit root, which amounts to first differencing, whereas if the root were local to unity, the appropriate procedure would be to use quasi differencing. However, since the value of the coefficient in the locality of unity is unknown (otherwise, there would be no need for a test), it can only be estimated or guessed. The procedure suggested by Elliott et al. (1995) is to use a value of the localizing coefficient in the quasi-differencing process for which asymptotic power is calculated by simulation to be around 50 per cent, a setting which depends on the precise model for estimation that is being used. This procedure, which is commonly known as generalized least squares (GLS) detrending (although the terminology is a misnomer because quasi-differencing not full GLS is used to accomplish trend elimination) is then asymptotically approximately point optimal in the sense that its power function touches the asymptotic power envelope at that value. Simulations show that this method has some advantage in finite samples, but it is rarely used in empirical work in practice, partly because of the inconvenience of using specialized tables for the critical values of the resulting test and partly because settings for the localizing coefficient are arbitrary and depend on the form of the empirical model.

Some unit root tests based on standard limit distribution theory have been developed. Phillips and Han (2008), for example, give an autoregressive coefficient estimator whose limit distribution is standard normal for all stationary, unit root and local to unity values of the autoregressive coefficient. This estimator may be used to construct tests and valid confidence intervals, but tests suffer power loss because the rate of convergence of the estimator is $\sqrt{n}$ uniformly over these parameter values. So and Shin (1999) and Phillips et al. (2004) showed that certain nonlinear instrumental variable estimators, such as the Cauchy estimator, also lead to t-tests for a unit root which have an asymptotic standard normal distribution. Again, these procedures suffer power loss from reduced convergence rates (in this case, $n^{1/4}$), but have the advantage of uniformity and low bias. Bias is a well known problem in autoregressive estimation and many procedures for addressing the problem have been considered. It seems that bias reduction is particularly advantageous in the case of unit root tests in panel data, where cross-section averaging exacerbates bias effects when the time dimension is small. Some simulation and indirect inference procedures for bias removal have been successfully used both in autoregressions (Andrews 1993; Gouriéroux et al. 2000) and in panel dynamic models (Gouriéroux, Phillips and Yu 2006).

Semiparametric unit root tests are among the most commonly used unit root tests in practical work and are appealing in terms of their generality and ease of use. Tests in this class employ nonparametric methods to model and estimate the contribution from the error process u_t in (2), allowing for both autocorrelation and heterogeneity. These tests and the use of functional limit theory methods in econometrics, leading to the limit formulae (20), were introduced in Phillips (1987a). Direct least squares regression on

$$\Delta y_t = a y_{t-1} + u_t \tag{24}$$

gives an estimate of the coefficient and its t-ratio in this equation. These two statistics are then corrected to deal with serial correlation in u_t by employing an estimate of the variance of u_t and its long-run variance. The latter estimate may be obtained by a variety of kernel-type HAC or other spectral estimates (such as autoregressive spectral estimates) using the residuals $\hat{u}_t$ of the OLS regression on (24). Automated methods of bandwidth selection (or order selection in the case of autoregressive spectral estimates) may be employed in computing these HAC estimates and these methods typically help to reduce size distortion in unit root testing (Lee and Phillips 1994; Stock 1994a; Ng and Perron 1995, 2001). However, care needs to be exercised in the use of automated procedures in the context of stationarity tests such as the KPSS procedure to avoid test inconsistency (see Lee 1996; Sul et al. 2006).

This semiparametric approach leads to two test statistics, one based on the coefficient estimate, called the $Z(a)$ test, the other based on its t-ratio, called the $Z(t)$ test. The limit distributions of these statistics are the same as those given in (20) for the ADF coefficient and t-ratio tests, so the tests are asymptotically equivalent to the corresponding ADF tests. Moreover, the local power functions are also equivalent to those of the Dickey–Fuller and ADF tests, so that there is no loss in asymptotic power from the use of nonparametric methods to address autocorrelation and heterogeneity (Phillips 1987b). Similar semiparametric corrections can be applied to the components models (21) and (22) leading to generally applicable LM tests of stationarity ($\sigma_2 = 0$) and stochastic trends ($a = 0$).

The Z tests were extended in Phillips and Perron (1988) and Ouliaris, Park and Phillips (1989) to models with drift, and by Perron (1989) and Park and Sung (1994) to models with structural breaks in the drift or deterministic component. An important example of the latter is the trend function

$$h_t = \sum_{j=0}^{p} f_j t^j + \sum_{j=0}^{p} f_{m,j} t_m^j, \quad \text{where} \quad t_m^j = \begin{cases} 0 & t \in \{1, \ldots, m\} \\ (t-m)^j & t \in \{m+1, \ldots, n\} \end{cases} \tag{25}$$

which allows for the presence of a break in the polynomial trend at the data point $t = m + 1$. Collecting the individual trend regressors in (25) into the vector x_t, there exists a continuous function $X(r) = (1, r, \ldots, r^p)'$ such that $D_n^{-1} x_{[nr]} \to X(r)$ as $n \to \infty$ uniformly in $r \in [0,1]$, where $D_n = \mathrm{diag}(1, n, \ldots, n^p)$. If $\mu = \lim_{n \to \infty}(m/n) > 0$ is the limit of the fraction of the sample where the structural change occurs, then the limiting trend function $X_\mu(r)$ corresponding to (25) has a similar break at the point μ. All the unit root tests discussed above continue to apply as given for such broken trend functions with appropriate modifications to the limit theory to incorporate the limit function $X_\mu(r)$. Indeed, (25) may be extended further to allow for multiple break points in the sample and in the limit process. The tests may be interpreted as tests for the presence of a unit root in models where broken trends may be present in the data. The alternative hypothesis in this case is that the data are stationary about a broken deterministic trend of degree p.

In order to construct unit root tests that allow for breaking trends like (25) it is necessary to specify the break point m. (Correspondingly, the limit theory depends on $X_\mu(r)$ and therefore on μ.) In effect, the break point is exogenously determined. Perron (1989) considered linear trends with single break points intended to capture the 1929 stock market crash and the 1974 oil price shock in this way. An alternative perspective is that any break points that occur are endogenous to the data and unit root tests should take account of this fact. In this case, alternative unit root tests have been suggested (for example, Banerjee et al. 1992; Zivot and Andrews 1992) that endogenize the break point by choosing the value of m that gives the least favourable view of the unit root hypothesis. Thus, if $ADF(m)$ denotes the ADF statistic given by the t-ratio for α in the ADF

regression (19) with a broken trend function like (25), then the trend break ADF statistic is

$$ADF(\hat{m}) = \min_{\underline{m} \le m \le \overline{m}} ADF(m), \quad \text{where} \quad \underline{m} = \left[n\underline{\mu}\right], \overline{m} = [n\overline{\mu}], \tag{26}$$

for some $0 < \underline{\mu} < \overline{\mu} < 1$. The limit theory for this trend break *ADF* statistic is given by

$$ADF(\hat{m}) \Rightarrow \inf_{\mu \in \left[\underline{\mu}, \overline{\mu}\right]} \left[\int_0^1 W_{X_\mu} dW\right] \left[\int_0^1 W_{X_\mu}^2\right]^{-1/2}, \tag{27}$$

where W_X is detrended standard Brownian motion defined by $W_X(r) = W(r) - \left[\int_0^1 WX\right]\left[\int_0^1 XX\right]^{-1} X(r)$. The limit process $X_\mu(r)$ that appears in the functional W_{X_μ} is dependent on the trend break point μ over which the functional is minimized. Similar extensions to trend breaks are possible for other unit root tests and to multiple breaks (Bai 1997; Bai and Perron 1998, 2006; Kapetanios 2005). Critical values of the limiting test statistic (27) are naturally further out in the tail than those of the exogenous trend break statistic, so it is harder to reject the null hypothesis of a unit root when the break point is considered to be endogenous.

Asymptotic and finite sample critical values for the endogenized trend break *ADF* unit root test are given in Zivot and Andrews (1992). Simulations studies indicate that the introduction of trend break functions leads to further reductions in the power of unit root tests and to substantial finite sample size distortion in the tests. Sample trajectories of a random walk are often similar to those of a process that is stationary about a broken trend for some particular breakpoint (and even more so when several break points are permitted in the trend). So continuing reductions in the power of unit root tests against competing models of this type is to be expected and discriminatory power between such different time series models is typically low. In fact, the limit Brownian motion process in (9) can itself be represented as an infinite linear random combination of deterministic functions of time, as discussed in Phillips (1998), so there are good theoretical reasons for anticipating this outcome. Carefully chosen trend stationary models can always be expected to provide reasonable representations of given random walk or unit root data, but such models are certain to fail in post-sample projections as the post-sample data drifts away from any given trend or broken trend line. Phillips (1998, 2001) explores the impact of these considerations in a systematic way.

From a practical standpoint, models with structural breaks attach unit weight and hence persistence to the effects of innovations at particular times in the sample period. In effect, break models simply dummy out the effects of certain observations by parameterizing them as persistent effects. To the extent that persistent shocks of this type occur intermittently throughout the entire history of a process, these models are therefore similar to models with a stochastic trend. However, if only one or a small number of such breaks occur then the process does have different characteristics from that of a stochastic trend. In such cases, it is often of interest to identify the break points endogenously and relate such points to institutional events or particular external shocks that are know to have occurred.

More general nonparametric tests for a unit root are also possible. These rely on frequency domain regressions on (24) over all frequency bands (Choi and Phillips 1993). They may be regarded as fully nonparametric because they test in a general way for coherency between the series y_t and its first difference Δy_t. Other frequency domain procedures involve the estimation of a fractional differencing parameter and the use of tests and confidence intervals based on the estimate. The time series y_t is fractionally integrated with memory parameter d if $(1 - L)^d y_t = u_t$ and u_t is a stationary process with spectrum $f_u(\lambda)$ that is continuous at the origin with $f_u(0) > 0$, or a (possibly mildly heterogeneous) process of the form given in (3). Under some rather weak regularity conditions, it is possible to estimate d consistently by semiparametric methods irrespective of the value of d. Shimotsu and Phillips (2005) suggest an exact local Whittle

estimator $\hat{d}$ that is consistent for all d and for which $\sqrt{n}(\hat{d}-d) \Rightarrow N\left(0, \frac{1}{4}\right)$, extending earlier work by Robinson (1995) on local Whittle estimation in the stationary case where $|d| < 1$. These methods are narrow band procedures focusing on frequencies close to the origin, so that long run behaviour is captured. The Shimotsu–Phillips estimator may be used to test the unit root hypothesis $\mathcal{H}_0 : d = 1$ against alternatives such as $\mathcal{H}_1 : d < 1$. The limit theory may also be used to construct valid confidence intervals for d.

The $Z(a)$, $Z(t)$ and ADF tests are the most commonly used unit root tests in empirical research. Extensive simulations have been conducted to evaluate the performance of the tests. It is known that the $Z(a)$, $Z(t)$ and ADF tests all perform satisfactorily except when the error process u_t displays strong negative serial correlation. The $Z(a)$ test generally has greater power than the other two tests but also suffers from more serious size distortion. All of these tests can be used to test for the presence of cointegration by using the residuals from a cointegrating regression. Modification of the critical values used in these tests is then required, for which case the limit theory and tables were provided in Phillips and Ouliaris (1990) and updated in MacKinnon (1994).

While the Z tests and other semiparametric procedures are designed to cope with mildly heterogeneous processes, some further modifications are required when there is systematic time-varying heterogeneity in the error variances. One form of systematic variation that allows for jumps in the variance has the form $E\left(\varepsilon_t^2\right) = \sigma_t^2 = \sigma^2 g\left(\frac{t}{n}\right)$, where the variance evolution function $g\left(\frac{t}{n}\right)$ may be smooth except for simple jump discontinuities at a finite number of points. Such formulations introduce systematic time variation into the errors, so that we may write $\varepsilon_t = g\left(\frac{t}{n}\right)\eta_t$, where ζ_t is a martingale difference sequence with variance $E\zeta_t^2 = \sigma^2$. These evolutionary changes then have persistent effects on partial sums of ε_t, thereby leading to alternate functional laws of the form

$$n^{-1/2} Y_{[nr]} \Rightarrow B_g(r) = \int_0^r g(s) dB(s),$$

in place of (8). Accordingly, the limit theory for unit root tests changes and some nonparametric modification of the usual tests is needed to ensure that existing asymptotic theory applies (Beare 2006) or to make appropriate corrections in the limit theory (Cavaliere 2004; Cavaliere and Taylor 2007) so that there is less size distortion in the tests.

An extension of the theory that is relevant in the case of quarterly data is to the seasonal unit root model

$$\left(1 - L^4\right) y_t = u_t. \tag{28}$$

Here, the polynomial $1 - L^4$ can be factored as $(1 - L)(1 + L)(1 + L^2)$, so that the unit roots (or roots on the unit circle) in (28) occur at 1, -1, i, and $-i$, corresponding to the annual $(L = 1)$ frequency, the semi-annual $(L = -1)$ frequency, and the quarter and three quarter annual $(L = i, -i)$ frequency respectively. Quarterly differencing, as in (28), is used as a seasonal adjustment device, and it is of interest to test whether the data supports the implied hypothesis of the presence of unit roots at these seasonal frequencies. Other types of seasonal processes, say monthly data, can be analysed in the same way. Tests for seasonal unit roots within the particular context of (28) were studied by Hylleberg et al. (1990), who extended the parametric *ADF* test to the case of seasonal unit roots. In order to accommodate fourth differencing, the autoregressive model is written in the new form

$$\begin{aligned} \Delta_4 y_t = {} & \alpha_1 y_{1t-1} + \alpha_2 y_{2t-1} + \alpha_3 y_{3t-2} \\ & + \alpha_4 y_{3t-1} + \sum\nolimits_{i=1}^{p} \phi_i \Delta_4 y_{t-i} + \varepsilon_t, \end{aligned} \tag{29}$$

where $\Delta_4 = 1 - L^4$, $y_{1t} = (1 + L)(1 + L^2)y_t$, $y_{2t} = -(1 - L)(1 + L^2)y_t$, and $y_{3t} = -(1 - L^2)y_t$. The transformed data y_{1t}, y_{2t}, y_{3t} retain the unit root at the zero frequency (long run), the semi-annual frequency (two cycles per year), and the annual frequency (one cycle per year). When $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$, there are unit roots at the zero and seasonal frequencies. To test the hypothesis of a unit root $(L = 1)$ in this seasonal model, a t-ratio test of $\alpha_1 = 0$ is used. Similarly, the test for

a semi-annual root ($L = -1$) is based on a t-ratio test of $\alpha_2 = 0$, and the test for an annual root on the t-ratios for $\alpha_3 = 0$ or $\alpha_4 = 0$. If each of the α's is different from zero, then the series has no unit roots at all and is stationary. Details of the implementation of this procedure are given in Hylleberg et al. (1990), the limit theory for the tests is developed in Chan and Wei (1988), and Ghysels and Osborne (2001) provide extensive discussion and applications.

Most empirical work on unit roots has relied on classical tests of the type described above. But Bayesian methods are also available and appear to offer certain advantages like an exact finite sample analysis (under specific distributional assumptions) and mass point posterior probabilities for break point analysis. In addressing the problem of trend determination, traditional Bayes methods may be employed such as the computation of Bayesian confidence sets and the use of posterior odds tests. In both cases prior distributions on the parameters of the model need to be defined and posteriors can be calculated either by analytical methods or by numerical integration. If (18) is rewritten as

$$y_t = \rho y_{t-1} + \sum_{1}^{k-1} \phi_i \Delta y_{t-i} + e_t \qquad (30)$$

then the posterior probability of the nonstationary set $\{\rho \geq 1\}$ is of special interest in assessing the evidence in support of the presence of a stochastic trend in the data. Posterior odds tests typically proceed with 'spike and slab' prior distributions (π) that assign an atom of mass such as $\pi(\rho = 1) = \theta$ to the unit-root null and a continuous distribution with mass $1 - \theta$ to the stationary alternative, so that $\pi(-1 < \rho < 1) = 1 - \theta$. The posterior odds then show how the prior odds ratio $\theta/(1 - \theta)$ in favour of the unit root is updated by the data.

The input of information via the prior distribution, whether deliberate or unwitting, is a major reason for potential divergence between Bayesian and classical statistical analyses. Methods of setting an objective correlative in Bayesian analysis through the use of model-based, impartial reference priors that accommodate nonstationarity are therefore of substantial interest. These were explored in Phillips (1991a), where many aspects of the subject are discussed. The subject is controversial, as the attendant commentary on that paper and the response (Phillips 1991b) reveal. The simple example of a Gaussian autoregression with a uniform prior on the autoregressive coefficient ρ and with an error variance σ^2 that is known illustrates one central point of controversy between Bayesian and classical inference procedures. In this case, when the prior on ρ is uniform, the posterior for ρ is Gaussian and symmetric about the maximum likelihood estimate $\hat{\rho}$ (Sims and Uhlig 1991), whereas the sampling distribution of $\hat{\rho}$ is biased downwards and skewed with a long left-hand tail. Hence, if the calculated value of $\hat{\rho}$ were found to be $\hat{\rho} = 1$, then Bayesian inference effectively assigns a 50 per cent posterior probability to stationarity $\{|\rho| < 1\}$, whereas classical methods, which take into account the substantial downward bias in the estimate $\hat{\rho}$, indicate that the true value of ρ is much more likely to be in the explosive region $\{\rho > 1\}$.

Another major point of difference is that the Bayesian posterior distribution is asymptotically Gaussian under very weak conditions, which include cases where there are unit roots ($\rho = 1$), whereas classical asymptotics for $\hat{\rho}$ are non-standard, as in (20). These differences are explored in Kim (1994), Phillips and Ploberger (1996) and Phillips (1996). The unit root case is one of very few instances where Bayesian and classical asymptotic theory differ. The reason for the difference in the unit root case is that Bayesian asymptotics rely on the local quadratic shape of the likelihood and condition on a given trajectory, whereas classical asymptotics rely on functional laws such as (9), which take into account the persistence in unit root data which manifest in the limiting trajectory.

Empirical illustrations of the use of Bayesian methods of trend determination for various macroeconomic and financial time series are given in DeJong and Whiteman (1991a, b), Schotman and van Dijk (1991) and Phillips (1991a, 1992), the latter implementing an objective model-based approach. Phillips and Ploberger (1994, 1996) develop Bayes tests, including an asymptotic

information criterion PIC (posterior information criterion) that extends the Schwarz (1978) criterion BIC (Bayesian information criterion) by allowing for potential nonstationarity in the data (see also Wei 1992). This approach takes account of the fact that Bayesian time series analysis is conducted conditionally on the realized history of the process. The mathematical effect of such conditioning is to translate models such as (30) to a 'Bayes model' with time-varying and data-dependent coefficients, that is,

$$y_{t+1} = \hat{\rho}_t y_t + \sum_{1}^{k-1} \hat{\phi}_{it} \Delta y_{t-i} + e_t, \qquad (31)$$

where $\left(\hat{\rho}_t, \hat{\phi}_{it}; i = 1, \ldots, k-1\right)$ are the latest best estimates of the coefficients from the data available to point 't' in the trajectory. The 'Bayes model' (31) and its probability measure can be used to construct likelihood ratio tests of hypotheses such as the unit root null $\rho = 1$, which relate to the model selection criterion PIC. Empirical illustrations of this approach are given in Phillips (1994, 1995).

Nonstationarity is certainly one of the most dominant and enduring characteristics of macroeconomic and financial time series. It therefore seems appropriate that this feature of the data be seriously addressed both in econometric methodology and in empirical practice. However, until the 1980s this was not the case. Before 1980 it was standard empirical practice in econometrics to treat observed trends as simple deterministic functions of time. Nelson and Plosser (1982) challenged this practice and showed that observed trends can be better modelled if one allows for stochastic trends even when there is some deterministic drift. Since their work there has been a continuing reappraisal of trend behaviour in economic time series and substantial development in the econometric methods of nonstationary time series. But the general conclusion that stochastic trends are present as a component of many economic and financial time series has withstood extensive empirical study.

This article has touched only a part of this large research field and traced only the main ideas involved in unit root modelling and statistical testing. This overview also does not cover the large and growing field of panel unit root testing and panel stationarity tests. The reader may consult the following review articles devoted to various aspects of the field for additional coverage and sources: (*a*) on unit roots: Phillips (1988b), Diebold and Nerlove (1990), Dolado et al. (1990), Campbell and Perron (1991), Stock (1994b), Phillips and Xiao (1998), and Byrne and Perman (2006); (*b*) on panel unit roots: Phillips and Moon (1999), Baltagi and Kao (2000), Choi (2001), Hlouskova and Wagner (2006); and (*c*) special journal issues of the *Oxford Bulletin of Economics and Statistics* (1986; 1992), *the Journal of Economic Dynamics and Control* (1988), *Advances in Econometrics* (1990), *Econometric Reviews* (1992), and *Econometric Theory* (1994).

See Also

- ARCH Models
- Bayesian Time Series Analysis
- Cointegration
- Markov Processes
- Martingales
- Present Value
- Statistical Inference

Bibliography

Andrews, D.W.K. 1993. Exactly median-unbiased estimation of first-order autoregressive/unit root models. *Econometrica* 61: 139–166.

Andrews, D.W.K., and P. Guggenberger. 2006. *Asymptotics for stationary very nearly unit root processes*. Mimeo: Yale University.

Bai, J. 1997. Estimating multiple break one at a time. *Econometric Theory* 13: 315–352.

Bai, L., and P. Perron. 1998. Estimating and testing linear models with multiple structural changes. *Econometrica* 66: 47–78.

Bai, L., and P. Perron. 2006. Multiple structural change models: A simulation analysis. In *Econometric theory and practice*, ed. Dean Corbae, B.E. Hansen, and S.N. Durlauf. New York: Cambridge University Press.

Baltagi, B.H., and C. Kao. 2000. Nonstationary panels, cointegration in panels and dynamic panels: A survey. *Advances in Econometrics* 15: 7–51.

Banerjee, A., R. Lumsdaine, and J. Stock. 1992. Recursive and sequential tests of the unit root and trend break hypotheses: Theory and international evidence. *Journal of Business and Economic Statistics* 10: 271–287.

Beare, B.K. 2006. *Unit root testing with unstable volatility.* Mimeo: Yale University.

Berkes, I., and L. Horváth. 2006. Convergence of integral functionals of stochastic processes. *Econometric Theory* 22: 304–322.

Beveridge, S., and C.R. 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.

Bhargava, A. 1986. On the theory of testing for unit roots in observed time series. *Review of Economic Studies* 52: 369–384.

Box, G.E.P. and Jenkins, G.M. 1976. *Time series analysis: Forecasting and control*, rev. edn. San Francisco: Holden Day.

Byrne, J.P., and R. Perman. 2006. *Unit roots and structural breaks: A survey of the literature*. Mimeo: University of Strathclyde.

Campbell, J.Y., and P. Perron. 1991. Pitfalls and opportunities: What macroeconomists should know about unit roots (with discussion). In *NBER macroeconomics annual 1991*, ed. O.J. Blanchard and S. Fischer. Cambridge, MA: MIT Press.

Campbell, J.Y., and R.J. Shiller. 1988. Interpreting cointegrated models. *Journal of Economic Dynamics and Control* 12: 505–522.

Cavaliere, G. 2004. Unit root tests under time-varying variances. *Econometric Reviews* 23: 259–292.

Cavaliere, G., and A.M.R. Taylor. 2007. Testing for unit roots in time series models with non-stationary volatility. *Journal of Econometrics* 140(2): 919–947.

Chan, N.H., and C.Z. Wei. 1987. Asymptotic inference for nearly nonstationary AR (1) processes. *Annals of Statistics* 15: 1050–1063.

Chan, N.H., and C.Z. Wei. 1988. Limiting distributions of least squares estimates of unstable autoregressive processes. *Annals of Statistics* 16: 367–401.

Choi, I. 2001. Unit roots for panel data. *Journal of International Money and Finance* 20: 249–272.

Choi, I., and P.C.B. Phillips. 1993. Testing for a unit root by frequency domain regression. *Journal of Econometrics* 59: 263–286.

Christiano, L.J. 1992. Searching for a break in GNP. *Journal of Business & Economic Statistics* 10: 237–250.

de Jong, R. 2004. Addendum to 'asymptotics for nonlinear transformations of integrated time series'. *Econometric Theory* 20: 623–635.

DeJong, D.N., and C.H. Whiteman. 1991a. Reconsidering trends and random walks in macroeconomic time series. *Journal of Monetary Economics* 28: 221–254.

DeJong, D.N., and C.H. Whiteman. 1991b. The temporal stability of dividends and stock prices: evidence from the likelihood function. *American Economic Review* 81: 600–617.

Dickey, D.A. 1976. Estimation and hypothesis testing in nonstationary time series. Ph.D. thesis, Iowa State University.

Dickey, D.A., and W.A. Fuller. 1979. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association* 74: 427–431.

Dickey, D.A., and W.A. Fuller. 1981. Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica* 49: 1057–1072.

Diebold, F.X., and M. Nerlove. 1990. Unit roots in economic time series. *Advances in Econometrics* 8: 3–70.

Dolado, J.J., T. Jenkinson, and S. Sosvilla-Rivero. 1990. Cointegration and unit roots. *Journal of Economic Surveys* 4: 249–273.

Duffie, D. 1988. *Security markets: Stochastic models*. San Diego: Academic Press.

Elliott, G., T.J. Rothenberg, and J.H. Stock. 1995. Efficient tests of an autoregressive unit root. *Econometrica* 64: 813–836.

Engle, R.F., and T. Bollerslev. 1986. Modeling the persistence of conditional variances. *Econometric Reviews* 5: 1–50.

Engle, R.F., and C.W.J. Granger. 1987. Co-integration and error correction: Representation, estimation, and testing. *Econometrica* 55: 251–276.

Evans, G.B.A., and N.E. Savin. 1981. Testing for unit roots: 1. *Econometrica* 49: 753–779.

Evans, G.B.A., and N.E. Savin. 1984. Testing for unit roots: 2. *Econometrica* 52: 1241–1269.

Fuller, W.A. 1976. *Introduction of statistical time series*. New York: Wiley.

Ghysels, E., and D.R. Osborn. 2001. *Econometric analysis of seasonal time series*. Cambridge: Cambridge University Press.

Giraitis, L., and P.C.B. Phillips. 2006. Uniform limit theory for stationary autoregression. *Journal of Time Series Analysis* 27: 51–60.

Gouriéroux, C., E. Renault, and N. Touzi. 2000. Calibration by simulation for small sample bias correction. In *Simulation-based inference in econometrics: Methods and applications*, ed. R.S. Mariano, T. Schuermann, and M. Weeks. Cambridge: Cambridge University Press.

Gouriéroux, C., P.C.B. Phillips, and J. Yu. 2007. Indirect inference for dynamic panel models. *Discussion Paper No. 1550*, Cowles Foundation, Yale University.

Hall, R.E. 1978. Stochastic implications of the life cycle-permanent income hypothesis. *Journal of Political Economy* 86: 971–987.

Hall, A. 1989. Testing for a unit root in the presence of moving average errors. *Biometrika* 76: 49–56.

Hall, P., and C.C. Heyde. 1980. *Martingale limit theory and its application*. New York: Academic Press.

Harris, D., S. Leybourne, and B. McCabe. 2007. Modified KPSS tests for near integration. *Econometric Theory* 23: 355–363.

Hlouskova, J., and M. Wagner. 2006. The performance of panel unit root and stationarity tests: Results from a

large scale simulation study. *Econometric Reviews* 25: 85–116.

Hong, S.H., and P.C.B. Phillips. 2005. Testing linearity in cointegrating relations with an application to purchasing power parity. Discussion Paper No. 1541, Cowles Foundation, Yale University.

Hu, L., and P.C.B. Phillips. 2004. Dynamics of the federal funds target rate: a nonstationary discrete choice approach. *Journal of Applied Econometrics* 19: 851–867.

Hylleberg, S., R.F. Engle, C.W.J. Granger, and S. Yoo. 1990. Seasonal integration and cointegration. *Journal of Econometrics* 44: 215–238.

Ibragimov, R., and P.C.B. Phillips. 2007. Regression asymptotics using martingale convergence methods. *Discussion Paper No. 1473*, Cowles Foundation, Yale University.

Johansen, S. 1988. Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control* 12: 231–254.

Kapetanios, G. 2005. Unit root testing against the alternative hypothesis of up to m structural breaks. *Journal of Time Series Analysis* 26: 123–133.

Karlsen, H.A., T. Myklebust, and D. Tjøstheim. 2007. Nonparametric estimation in a nonlinear cointegration model. *Annals of Statistics* 35(1).

Kasparis, I. 2004. *Detection of functional form misspecification in cointegrating relations*. Mimeo: University of Nottingham.

Kim, J.Y. 1994. Bayesian asymptotic theory in a time series model with a possible nonstationary process. *Econometric Theory* 10: 764–773.

Kwiatkowski, D., P.C.B. Phillips, P. Schmidt, and Y. Shin. 1992. Testing the null of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics* 54: 159–178.

Lee, J.S. 1996. On the power of stationary tests using optimal bandwidth estimates. *Economics Letters* 51: 131–137.

Lee, C.C., and P.C.B. Phillips. 1994. *An ARMA prewhitened long-run variance estimator*. Mimeo: Yale University.

Leeb, H., and B. Potscher. 2001. The variance of an integrated process need not diverge to infinity and related results on partial sums of stationary processes. *Econometric Theory* 17: 671–685.

MacKinnon, J.G. 1994. Approximate asymptotic distribution functions for unit-root and cointegration tests. *Journal of Business and Economic Statistics* 12: 167–176.

Müller, U. 2005. Size and power of tests for stationarity in highly autocorrelated time series. *Journal of Econometrics* 128: 195–213.

Müller, U., and G. Elliott. 2003. Tests for unit roots and the initial condition. *Econometrica* 71: 1269–1286.

Nelson, D.B. 1990. Stationarity and persistence in the GARCH (1, 1) model. *Econometric Theory* 6: 318–334.

Nelson, C.R., and C. Plosser. 1982. Trends and random walks in macroeconomic time series: Some evidence and implications. *Journal of Monetary Econometrics* 10: 139–162.

Ng, S., and P. Perron. 1995. Unit root tests in ARMA models with data dependent methods for the selection of the truncation lag. *Journal of the American Statistical Association* 90: 268–281.

Ng, S., and P. Perron. 2001. Lag length selection and the construction of unit root tests with good size and power. *Econometrica* 69: 1519–1554.

Ouliaris, S., J.Y. Park, and P.C.B. Phillips. 1989. Testing for a unit root in the presence of a maintained trend. In *Advances in econometrics and modelling*, ed. B. Raj. Norwell, MA: Kluwer.

Park, J.Y. 1990. Testing for unit roots and cointegration by variable addition. *Advances in Econometrics* 8: 107–133.

Park, J.Y., and P.C.B. Phillips. 1988. Statistical inference in regressions with integrated processes: Part I. *Econometric Theory* 4: 468–497.

Park, J.Y., and P.C.B. Phillips. 1989. Statistical inference in regressions with integrated processes: Part II. *Econometric Theory* 5: 95–131.

Park, J.Y., and P.C.B. Phillips. 1999. Asymptotics for nonlinear transformations of integrated time series. *Econometric Theory* 15: 269–298.

Park, J.Y., and P.C.B. Phillips. 2000. Nonstationary binary choice. *Econometrica* 68: 1249–1280.

Park, J.Y., and P.C.B. Phillips. 2001. Nonlinear regressions with integrated time series. *Econometrica* 69: 1452–1498.

Park, J.Y., and J. Sung. 1994. Testing for unit roots in models with structural change. *Econometric Theory* 10: 917–936.

Perron, P. 1989. The great crash, the oil price shock and the unit root hypothesis. *Econometrica* 57: 1361–1401.

Phillips, P.C.B. 1986. Understanding spurious regressions in econometrics. *Journal of Econometrics* 33: 311–340.

Phillips, P.C.B. 1987a. Time series regression with a unit root. *Econometrica* 55: 277–301.

Phillips, P.C.B. 1987b. Towards a unified asymptotic theory of autoregression. *Biometrika* 74: 535–547.

Phillips, P.C.B. 1988a. Regression theory for near-integrated time series. *Econometrica* 56: 1021–1044.

Phillips, P.C.B. 1988b. Multiple regression with integrated processes. In *Statistical inference from stochastic processes*, ed. N.U. Prabhu. *Contemporary Mathematics* 80: 79–106.

Phillips, P.C.B. 1991a. To criticize the critics: An objective Bayesian analysis of stochastic trends. *Journal of Applied Econometrics* 6: 333–364.

Phillips, P.C.B. 1991b. Bayesian routes and unit roots: de rebus prioribus semper est disputandum. *Journal of Applied Econometrics* 6: 435–474.

Phillips, P.C.B. 1992. The long-run Australian consumption function reexamined: An empirical exercise in Bayesian inference. In *Long run equilibrium and*

macroeconomic modelling, ed. C. Hargreaves. Cheltenham: Edward Elgar.

Phillips, P.C.B. 1994. Model determination and macroeconomic activity. Fisher–Schultz Lecture to the European Meetings of the Econometric Society, Maastricht. Discussion Paper No. 1083, Cowles Foundation, Yale University.

Phillips, P.C.B. 1995. Bayesian model selection and prediction with empirical applications. *Journal of Econometrics* 69: 289–332.

Phillips, P.C.B. 1996. Econometric model determination. *Econometrica* 64: 763–812.

Phillips, P.C.B. 1998. New tools for understanding spurious regressions. *Econometrica* 66: 1299–1326.

Phillips, P.C.B. 2001. New unit root asymptotics in the presence of deterministic trends. *Journal of Econometrics* 11: 323–353.

Phillips, P.C.B. 2006. *When the tail wags the unit root limit distribution*. Mimeo: Yale University.

Phillips, P.C.B., and S.N. Durlauf. 1986. Multiple time series regression with integrated processes. *Review of Economic Studies* 53: 473–496.

Phillips, P.C.B., and C. Han. 2007. Gaussian inference in AR(1) time series with or without a unit root. *Econometric Theory* 24(3).

Phillips, P.C.B., and B.E. Hansen. 1990. Statistical inference in instrumental variables regression with I(1) processes. *Review of Economic Studies* 57: 99–125.

Phillips, P.C.B., and T. Magdalinos. 2007. Limit theory for moderate deviations from a unit root. *Journal of Econometrics* 136: 115–130.

Phillips, P.C.B., and H.R. Moon. 1999. Linear regression limit theory for nonstationary panel data. *Econometrica* 67: 1057–1111.

Phillips, P.C.B., and H.R. Moon. 2000. Nonstationary panel data analysis: an overview of some recent developments. *Econometric Reviews* 19: 263–286.

Phillips, P.C.B., and S. Ouliaris. 1990. Asymptotic properties of residual based tests for cointegration. *Econometrica* 58: 165–193.

Phillips, P.C.B., and P. Perron. 1988. Testing for a unit root in time series regression. *Biometrika* 75: 335–346.

Phillips, P.C.B., and W. Ploberger. 1994. Posterior odds testing for a unit root with data-based model selection. *Econometric Theory* 10: 774–808.

Phillips, P.C.B., and W. Ploberger. 1996. An asymptotic theory of Bayesian inference for time series. *Econometrica* 64: 381–413.

Phillips, P.C.B., and V. Solo. 1992. Asymptotics for linear processes. *Annals of Statistics* 20: 971–1001.

Phillips, P.C.B., and Z. Xiao. 1998. A primer on unit root testing. *Journal of Economic Surveys* 12: 423–469.

Phillips, P.C.B., J.Y. Park, and Y. Chang. 2004. Nonlinear instrumental variable estimation of an autoregression. *Journal of Econometrics* 118: 219–246.

Ploberger, W. 2004. A complete class of tests when the likelihood is locally asymptotically quadratic. *Journal of Econometrics* 118: 67–94.

Pötscher, B.M. 2004. Nonlinear functions and convergence to Brownian motion: Beyond the continuous mapping theorem. *Econometric Theory* 20: 1–22.

Robinson, P.M. 1995. Gaussian semiparametric estimation of long range dependence. *Annals of Statistics* 23: 1630–1661.

Said, S.E., and D.A. Dickey. 1984. Testing for unit roots in autoregressive moving average models of unknown order. *Biometrika* 71: 599–608.

Sargan, J.D., and A. Bhargava. 1983. Testing residuals from least squares regression for being generated by the Gaussian random walk. *Econometrica* 51: 153–174.

Schmidt, P., and P.C.B. Phillips. 1992. LM tests for a unit root in the presence of deterministic trends. *Oxford Bulletin of Economics and Statistics* 54: 257–287.

Schotman, P., and H.K. van Dijk. 1991. A Bayesian analysis of the unit root in real exchange rates. *Journal of Econometrics* 49: 195–238.

Shimotsu, K., and P.C.B. Phillips. 2005. Exact local whittle estimation of fractional integration. *Annals of Statistics* 33: 1890–1933.

Sims, C.A., and H. Uhlig. 1991. Understanding unit rooters: A helicopter tour. *Econometrica* 59: 1591–1599.

So, B.S., and D.W. Shin. 1999. Cauchy estimators for autoregressive processes with applications to unit root tests and confidence intervals. *Econometric Theory* 15: 165–176.

Solo, V. 1984. The order of differencing in ARIMA models. *Journal of the American Statistical Association* 79: 916–921.

Stock, J.H. 1991. Confidence intervals for the largest autoregressive root in US macroeconomic time series. *Journal of Monetary Economics* 28: 435–459.

Stock, J. 1994a. Deciding between I(1) and I(0). *Journal of Econometrics* 63: 105–131.

Stock, J.H. 1994b. Unit roots, structural breaks and trends. In *Handbook of econometrics*, ed. R.F. Engle and D. McFadden, Vol. 4. Amsterdam: North-Holland.

Stock, J.H. 1999. A class of tests for integration and cointegration. In *Cointegration, causality and forecasting: A Festschrift in honour of Clive W. J. Granger*, ed. R.F. Engle and H. White. Oxford: Oxford University Press.

Stock, J.H., and M.W. Watson. 1988. Variable trends in economic time series. *Journal of Economic Perspectives* 2(3): 147–174.

Sul, D., P.C.B. Phillips, and C.-Y. Choi. 2006. Prewhitening bias in HAC estimation. *Oxford Bulletin of Economics and Statistics* 67: 517–546.

Wang, Q., and P.C.B. Phillips. 2006. Asymptotic theory for local time density estimation and nonparametric cointegrating regression. Discussion Paper no. 1594, Cowles Foundation, Yale University.

Wei, C.Z. 1992. On predictive least squares principles. *Annals of Statistics* 20: 1–42.

Xiao, Z., and P.C.B. Phillips. 1998. An ADF coefficient test for a unit root in ARMA models of unknown order

with empirical applications to the U.S. economy. *Econometrics Journal* 1: 27–43.
Zivot, E., and D.W.K. Andrews. 1992. Further evidence on the great crash, the oil price shock and the unit root hypothesis. *Journal of Business and Economic Statistics* 10: 251–270.

United States, Economics in (1776–1885)

Stephen Meardon

Abstract

Economics in the United States before 1885 was not a discipline of credentialled professionals. Its debates were published in political pamphlets and newspaper editorials as well as college textbooks and scholarly treatises. Its principal project was to set the boundaries of a doctrinal system of economic liberalism and determine the appropriate functions of government. The system was characterized by the sanctity of private property, the celebration of individual labour, and the harmony of the economic and moral orders. Those parts of the system that were imported from abroad were adapted, sometimes ingeniously, to economic and political circumstances at home.

Keywords

American Economic Association; American Free Trade League; Banknotes; Bastiat, F.; Bryant, W. C.; Cardozo, J. N.; Carey, H. C.; Carey, M.; Catallactics; Combinations; Economic liberalism; Ely, R. T.; Enlightenment; Finney, C.; Franklin, B.; Free banking; Free labour doctrine; Free trade; George, H.; Great Awakening; Hamilton, A.; James, E.; Jefferson, T.; Land tax; Leavitt, J.; Liberty Party; List, F.; Madison, Bishop J.; Malthus's theory of population; McCulloch, J. R.; Physiocracy; Paine, T.; Paper money; Patten, S. N.; Perry, A. L.; Petty, W.; Political economy; Protectionism; Public works; Raymond, D.; Rent; Ricardo, D.; Say, J.-B.; Second Great Awakening; Slavery; Smith, A.; Tucker, G.; Turgot, A.R. J.; United States, economics in; Vattel, E.; Wayland, F.; Wealth; Wells, D. A

JEL Classifications

B1

Pre-independence to 1820

Before US independence, the middle Atlantic colonies offered more fertile ground than New England or the South for ideas of political economy from Britain and the European Continent to take root. Eighteenth-century Philadelphia was rivalled only by New York and Boston in its population, which grew from over 4000 in 1700 to over 28,000 in 1790. It was unrivalled in its cosmopolitanism. The religious diversity of Pennsylvania promoted tolerance of notions that, especially from the perspective of the New England theocracy, were tainted by deism or secularism. One consequence was a thriving press.

Philadelphia, therefore, was the destination of Benjamin Franklin (1706–1790) when he ran away from his printer's apprenticeship in Boston at the age of 17. The many ideas he published in his new situation included his adaptation of the work of Sir William Petty to American commercial policy. Franklin's *A Modest Enquiry Into the Nature and Necessity of a Paper Currency* (1729), following Petty's *Discourse on Political Arithmetic* (1690), emphasized the importance of the quantity of specie in circulation to commercial prosperity, and a favourable balance of trade to the quantity of specie. Beyond Petty, Franklin argued that issuance of paper money would at once substitute for the specie that was then lacking and, by stimulating production and exports, improve the balance of trade – and thereby garner more specie. Paper money, though commonly disdained, could effect a virtuous circle of commercial vitality and specie accumulation.

In 1757, when Franklin was renowned in America and abroad as a scientist and statesman, he was appointed the Pennsylvania Assembly's

emissary to the Crown. The governor of the colony had vetoed the issuance of paper money and other fiscal measures desired by the colonists; Franklin was to convey their complaints. Following the onerous internal taxes established by the Stamp Act of 1765 and the import duties of the Townshend Acts of 1767, they would have more cause to complain. The popular response was 'No taxation without representation' coupled with the colonists' mutual agreement to refuse the importation of British goods. Franklin's own response, more nuanced but to the same effect, drew upon the Physiocratic notion of the pre-eminence of agriculture that had been impressed upon him in France by Anne Robert Jacques Turgot. In his *Positions to be Examined, concerning National Wealth* (1769), Franklin explained that trade is fair when both parties know the value of the traded goods, meaning the value of labour embodied in them; and value is most knowable when the goods consist of agriculture, not manufactures, because agriculture is derived more directly from labour. The implication was that by creating a captive market in America for British manufactured goods, Britain was attempting to hoodwink the colonists. The colonists' eschewal of legal British trade and their substitution of smuggled foreign goods, therefore, were justified. Besides, Franklin added, the frugality necessitated by the boycott would make domestic farmers and tradesmen more productive (Dorfman 1966, vol. 1, 191–3).

While Franklin engaged the political-economic ideas of the Enlightenment in public argument, he also sought to disseminate them in higher education. His Academy of Philadelphia, later to become the University of Pennsylvania, commenced classes in 1751 with a non-sectarian board of trustees, a progressive Scottish cleric as president, and a curriculum including governance and commerce. The institution's secular orientation and curriculum stood in contrast to those of Harvard (1636), the College of William and Mary (1693), and Yale (1701), which were aligned closely with the Congregational or Anglican churches, directed to the training of ministers and missionaries, and uncongenial to the innovations in moral philosophy and legal and commercial thought from Britain and Europe. Even Princeton (1746), Columbia (1754), and Brown (1764), which were intended to embody a new spirit of religious tolerance, succumbed ultimately to a more constrained orthodoxy. Indeed, so too did Franklin's Academy. But the Revolutionary war shook the orthodoxy. Franklin's curricular plan made inroads at William and Mary in 1779 (O'Connor 1944, pp. 64–6). Bishop James Madison (1749–1812), the College's president, cousin of the fourth US president, and a man of radical political views notwithstanding his prominence in the Episcopal Church, became America's first teacher of political economy (O'Connor 1944, pp. 20–1).

Madison is believed to have relied upon Emer de Vattel's *The Law of Nations, or the Principles of Natural Law Applied to the Conduct and to the Affairs of Nations and of Sovereigns* (1758). The book stood out for its maxims of sovereignty more than commercial policy. To Vattel, the sovereignty of the state is inalienable. While in some times and places the head of state may be a prince, nevertheless, 'the State is not, and can not be, a patrimony' (1758, pp. 29, 33). The sovereign is obliged to serve the state's interests, not his own. Although they are peripheral, the book also includes chapters on general commerce, the financing of public roads and canals, money and exchange, and commerce between nations. Vattel makes brief but notable mention of the desirability of defraying the expense of roads and canals from the general revenue of the entire nation, and of the ruler's obligation to encourage trade that is beneficial but to restrict that which is hurtful. In the latter category is that 'ruinous' trade which results in a negative balance, entailing more gold and silver leaving the country than entering it (1758, pp. 44, 43).

In Vattel's work, the political thought of the enlightenment is bonded to mercantile economic thought. Neither was particularly avant-garde for the 1770s. Thomas Paine (1737–1809) did not steel the colonists' patriotic will with 25 editions and 150,000 copies of his *Common Sense* (1776) by arguing merely that King George had not fulfilled adequately a sovereign's obligations to his subjects. Paine remonstrated that monarchy yields tyranny and war; only republican government

would serve the cause of liberty. The differences of Paine from Vattel in political economy are equally significant. 'Our plan is commerce,' Paine declared. It was no time to ponder which particular avenues and products and circumstances of trade were hurtful and which were not; it was time to throw off the commercial shackles of Britain altogether and trade with all the world. 'Trade flourishes best when it is free', he continued a short while later, 'and it is weak policy to attempt to fetter it' (Paine 1778, p. 153).

Paine wrote polemics for the literate multitudes. His work was not the stuff of college curricula; Adam Smith's *Wealth of Nations* was fitter for that purpose. Smith exploded the 'absurd' balance-of-trade doctrine (Smith 1789, p. 456), illuminating in more scholarly fashion than did Paine the virtues of commercial freedom, whether international or domestic, and nudging England towards the 'natural system of perfect liberty and justice' in the administration of its colonial trade (1789, p. 572). Regarding roads and canals and other means of facilitating commerce, Smith was disinclined to see them financed with the general revenue of the state. More consistent with his system was their financing by tolls or by local or provincial taxes (1789, pp. 682–3). The *Wealth of Nations* was taken up at William and Mary sometime between the mid-1780s and the 1790s. It was read, debated, and absorbed in America several years earlier, however: an edition was published in Philadelphia within a few years of the Revolution (O'Connor 1944, pp. 21–2).

Questions of foreign trade policy, and the role of the federal government in undertaking public works and institutions for promoting commerce, dominated the economic debate of the early republic. They were argued passionately because it was believed that upon their answers depended not only the nation's prosperity, but its character and survival. Thomas Jefferson (1743–1826) sought free trade between America and Europe less for the goods that Americans could obtain than for the occupations they would undertake in consequence. In his *Notes on the State of Virginia* (1785), Jefferson observed that, if trade had not been interrupted and manufactures stimulated by the Revolutionary War, Americans in greater numbers would work the land and buy their textiles from abroad. The advantage of making a living from the land lay in the republican virtues that husbandry instilled in the smallholder. 'Let our workshops remain in Europe', Jefferson offered; 'the loss by the transportation of commodities across the Atlantic will be made up in happiness and permanence of government' (Spiegel 1960, pp. 42–3).

Alexander Hamilton (1755–1804), Jefferson's ideological rival, did not repudiate the Virginian's vision of the free trade of American agriculture for European manufactures. He just denied that, under the circumstances, it was realizable. The European powers regularly imposed barriers to the importation of American products, and the several American states had little power to persuade or compel them to do otherwise. Hamilton's answer was to change the circumstances of American government. The unlikelihood of the Jeffersonian ideal became an argument for a strong central authority. In *Federalist* No. 11 (1788), Hamilton argued that adoption of the Constitution would create a central government with the power to bargain for commercial privileges abroad. If instead disunion persisted, 'we should then be compelled to content ourselves with the first price of our commodities, and to see the profits of our trade snatched from us to enrich our enemies and persecutors.'

Hamilton envisioned a larger manufacturing sector than did Jefferson, but he did not admit that the growth of manufactures would retard agriculture. In his famous *Report on Manufactures* (1791), composed during his term as the nation's first Secretary of the Treasury, he offered two reasons. First, even if manufactures slowed the extensive cultivation of land, they would 'promote a more steady and vigorous cultivation of the lands occupied'. Second, manufactures probably would not slow the extensive cultivation of land, anyway. Immigrants drawn to America by manufacturing jobs would eventually leave them for agriculture, attracted by the promise of independent proprietorship of land (Spiegel 1960, p. 33).

The question that followed was, 'Will manufacture develop by itself?' Hamilton thought not.

The scarcity of labour, high wages, and lack of capital conspired against American manufacturers. To survive against European competition would be difficult – but not hopeless. Raw materials were more abundant in the United States and, although wages were indeed higher, the greater expense of labour was counterbalanced by the expenses of transportation and customs duties for those who would import foreign goods. The scarcity of capital could be addressed by fostering a banking system and by the Treasury's issuance of debt, which would serve as a form of money. To Hamilton, as to Paine, the business of America was commerce; but Hamilton was not as squeamish as Paine and his ilk, including Jefferson, about using the powers of the newly constituted federal government to build the institutions and infrastructure of commerce and to encourage industries that might otherwise languish.

Hamilton's name was invoked often in ensuing controversies that exercised Americans and their representatives. Among them were questions of the federal government's role in financing 'internal improvements', including the national turnpike stretching from the Potomac to the Ohio River, authorized in 1806; the First Bank of the United States, the charter of which expired in 1811; the Second Bank of the United States, the fate of which was decided in the presidential election of 1832; and the 'American System' of tariff protection for domestic manufactures, which sparked debate through most of the 19th century. One of the many whom Hamilton inspired was Daniel Raymond (1786–1849), who became in 1820 the first American author of a systematic treatise on political economy.

From 1820 to the Civil War

Raymond penned his *Thoughts on Political Economy* in his free time as an underemployed Baltimore lawyer (Spiegel 1960, p. 55). His careful definition of national wealth as 'a capacity for acquiring the necessaries and comforts of life', not the value of tangible assets, was intended to combat the idea that the United States was wealthy by virtue mainly of its climate, soil, and nearly limitless territory. Raymond argued that other things mattered too, and more: population density, the distribution of private wealth, the capital and technology applied to agriculture, transport, and other industries, and above all 'the industrious habits of the people' (Spiegel 1960, p. 60). The list corresponded closely to the priorities for government action envisioned by Hamilton.

Raymond's distinction in American economics is that he was first in a class, not that he was influential. More credit for influence is due to Mathew Carey (1760–1839), an Irish publisher and controversialist who emigrated to Philadelphia in 1784 and soon re-established himself in the same vocations. On political economy, he wrote prolifically – but not systematically in the manner of Raymond, whose work he sponsored (Spiegel 1960, p. 55). Carey was concerned mainly with one issue: the tariff. His *Addresses of the Philadelphia Society for the Promotion of National Industry* (1819) responded to the country's financial crisis of 1819 with a call for further tariff protection for domestic industry. Although the tariff bill of 1820 did not pass, the protectionists' champion in Congress, Henry Clay, acknowledged Carey's importance to the ferment of the tariff of 1824 (Dorfman 1966, vol. 1, pp. 384, 389–90). Carey's influence was also exercised importantly, if indirectly, through the writings of Friedrich List (1789–1846), the German thinker who was renowned posthumously in his country and abroad for his protectionist 'national system' of political economy. List devised his system during a five-year sojourn in Pennsylvania from 1825 to 1830, in which time he became acquainted with Carey's ideas, lent them his own voice and authority, and propagated them through the protectionist periodical, the *National Gazette* (Dorfman 1966, vol. 2, p. 577).

The protectionist trend in trade policy that gathered strength in the mid-Atlantic states caused alarm in the South. Jacob Newton Cardozo (1786–1873), publisher of Charleston's *Southern Patriot,* addressed the tariff and other questions in light of David Ricardo's system. An American edition of Ricardo's *Principles of Political Economy and Taxation* was published in 1819;

another avenue of his large influence was the *Encyclopædia Britannica* article on 'Political Economy' by John Ramsay McCulloch, Ricardo's chief expositor in Britain. The article was republished in 1825 with introduction, summary, and annotations by Columbia College's Reverend John McVickar (O'Connor 1944, p. 136).

Cardozo's *Notes on Political Economy* appeared in 1826. To Cardozo, the unenlightened social arrangements of Europe – above all, impediments to the alienation of land – had obscured to Ricardo what was evident in the United States. In the absence of unwise customs and legislation, rent was never pernicious. Indeed, some payments that were nominally rent, but were really the returns to the ingenuity of the cultivator, were positively benign. Conversely, where rent was pernicious, it was due to legislation that fomented, in one way or another, 'monopoly'. In the United States, such legislation was the protective tariff. The genius of Cardozo's argument was that indicted protectionism even more forcefully than did Ricardo's system, thereby serving the interests of the cotton-exporting South and the entrepôt of Charleston, while casting Southern landowners in a productive rather than a parasitic role (Dorfman 1966, vol. 1, pp. 551–66).

Bank regulation and the sustenance of slavery preoccupied the South as much as the bane of protectionism. The question of greater urgency varied with the occasion. The 1828 'Tariff of Abominations' and the crisis stemming from South Carolina's 'nullification' of the law in 1832 were occasions of substantive political–economic debate – but so was President Jackson's veto, in 1832, of the proposed re-charter of the Second Bank of the United States. While the Democratic and Whig parties agreed that the currency should be convertible for gold, there were large differences between and even within them regarding the appropriate degrees competition and regulation of banks, especially those involved in issuing currency. They also differed on the function of the Second Bank of the United States as competitor, usurper, and regulator in the domain of private and state banks. Jackson's objection to the Second Bank of the United States was that its shareholders enjoyed a monopoly of the Federal government's bank transactions. Cardozo shared Jackson's anti-monopoly convictions, but appreciated the Bank's potential as a regulator of the note issuance of other banks, and thus of the recurring cycles of credit expansion, inflation, and contraction. Banks of deposit and discount – 'banking in its legitimate meaning' – should be subject to free competition, according to Cardozo, but banks of issue should be subject to severe restriction (Leiman 1966, p. 112). This exception to perfect freedom served, by means of a stable currency, to protect the value of property and promote trade. Cardozo's justified similarly his support for slavery. Abolitionism was a 'conspiracy against property', while slavery heightened the United States' comparative advantage in agriculture and its gains from free trade (Leiman 1966, p. 176–7).

By the 1820s and 30s, the inviolability of property and free trade were touchstones of economic thought not only in the South. In the North, too, notwithstanding its traditional allegiance to Hamiltonian federalism, the ideas had wide currency. Jefferson's Embargo Act of 1807, which prohibited exports in the (disappointed) expectation of compelling Britain and France to rescind their restraints on American trade, provoked ferocious opposition in the Northeast. Ideas of free trade, and economic liberalism more generally, sank their roots, notwithstanding the irony of their having been planted in opposition to one of their leading advocates. William Cullen Bryant (1794–1878), who began in 1826 his life's work as editor of the *New York Evening Post,* and concurrently one of the nation's most renowned poets and most prominent spokesmen for unfettered commercial freedom, cut his literary and economic teeth during the embargo. The first poem he published as a boy in rural Massachusetts was a widely circulated satire on the subject. 'In vain Mechanics ply their curious art, / And bootless mourn the interdicted mart', complained the 13-year-old Bryant. Jefferson's diplomacy, he mocked, 'His grand *'restrictive energies'* employs, / And wisely regulating trade – destroys' (Bryant 1808).

To gain acceptance in Northern colleges, however, economic liberalism had to be disassociated

from French political radicalism and garbed in religious piety. The first event was already accomplished in the mid-1790s, as the enthusiasm of American partisans of the cause of the French Revolution diminished. The second was the outcome of a longer process that started at the same time and gathered momentum as political economy made its way into college curricula as a subject in its own right. Most colleges introduced political economy as such, and explicitly, in the 1820s: although Harvard, Columbia, and Princeton did so between 1817 and 1819, they were followed by Dickinson, Pittsburgh, Bowdoin, Amherst, Yale, Rutgers, the US Military Academy, Geneva (later renamed Hobart), Williams, Dartmouth, Brown, Union, and Hamilton, all in the years 1822–1827 (O'Connor 1944, p. 100). In most cases the favoured text was Jean Baptiste Say's *Treatise on Political Economy.* Because Say disapproved of several of the social changes in France since the Revolution, he passed the political test. Because he was a protestant Huguenot, he passed the religious one – although it was a close call (O'Connor 1944, pp. 120, 133–4).

The religious test became more stringent in the 1830s. Religious revivals had recurred with varying intensity since the Great Awakening of the 1730s–1750s, but none since had swept through the country with as much intensity as those commencing in western New York in 1826 and spreading outward, nationwide, through roughly the decade that followed (Cole 1954, pp. 75–6). 'Harvests' of new souls were most bountiful among the middle class and young in small towns and rural areas (Cole 1954, p. 80). Colleges were especially fertile ground: the demographics were right, and in most cases so too was the geography. College presidents, besides, saw revivals as a way of boosting their institutions' visibility and prestige. Students and faculty were particularly amenable to the particular character of the Second Great Awakening. In past eras evangelism sought the salvation of the individual; now it sought that of the nation, even the world. The most popular evangelist of the period, Charles Finney, called upon his listeners to strive for temperance, moral reform in general, and the abolition of slavery (Cole 1954, p. 77). The last of these causes, even within the flock, was controversial. That the awakened moral sensibilities demanded an end to slavery, there was consensus; that they required the North to impose abolition upon an unwilling South, there was not. As evangelical religion asserted itself in the same citadels where political economy was establishing its foothold, the subject required a text reflecting at least the sensibilities, if not always the objectives, of its teachers and students.

That text was *Elements of Political Economy* (1837a, b) by Reverend Francis Wayland (1796–1865), president of Brown University. Wayland, who had already published his *Elements of Moral Science* in 1835, was the exemplary teacher of political economy for two generations of antebellum men. Political economy partook of moral philosophy; the two subjects were taught proximately, but separately, in the students' junior or senior year. In the *Moral Science* students learned the nature of their moral constitution, including the power and limits of their unaided conscience to discern right from wrong in their private and public conduct. Inasmuch as conscience was limited, there was recourse first to natural religion and finally to revealed religion. Revealed religion was the study of God's word as manifested in the Bible; natural religion was the study of God's design as manifested in the consequences of human behaviour. Taking intemperance as an example, Wayland illustrated that one could survey the vice and poverty that resulted from the consumption of liquor, consider that God has a design relating all causes to their effects, and thereby infer that God's design forbids intemperance 'as though He had said so by a voice from heaven' (Wayland 1837a, b, p. 120). Of course, one could study other sorts of behaviour and their consequences – production, exchange, distribution, and individual and public consumption – and infer God's design therein, too. To do so required another volume: that was the thrust of Wayland's *Political Economy.*

Wayland taught that the economic role of government is 'to construct the arrangements of society as to give free scope to the laws of Divine Providence' (O'Connor 1944, p. 189). Man must

be free to produce what he will and dispose of the product as he pleases. International trade should be unencumbered by protective tariffs. Although the incorporation of banks should be made conditional on arrangements to ensure the convertibility of their notes into specie, nobody who is willing to accept the conditions should be refused a bank charter. Poor laws supporting those capable of work were founded on the premise the rich were obliged to support the poor by law instead of by charity, and as such were a violation of property. Combinations of workers, like combinations of capital owners, were oppressive, even tyrannical. All that man required was the right to sell his labour and invest his capital freely. Legislation granting one party or the other any additional privileges was 'impolicy' (Wayland 1841, pp. 108–23).

As a treatise in political economy, Wayland's book was rivalled, after 1848, by John Stuart Mill's *Principles of Political Economy.* Mill conveyed the requisite ethical tone, if not a religious one, and allowed sufficient qualifications to the cases for free trade and strict constraints on banknote issuance to appeal to a broader swathe of political opinion than did Wayland. But Mill's book was too lengthy to serve as a college text (Dorfman 1966, vol. 2, p. 710). In that role, Wayland's was unrivalled. The Elements of Political Economy went through no fewer than 24 printings in Boston and New York between 1837 and 1875, with three more in London and a translation to Hawaiian (O'Connor 1944, p. 174).

For his large readership and his tightly connected system of economic liberty, Wayland has been designated the 'Ricardo of evangelists' (Cole 1954, p. 178). The designation is not, notably, the 'evangelist of Ricardo'. The last would not apply, for in one crucial respect Wayland was not the system's most fervent expositor. Those who adhered to the doctrine that man's liberty to reap the fruits of his own industry was sacrosanct, and who followed its implications wherever they might lead, found their way in short order to the question of slavery and determined that it could not be countenanced. 'Free labor' was their slogan. Wayland did not travel with them so far: he was silent on the slavery question for the better part of a generation. He believed that the Constitution left slavery a matter for states to decide (Dorfman 1966, vol. 1, p. 760) – but even as a matter of state law, Wayland was not exercised by the subject.

The Reverend Joshua Leavitt (1794–1873) proselytized a more thoroughgoing interpretation of the free labour doctrine. In the same year that Wayland first published the *Elements of Political Economy,* Leavitt became editor of the *Emancipator*, an organ urging free trade, cheap postage, temperance, and above all, the demise of slavery by political action. The *Emancipator* also advocated the election of James G. Birney of the newly formed Liberty Party as President of the United States in 1840 and 1844 (Cole 1954, p. 40). Although the Liberty Party was ineffectual and Birney unsuccessful, Leavitt's periodical and the presidential campaigns it championed marked the beginnings of a longer effort. Its objectives were at once to intensify opposition to slavery and to peel the opponents' support away from the dominant Whig and Democratic parties.

The effort began to show effects once the Liberty Party was eclipsed by the Free Soil Party in 1848 and 1852. Northern 'Barnburner' Democrats, who sided with the majority of their party in opposition to the protective tariff, national bank, and internal improvements, but disfavoured its pro-slavery orientation, drifted towards the Free Soilers. So did radical 'Conscience Whigs', who could not abide their own party's concessions to the Slave Power. In the run-up to the election of 1852, while Democrats were able temporarily to quell their internal conflict, the cleavage among Whigs widened. The radicals demanded political abolition; the conservative 'Cotton Whigs', who represented the party's manufacturing constituency, were loath to disrupt commerce (and the Northern mills' supply of southern cotton) by fuelling the sectional conflict. The rancour contributed to the Whigs' loss of the presidency. At that moment, the urgency for conservatives to answer the articulation of the free labour doctrine to political abolition was manifest.

The task was delicate: the Whig party cast itself as a friend of labour and opponent of slavery. The tariff, a central plank in its platform, had long been

promoted by Whig icons such as Henry Clay and Daniel Webster as a way of arming workers in 'an unequal contest with the pauper labor of Europe' (Eckes 1995, p. 24). Advocacy of labour did not require compromise of the interests of manufacturers: the party line maintained that in general, as in tariff legislation, the interests of labour and capital were in harmony (Foner 1970, pp. 20–1). Nor did opposition to slavery require concrete support for its abolition: inheritors of the mantle of Clay and Webster detested slavery but did not see fit to involve the federal government in the fundamental question of its existence, except in new states and territories where prior law did not apply. In order for the conservatives to save the Whig party without sacrificing their convictions, the free labour doctrine and the presumption of harmony of interests had to be retained, but both had to be interpreted to support the protective tariff and to oppose political abolition.

The task was taken up by 19th-century America's most original economic thinker, Henry Charles Carey (1793–1879), son of Mathew Carey. By 1852 the younger Carey was already an important framer of Whig doctrine. The title of his first major book, *The Harmony of Nature as Exhibited in the Laws which Regulate the Increase of Population and of the Means of Subsistence: and in the Identity of the Interests of the Sovereign and the Subject; the Landlord and the Tenant; the Capitalist and the Workman; the Planter and the Slave* (1836) suffices to describe its contents. His second, a two-volume *Principles of Political Economy* (1840), developed more fully his system: its centrepiece was a law of the progression of civilization that entailed a new refutation of the Ricardian theory of rent. His third, The Past, the Present, and the Future (1848), demonstrated that the same law prescribed tariff protection for the United States – an argument that he propounded relentlessly to the end of his life. What remained was to show what the law implied for slavery.

That Carey did in *The Slave Trade, Domestic and Foreign: Why it Exists and How it May be Extinguished* (1853). For the unacquainted reader he reiterated the law of progression. Contrary to Ricardo's teaching, Carey showed that the land that is cultivated first is not the most fertile, which is at the bottom of river valleys where the vegetation is dense. The first settlers to an area do not have the numbers or the technology to clear and till the bottom land: they settle instead at higher altitudes and cultivate less fertile land. As they eke out a living and their numbers grow, they need not confine their work to agriculture. Their employments diversify, and some produce manufactures. Manufacturers devise machines and techniques that are useful in agriculture, allowing the population to inhabit and cultivate more fertile land.

They move down the hillside; they produce more food; their numbers grow, and they diversify further their employments; their manufacturers invent better machines and techniques; they move further down the hillside and cultivate better land; and so on. Thus progress requires the development of technology that passes into agricultural use, whether by design or happenstance. Development of technology depends on the diversification of industry – or as Carey put it more vividly, 'the natural tendency of the loom and the anvil to seek to take their place by the side of the plough and the harrow' (Carey 1853, p. 50).

The natural tendency is interrupted at a country's peril. The United States inflicted upon itself such an interruption, according to Carey, in so far as it relaxed its impediments to foreign trade, impelling settlers to go West and farm more extensively for export rather than planting roots and farming more intensively for the home market. But that was not the only consequence of excessive foreign trade. Amongst a geographically dispersed population with few and small population centres, land tenure was characterized by larger parcels, which were more congenial to the cultivation of crops by slave labour, and correspondingly the demand for free labour was small (Carey 1853, p. 51).

The extinction of slavery, then, was not to be achieved by political abolition.

The sudden reversal of the relationship between master and slave would produce indolence in the latter and cause the ruin of both (Carey 1853, p. 23–4). It was to be achieved by abolishing the economic conditions that fostered slavery. The conditions were reducible to the small demand for free labour. To increase the

demand, it was necessary to ensure that land was cultivated intensively, in small plots, and for sale in a home market that also comprised manufacturers. To ensure that outcome, it was necessary to restrain foreign trade with a protective tariff. Only with the traditional programme of the Whig Party, not the radical one of the Free Soil Party, would both master and slave be set free: the master from his dependence on distant markets, the slave from his shackles.

Carey's system ultimately failed to be the cement to hold together the Whig Party, but not because it was unpersuasive under the circumstances when he wrote it. Circumstances changed. The Missouri Compromise of 1820 had prohibited the future admission of slave states above latitude 36°30′. Because the compromise rendered the slavery question irrelevant in most of the territory into which the United States would expand, moderate opponents of slavery were willing to leave the question up to state legislatures in the remaining areas. At least one professed opponent of slavery even argued, paradoxically, for its *expansion* into those areas and beyond. George Tucker (1775–1861), who introduced political economy to curriculum of the University of Virginia in 1825, had earlier served his district in the US House. There, in 1820, he borrowed Malthusian population theory to demonstrate that as the nation expanded westward the means of subsistence would increase, so the number of slaves could be expected to grow. If slaves were confined to the South while only whites migrated to the new areas, the stage would be set for a violent confrontation. The slaves, growing more populous relative to whites in the South if not elsewhere, would rise up against their masters. If slavery were allowed instead to spread westward, then the value of white labour would fall, the value of slave labour would follow, the price of slaves would fall below the cost of their maintenance, and the slaves would eventually be freed voluntarily (Dorfman 1966, vol. 2, p. 544). Carey's *The Slave Trade* was but a more comprehensive (and more plausible) justification for maintaining the same inclination manifested for so long by Tucker and many others: to oppose slavery while conciliating slaveholders. At last, just one year after the publication of Carey's tract, the position was clearly untenable. The Kansas–Nebraska Act, signed into law by President Franklin Pierce in 1854, allowed the slavery question to be put to a vote in the two newly organized territories, both of which were north of 36°30′, and to admit them to the Union as slave states or free according to the voters' wishes. In effect, the Act repealed the Missouri Compromise; and Tucker's decades-old notion that such a result would promote, not for ever postpone, slavery's demise could no longer be believed. Erstwhile conservative Whigs, with Carey in the front ranks, dissolved their party and joined the Conscience Whigs and Barnburner Democrats to organize the Republican Party.

Yet Carey's free-labour protectionism had become obsolete only in part. Former protectionist Whigs dominated the new Republican Party in financial matters, including tariff policy. After the outbreak of war, they instituted an irredeemable currency and a succession of tariff increases, mainly for war finance. Those who understood the free labour doctrine to have the opposite implications for money and tariffs did not oppose the new legislation: considering the stakes, they acknowledged its expediency. But this marriage of convenience between adherents to otherwise rival political–economic doctrines lasted only as long as the war.

From the Civil War to the American Economic Association

The slavery debate was settled with the war's end, but the money and tariff controversies erupted anew. Popular discontentment was aroused immediately by the myriad internal and external war taxes. The Republican Congress established a revenue commission to study the tax system and recommend an overhaul; its chairman (and after the first year, the sole commissioner) was David A. Wells (1828–1898), known to be a disciple of Henry Carey. From 1865 through 1867, in the first three of his five annual reports, Wells took care not to antagonize his mentor or his patrons. The dominant view in the Republican Party was that

tariffs, necessary for revenue in wartime, were opportune for protection in peacetime. Wells recommended accordingly that the internal excises should be dismantled, but the tariffs, which now yielded revenues of over 45 per cent of the value of imports, should be maintained.

Wells's report for 1868 marked a change of his thinking, and more. Favourable and unfavourable reviewers alike read it as an insider's repudiation of the American System. By attempting 'indiscriminate or universal protection', Wells determined, the protective tariff rendered 'all protection a nullity' because the iron and steel industries' output was the textile manufactures' input (Wells 1869, p. 3). The problem could not be solved simply by raising further the textile tariff: any modification of the tariff law would incite a general scramble for more protection. Given the political reality, Wells reasoned that it was better to determine tariffs by a simple and invariant principle than to attempt ad hoc changes. The principle was a tariff for revenue, not protection. Wells's stand for it, which only hardened in his reports for 1869 and 1870, widened the divide within the Republican coalition. Arthur Latham Perry (1830–1905) of Williams College, an evangelist for free trade in the mould of Leavitt, congratulated Wells for having written 'our Bible in onslaughts against the monopolists'; Henry Carey published a missive likening Wells to Judas Iscariot.

Although Wells arrived at the revenue tariff position by way of expediency, he occupied it thereafter as a sincere and doctrinaire exponent of free trade. There he joined William Cullen Bryant, who appended in 1866 the presidency of the American Free Trade League (AFTL) to his duties as newspaper editor and publisher; and Perry, who barnstormed for the AFTL after completing in the same year his *Elements of Political Economy,* the successor to Wayland's work as the principal American textbook on the subject. Bryant, a septuagenarian at the war's end, was of an older generation whose powers were waning; Wells and Perry, in their thirties, were more representative of the spirit of the post-bellum decade. Free trade was the cause that they championed most actively, but they were led to it by attitudes and methods of broader significance.

Both Wells and Perry were schooled in the clerical system of Wayland and were moved by the evangelical impulse of reform, but they were also young enough to partake of the fascination with science that characterized the American mind of the 1840s and 1850s. The fascination was stoked by the appearance of Darwin's *On the Origin of Species* in 1859, but, like the inquiry into biological adaptation to which Darwin contributed, was present much earlier. The famed Swiss geologist and zoologist Louis Agassiz arrived in the United States in 1846, drew thousands to his public lectures in Boston, and inspired textile magnate Abbott Lawrence to establish a scientific school at Harvard for him to lead. Wells was among the first four graduates of the Lawrence Scientific School. There he learned the talents of empirical observation and statistical argument that he displayed as Commissioner of the Revenue. Perry's scientific inclinations were manifested differently from Wells's but were consonant with them.

Perry himself served up few statistics, but he sought to purge political economy of metaphysical presumptions that were unanswerable by statistical measurement.

The goal could be achieved, he thought, by circumscribing the science, which had been stymied by preoccupation with 'wealth'. Because the word admitted so many meanings, none of them precise, whatever thing it named was hard to measure. Because the thing was hard to measure, the word was 'the bog whence most of the mists have arisen which have beclouded the whole subject' and was 'totally unfit for any scientific purpose whatever' (Perry 1866, p. 29). A better word than 'wealth' was 'value'. Value was no sooner determined than it was measurable; one had only to relinquish any hope of finding an invariant standard of it. The value of a thing was always relative to the other thing for which it was exchanged, and was determined only when the exchange was made. The amount of money exchanged depended on the exchanging parties' estimates of their desires and the efforts required for their satisfaction. Refashioned thus as the science of exchanges, political economy would concern itself only with things that are measurable in units of money. 'Catallactics', or perhaps

'economics', was the more accurate name for such a science, Perry allowed – but in this particular question of terminology he was less fastidious. Regardless of the science's name, once metaphysics was ostensibly exiled from it, one groped with difficulty within its scope for justification of government constraints of exchange. Therein lay Perry's enthusiasm for free trade, his unsurpassed appreciation of Frédéric Bastiat, and his more general presumption in opposition to commercial legislation, foreign or domestic.

While Wells's affinity for statistics and Perry's redefinition of political economy reflected the scientific aspirations of the post-bellum generation, the American Social Science Association (ASSA) reflected the union of those aspirations with the generation's reformist impulse. The impulse drew urgency from the tremendous economic transformation that the war had merely paused, not reversed. Fewer than 30 railroad miles were in operation in the United States in 1830; in 1860, the number exceeded 30,000. The railroad hastened migration from the eastern states to the West, but it also changes patterns of life and work within states. Urbanization, which encompassed only nine per cent of the population in 1830, swelled to 20 per cent in 1860 and continued upwards. From its founding in 1865, the ASSA directed itself to inquiry into the attendant problems: sanitary conditions, relief of the poor, prevention of crime, reform of criminal law, prison discipline, treatment of the insane, in addition to other matters in the domain of 'social science' (Haskell 1977, p. 98). Its organization, following that of the British association upon which it was modelled, consisted originally of four departments: Education, Public Health, Jurisprudence, and Economy, Trade, and Finance – the latter of which was concerned with issues ranging from the hours of work to prostitution and intemperance, public libraries, tariffs and other taxes, the national debt, regulation of markets, and the currency (pp. 104–5). Because participants generally shared the classical liberal assumptions of Wells and Perry – indeed, Wells was an early head of the Economy, Trade, and Finance Department and later president of the Association, and Perry was a regular contributor – their inquiries into these subjects tended not to yield proposals for ambitious programmes of government subsidization or regulation. Instead they were aimed (as the Association's constitution put it) at the diffusion of 'sound principles' and at bringing people together ' for the purpose of obtaining by discussion the real elements of Truth' (p. 101).

In matters of foreign trade, sound principles implied rejection of protectionism; in the currency, the resumption of specie payments. In the 'labor question', which Wells and Perry discussed at a round table in 1866 and again in 1867 (pp. 113–14), sound principles called perhaps for the collection of ample data regarding hours, wages, and conditions of work, but not their regulation on behalf of able-bodied and able-minded adults. Because the interests of workers and employers were in harmony, where poor work conditions existed they were the result either of a general lack of material progress or lack of knowledge by the actors. The solution to lack of material progress was to clear away any legislative impediments to it. The solution to lack of knowledge was for the ASSA to produce and disseminate it. In neither case was the solution to draw up redistributive 'class legislation' or other assaults on the prerogatives of property.

The panic of 1873, the depression in its wake, and the great railroad strike of 1877 undermined confidence in the solutions of the liberal reformers and made room for challengers. The labour question had metastasized into 'the social question'. The most original and widely read author to address it was Henry George (1839–1897), an autodidact working as a journalist in California when the publication of *Progress and Poverty* made him famous in 1879. George admired David Wells, and corresponded with him as early as 1871. He joined Wells in advocating free trade and denouncing 'monopoly' (Dorfman 1969, p. 142). But George was less deferential to property than was Wells, in whose circles he was viewed less of an ally than a danger.

George attributed the persistent poverty of labour amidst material progress to the unearned rewards of one particular kind of property. Land was the free gift of nature: although improvements on it were the work of man and the returns

to improvements were earned, the returns to land were not. Yet the returns to land grew inexorably as population and mechanical invention increased and cultivation was extended. The solution that George proposed was a single tax on land of 100 per cent of its annual value. Under such a system it would matter little if one held title to land, leased it to another, and paid a 100 per cent tax on the rent; or if instead the state held formal title to the land and leased it to (presumably) the same person who would rent it from a private owner. Private property in land could be formally retained, but in effect land would be appropriated by the state for the benefit of all.

At the same time, the right to all other property would be respected more scrupulously than before, because the single tax on land would obviate the need for all other taxes.

The political economists associated with liberal reform caught the scent of socialism from George's proposal, and they responded with alarm. Yet Henry George shared most of the values and even some of the legislative prescriptions with which the liberal reformers were most closely identified. He departed from them importantly only in his assumption of the illegitimacy of private ownership of one kind of possession. The challenges that would follow in the middle of the turbulent 1880s, from a new generation of economists including Edmund James (1855–1925), Simon Nelson Patten (1852–1922), and Richard T. Ely (1854–1943), were of another order entirely. The professional credentials of the young economists, which included doctoral studies in Germany and university positions in the United States, were different from Wells's and Perry's (let alone George's); and rather than sharing the values and prescriptions of their elders, they repudiated them thoroughly, defining themselves by opposition to economic liberalism. These economists were the founders in 1885 of the American Economic Association.

See Also

- ▶ American Economic Association
- ▶ Carey, Henry Charles (1793–1879)
- ▶ Catallactics
- ▶ Free Banking Era
- ▶ George, Henry (1839–1897)
- ▶ Land Tax
- ▶ List, Friedrich (1789–1846)
- ▶ Mill, John Stuart (1806–1873)
- ▶ Slavery
- ▶ Tariffs
- ▶ United States, Economics in (1885–1945)
- ▶ United States, Economics in (1945 to present)
- ▶ Wells, David Ames (1828–1898)

Bibliography

Bryant, W.C. 1808. *The embargo, or, sketches of the times: A satire*. Boston: E.G. House.

Cardozo, J.N. 1826. *Notes on political economy*. Charleston: A.E. Miller.

Carey, M. 1819. *Addresses of the Philadelphia society for the promotion of national industry*. Philadelphia: M. Carey and Son.

Carey, H.C. 1836. *The harmony of nature as exhibited in the laws which regulate the increase of population and of the means of subsistence: And in the identity of the interests of the sovereign and the subject; the landlord and the tenant; the capitalist and the workman; the llanter and the slave*. Philadelphia: Carey, Lea & Blanchard.

Carey, H.C. 1840. *Principles of political rconomy*. Philadelphia: Carey, Lea & Blanchard.

Carey, H.C. 1848. *The past, the present, and the future*. Philadelphia: Carey & Hart.

Carey, H.C. 1853. *The slave trade, domestic and foreign: Why it exists, and how it may be extinguished*. Philadelphia: A. Hart.

Cole, C.C. Jr. 1954. *The Social Ideas of the Northern Evangelists, 1826–1860*. New York: Columbia University Press.

de Vattel, E. 1758. *The law of nations, or the principles of natural law applied to the conduct and to the affairs of nations and of sovereigns*. Trans. Charles G. Fenwick. Vol. 3 of *The Classics of International Law: Le Droit des Gens*, ed. J.B. Scott. Washington, DC: Carnegie Institution, 1916.

Dorfman, J. 1966. *The Economic Mind in American Civilization, 1606–1865*. Vol. 2. New York: Augustus M. Kelley.

Dorfman, J. 1969. *The economic mind in American civilization, 1865–1918*. New York: Augustus M. Kelley.

Eckes, A.E. Jr. 1995. *Opening America's Market: U.S. foreign trade policy since 1776*. Chapel Hill: University of North Carolina Press.

Foner, E. 1970. *Free soil, free labor, free men: The ideology of the Republican Party before the Civil War*. New York: Oxford University Press.

Foner, E. 1976. *Tom Paine and revolutionary America*. New York: Oxford University Press.
Franklin, B. 1729. A modest enquiry into the nature and necessity of a paper currency. In *The writings of Benjamin Franklin*, ed. A.H. Smyth, Vol. 2. New York: Macmillan, 1907.
Franklin, B. 1769. Positions to be examined, concerning national wealth. In *The Writings of Benjamin Franklin*, ed. A.H. Smyth, Vol. 5. New York: Macmillan, 1907.
George, H. 1879. *Progress and poverty: An inquiry into the cause of industrial depressions, and of increase of want with increase of wealth – The remedy.* San Francisco: W. M. Hinton.
Hamilton, A. 1791. Report on manufactures. In *Encyclopedia of tariffs and trade in U.S. history*, ed. C.C. Northrup and E.C. Turney, Vol. 2. Westport: Greenwood Press, 2003.
Hamilton, A., J. Jay, and T. Jefferson. 1787–1788. *The Federalist.* Online. Available http://www.foundingfathers.info/federalistpapers. Accessed 3 Sept 2006.
Haskell, T.L. 1977. *The emergence of professional social science: The American Social Science Association and the nineteenth-century crisis of authority.* Baltimore: The Johns Hopkins University Press.
Jefferson, T. 1785. *Notes on the state of Virginia*, 1788. Philadelphia: Prichard and Hall.
Leiman, M.M. 1966. *Jacob N. Cardozo: Economic thought in the antebellum South*. New York: Columbia University Press.
McVickar, J. 1825. *Outlines of political economy.* New York: Wilder & Campbell.
O'Connor, M.J.L. 1944. *Origins of academic economics in the United States*, 1974. New York: Garland Publishing, Inc..
Paine, T. 1776. Common sense: Addressed to the inhabitants of America, etc. In *The complete writings of Thomas Paine*, ed. P.S. Foner, Vol. 1. New York: Citadel Press, 1945.
Paine, T. 1778. The American crisis: VII. In *The complete writings of Thomas Paine*, ed. P.S. Foner, Vol. 1. New York: Citadel Press, 1945.
Perry, A.L. 1866. *Elements of political economy.* New York: Charles Scribner's Sons.
Petty, Sir W. 1690. *Discourse on Political Arithmetic*. Reprinted in *Essays on Mankind and Political Arithmetic.* London*:* Cassell, 1888.
Raymond, D. 1820. *Thoughts on political economy.* Baltimore: F. Lucas.
Say, J.-B. 1821. *A treatise on political economy; or the production, distribution, and consumption of wealth.* Trans. from the 4th edn by C.R. Prinsep. Boston: Wells and Lilly.
Smith, A. 1789. In *An inquiry into the nature and causes of the wealth of nations*, 5th ed., ed. E. Cannan. New York: Modern Library, 1937.
Spiegel, H.W. 1960. *The rise of American economic thought*. Philadelphia: Chilton Company.
Wayland, F. 1835. *Elements of moral science*. New York: Cooke and Co..
Wayland, F. 1837a. *Elements of moral science*. 4th ed. Boston: Gould and Lincoln, 1859.
Wayland, F. 1837b. *Elements of political economy.* New York: Leavitt, Lord & Co..
Wayland, F. 1841. *Elements of political economy.* 4th ed. Boston: Gould and Lincoln, 1873.
Wells, D.A. 1869. *Report of the Special Commissioner of the Revenue for the Year 1868*. In House Ex. Doc. No. 16, 40th Cong., 3rd Session. Washington, DC: Government Printing Office.

United States, Economics in (1885–1945)

Bradley W. Bateman

Abstract

The history of American economics following the founding of the American Economic Association in 1885 is not a simple linear narrative of the triumph of neoclassical economics over historical economics. On the contrary, American economics remained a highly plural enterprise until the 1930s. Although there was strife in the 1880s over the proper method of doing economic research, this strife quickly gave way to a long period of détente. Only following the secularization of economics in the 1920s and the advent of the synthesis of neoclassical and Keynesian economics in the 1940s did this pluralism end.

Keywords

Adams, H. C.; American Economic Association; Anderson, B.; Arrow, K. J.; Bain, J.; Catchings, W.; Chamberlin, E.H.; Clark, J. B.; Clark, J. M.; Commons, J. R.; Cowles Commission; Cowles, A.; Currie, L.; Dantzig, G. B.; Douglas, P.; Dunbar, C.; Econometric Society; Elym, R. T.; Fetter, F.; Fisher, I.; Foster, W. T; German Historical School; Gibbs, W.; Gilbert, M.; Great Depression; Haavelmo, T.; Hadley, A. T.; Hamilton, W.; Hotelling, H.; Institutionalism; Knight, F. H.; Koopmans,

T. C.; Kuznets, S.; Laissez-faire; Laughlin, J. L.; Marginal School; Mason, E.; Mitchell, W. C.; National Bureau of Economic Research; Neoclassical economics; New Deal; Newcomb, S.; Patten, S. N.; Pollak Foundation for Economic Research (USA); Probabilistic revolution; Progressivism; Real-bills doctrine; Samuelson, P. A.; Seligman, E. R. A.; Snyder, C.; Sumner, W. G.; Sumner, W. G.; Taussig, F. W.; Underconsumptionism; United States, economics in; Veblen, T. I.; Walker, F. A.; Warburton, C.; Willis, H.P.; Young, A. A

JEL Classifications
B2

The 60-year period from the founding of the American Economic Association (AEA) in 1885 to 1945 marks the period when American economics was first professionalized and then came to take its characteristic, modern form. However, this story is not a simple linear narrative, as the events during this period were influenced by a series of historical contingencies and social forces that meant the outcome remained unpredictable until sometime during the last decade of the period.

Historians and economists alike have tended to believe that the story is somewhat more straightforward than it actually was. The historian Dorothy Ross (1991), for instance, has argued that the rise to prominence of John Bates Clark as America's premier (and first native) economic theorist at the turn to the 20th century marks the triumph of neoclassicism in American economics. More recently, Nancy Cohen (2002) has argued that neoclassicism came to dominate even earlier than Ross suggests, by as much as a decade. Both of these arguments are based, in part, on the well worn idea that American economics was characterized by a '*Methodenstreit*', or 'war of methods' in which the Marginalist School defeated the previously strong Historical School by the end of the 19th century. But while it is true that there was a crisis in American economics during the decade after the founding of the AEA, it had a somewhat different nature from that usually understood: it was much more about the purpose of economics than about method.

American Economics, Circa 1885

The first doctorate in political economy issued by an American university was in 1886 when Henry Carter Adams received his degree from Johns Hopkins University. Before that time, if an American wanted a doctorate in economics, it was necessary to travel to Europe to study. The vast majority who chose this route studied in Germany, to which 9,000 Americans travelled for this purpose between 1820 and 1920 (Herbst 1965). The most prominent of these young Americans went on to help found the AEA after their return, including Richard T. Ely, J.B. Clark and Simon Nelson Patten.

In Germany, these young economists found a profession that was characterized by scepticism towards Adam Smith's system of 'perfect liberty' and his arguments for laissez-faire. It now seems fair to say that the economic ideas that the Germans were most disturbed by were in fact a caricature of Smith: the 'Manchesterism' that they found so offensive was more the product of David Ricardo than Adam Smith, but their objection was deep and profound to what they took to be an argument unsuited for all nations at all times. They believed that Smith had made universalizing assumptions about human nature and human behaviour that were not universally correct. In the place of these assumptions, they intended to build a more empirically accurate economics based on careful historical study of how the individual behaviour and economic institutions of different societies had come about. Thus, rather than assuming that all people always weighed their options and made their choices so as to maximize their individual well-being, the leading German economists in the second half of the 19th century believed that it was necessary to study how cultural customs had evolved and how these shaped individual behaviour.

Known as the 'Historical School', the leading German economists at this time espoused the idea of '*Nationalökonomie*', or the study of how each

nation had come to its current position. These economists taught their students contemporary methods of historical study: careful archival work, the study of laws and customs, the collection of data, and the slow, inductive accumulation of knowledge. However, this did not preclude the study of marginal analysis. Karl Knies, at Heidelberg, who taught marginal techniques to the Austrians Friedrich von Wieser and Eugen Böhm-Bawerk, similarly taught the two most important Americans in Germany, John Bates Clark, the pioneer American marginalist, and Richard T. Ely, one of the leading representatives of American historical economics.

When Clark returned from Germany in 1875 and Ely in 1880 they found the American economics profession unprepared for much of what they had to offer. American economists were anxious about their own lack of theoretical sophistication. Writing on the occasion of the nation's centennial in 1876, Charles Dunbar, Professor of Economics at Harvard, had complained of the unoriginal and derivative character of American economics. Dunbar 'observed that American scholarship as yet contributed nothing to fundamental economic knowledge. In his reading, American economics to date had been derivative, stagnant, and sterile' (Barber 2003, p. 231). This may be have been an exaggeration, but it was the case that American economics texts at this time, mostly written by college presidents for undergraduate moral philosophy classes, lacked theoretical sophistication, and the young economists returning from Europe found an open field for the possibility of producing a new American economics.

Another reason for this lack of theoretical sophistication was that texts were written to propagate the virtues of free enterprise, hard work, and republican democracy, a troika of American virtues that posed a problem for the returning young economists. The crux of the problem was that the elders who controlled the newly created academic positions that the young economists hoped to enter expected anyone they hired to espouse the same beliefs regarding laissez-faire. Francis Amasa Walker, who taught at the Massachusetts Institute of Technology (MIT) and who served as the first president of the AEA, would reflect that laissez-faire, 'was not made the test of economic orthodoxy, merely. It was used to decide whether a man were an economist at all' (in Coats 1988, p. 362).

Unfortunately for men like Clark and Ely, this made employment in academe difficult, if not impossible, for on their immediate return from Germany they almost all adhered to Christian socialism. In part, this commitment reflected the enthusiasm they had acquired in Germany for social reform. But it also reflected an effort on their part to build an American '*Nationalökonomie*'. Trained to believe that each country had a unique culture and unique institutions, these young men latched onto evangelical Protestant Christianity, and tried to use it to fashion a native argument for economic reform.

Following the Civil War, after an era when the importance of hard work, free enterprise and the rights of capital had been universally preached (Bateman 2005). American Protestantism had begun to split into two 'parties' (cf. Marty 1986). The 'private party' focused on individual salvation and did not view the emerging economic conditions as a matter for Christian concern. The 'public party', on the other hand, focused on social reform while de-emphasizing the older evangelical concern with individual piety. In the new economic order of the post-bellum world these two strands of evangelical thought became emblematic of the two most common (and opposing) responses to the growth in industrial production, increasing urbanization, rapid immigration and growing inequality. Each group took the Bible as its primary text, but their purposes and understanding of the world could not have been more different. Whereas the private party Christians focused on individual piety, the public party wanted to build 'the Kingdom of God on this earth' through collective action and with the aid of the state.

The Progressive Era

At the centre of the public party of American Protestantism were the young economists who had returned from Germany. They had learned a new scientific ethic of empirical study and had made initial contact with marginal reasoning.

But balanced against this, was the fact, uncomfortable for their elders, that almost to a person these returning young economists were interested in changing the balance of power between working men and the owners of the new, vertically integrated enterprises that employed them. Thus, the young German-trained economists had some things that the profession craved, such as technical prowess, but they also manifested a political attitude that was unacceptable to much of the profession.

Laissez-faire in 19th-century America was not a technical argument for the efficiency of free markets, but rather a manifestation of a broader Protestant ethos. It meant a limited role for the state in the economy, but it also meant the unfettered right of capitalists to employ workers on the capitalists' terms; at the time of the Civil War, for instance, virtually throughout the United States it was still illegal to strike or to form a labour union. Thus, when the young, German-trained economists challenged laissez-faire, they were as likely to be arguing for the right of workers to strike as they were to be arguing for the municipal provision of water. In the last decades of the 19th century, then, laissez-faire meant both a limited role for the state and privilege for the prerogatives of capital. Men like Ely and Patten meant to challenge these ideas head-on and they took their warrant to do so as much from Christian scripture as they did from economic theory: they saw their work as economic theorists as an expression of their Christian commitments.

However, they generally met scorn from the older generation. Under the leadership of Ely, Patten, and Clark, the young advocates of labour and the possibility of beneficial state intervention formed the American Economic Association (AEA) in 1885, in part as an effort to provide support for one another and to seek an alternative form of professional recognition. But not all the AEA's members were economists; about a quarter of the members at the first meeting in Saratoga Springs, New York, were Protestant ministers. The young economists with doctorates could perform empirical research showing the extent of poverty, inequality and industrial dislocation in the economy; the ministers could preach the young economists' findings from their pulpits on Sunday morning to help motivate their parishioners to act to help build the Kingdom.

The formation of the AEA led to one of the most well-known exchanges between the older advocates of laissez-faire and the young economists who hoped to establish a role for the state in the functioning of the economy. William Graham Sumner at Yale University was the most high-profile advocate of laissez-faire at the time of the founding of the AEA, but one of Ely's colleagues at Johns Hopkins, Simon Newcomb, was also widely recognized. Newcomb was an acclaimed astronomer and mathematician who occasionally lectured on economics at Johns Hopkins and he wrote widely on economics in the popular press. Following the establishment of the AEA in 1885, *Science* magazine asked members of the old and new schools to debate their positions in a series of exchanges (reprinted in Adams, 1886). The exchange was often vitriolic, with Newcomb accusing Ely of being a socialist, no small charge in the immediate aftermath of the Haymarket riots. Often unnoticed in the debate, however, is the fact that both sides called on the authority of Alfred Marshall and William Stanley Jevons to establish their own points. More than anything in the debate, this last fact perhaps illustrates the extent to which the American economics profession at this time was animated by political differences over laissez-faire and the role of the state, rather than primarily by an argument over methods.

Much to the chagrin of Ely, the AEA did not long remain a haven for advocates of the state. With the pressure on the young economists to demonstrate they were not radicals who advocated violence, they felt it necessary to show themselves open to their older colleagues. And since the Association was also billing itself as the national organization for all economists, and because the young economists depended on the older generation for jobs and support, they hardly felt that they could deny the older economists membership in the new organization when they asked to join. Thus, by 1892, virtually all the older advocates of laissez-faire were members of the organization and the founding statement of

principles in favour of a role for the state and amelioration of economic dislocation, which Ely had helped draft in 1885 at the time of the founding, had been abandoned. Ely's response was to boycott the annual meeting that year.

At this point, it is undeniable there was great tension in the profession. It is important to realize, however, that, John Bates Clark, the premier marginalist, and Richard T. Ely, the premier historical economist, had worked together to found the AEA. It is true that Clark has dropped his interest in Christian socialism by 1885, but he had not turned completely against historical analysis. Nor, strictly speaking, was Ely an anti-marginalist; Ely had introduced marginal utility into his textbook writing as early as 1893. E.R.A. Seligman, another of the AEA founders, represents a good case study of the way that many economists at this time balanced marginalist and historical techniques in their work. The real conflict within the profession at this time was about the possible role of the state and whether the purpose of economics was to defend laissez-faire, rather than whether marginalism should be used in economic analysis (cf. Yonay 1998).

The diversity within the profession is also clearly manifest in the leading figures of the old school that emerged during the last two decades of the 19th century: Arthur Hadley, Frank Taussig, and J. Laurence Laughlin. These three men became the leaders respectively at Yale, Harvard and Chicago in the era when departments were emerging as the dominant institutions in the formation of the American economics profession. The work of these men represents three of the main economic problems facing the nation at the end of the 19th century (international trade and tariffs, railroads and monetary policy) and their theoretical eclecticism, while adhering to some form of the older cost-based theories of classical economics, demonstrates the evolving nature of the profession.

J. Laurence Laughlin was the most rigidly classical of these three 'young traditionalists' (Dorfman 1949). Despite writing a dissertation in history under Henry Adams, Laughlin gave no credence to the Historical School and was the last of the major figures in the old school to join the AEA (in 1904). Although not a prolific writer, his *History of Bimetallism in the United States* (1885) is a landmark study of the history of late 19th-century fights over currency and monetary policy. Ironically, however, after founding the *Journal of Political Economy* (1892), Laughlin opened the journal to economists of all stripes, and turned the book reviewing over to Thorstein Veblen for several years.

Frank Taussig is probably the most well remembered of the three, not least for his influence on the generation of students who enroled in his 'EC 11' graduate seminar at Harvard. Widely respected as a teacher, he was a gregarious person who was the first of the old school economists to join the AEA (in 1886). Although himself an adherent of classical economics, particularly of Ricardo and John Stuart Mill, Taussig was capacious in his analysis of economic problems and was often willing to see the legitimacy of other points of view. There is no better example of this than in his *Tariff History of the United States* (1888), an adaptation of his dissertation that went through many revisions in its subsequent editions; although he was a dedicated free trader, Taussig had a subtle understanding of the use of tariffs for revenue collection and appreciated, for instance, the arguments for sometimes protecting infant industries. Like Laughlin, Taussig was also an eclectic and important early journal editor. Charles Dunbar had founded the *Quarterly Journal of Economics* at Harvard (1886), but when Taussig became its editor, he opened his pages to economists of all views.

Arthur Hadley tackled one of the major economic issues in late 19th-century American capitalism, the railroads. His early reputation was based on his *Railroad Transportation* (1885), in which he argued that businesses with large fixed costs would not necessarily shut down when prices fell below the cost of production. Instead, he pointed out that as long as the firm could cover what are now known as variable costs and still make some contribution towards paying fixed costs such as interest payments, that they would stay in business. Hadley represents an intriguing figure for he acknowledged that his results on the role of fixed cost contradicted some of Ricardo's

arguments and he also accepted the basic validity of marginal utility reasoning. Thus, he seemed to have transcended classical economics in many regards. Yet his outlook was unmistakably laissez-faire, and he believed that there was no reason for further work in marginal utility theory, since its validity had been shown and that more work represented an unnecessary foray into psychology. In the end, his economics was driven largely by the study of costs, as in classical economics, and his conclusions did not stray far from the classic dogmas.

Members of the older generation were not alone in forming strong departments during the last decade of the 19th century. Johns Hopkins, which was founded in 1876 to establish higher education, especially graduate education, in the German style in the United States, was the first notable American graduate course in economics, producing the first Ph.D. granted in economics in the United States. The department's two notables were Simon Newcomb and Richard T. Ely. However, after his falling out with Newcomb, Ely left Johns Hopkins in 1892 for the University of Wisconsin, where he quickly assembled one of the top economics departments in the nation. Initially, Ely's move to Wisconsin was seen to be a possible precursor for regional factionalism in American economics. After Ely failed to attend the annual meeting of the AEA in 1892, there was concern among many Eastern economists that he would lead a movement from the Midwest to create a new professional organization that would promote his original challenge to laissez-faire.

However, following Ely's academic trial in Wisconsin in 1894 on charges of entertaining a union organizer in his home and of advocating socialism in his lectures, people from both camps began to look for common ground. Ely self-consciously chose to lower his profile after he was acquitted in his trial, and the advocates of laissez-faire began to realize that it was not good for the profession's credibility to have high-profile public disagreement. What followed at the turn of the century was the emergence of a kind of détente in American economics: leading figures in the profession continued to build strong departments around the country, but economists were granted a wide berth to examine social and economic conditions and to use the tools they saw best suited for the specific question at hand. Support for laissez-faire and government intervention were both accepted; what was expected was a rigorous approach to one's position.

One basis for this détente was undoubtedly the common Protestant background of most American economists at this time. Although Ely, Patten, John R. Commons, and Henry Carter Adams were asking American Protestants to turn away from their traditional position in favour of the prerogatives of the owners of capital, they were making the appeal on biblical grounds and were self-consciously appealing to the emerging public party of Protestantism. Since prominent members of the group of younger economists (for example, Ely and Adams) made it clear in their academic trials that they were not advocating the overthrow of the state, but rather the empirical study of the conditions of labourers, it became harder to demonize them for their impulse to seek what were clearly Christian ends. For their part, the older advocates on laissez-faire were willing to begin the difficult work of absorbing new theoretical techniques and accepted the call of the younger economists to undertake more empirical work and to let its results inform their understanding of the true effects of relatively untrammelled markets.

This détente between the younger and older economists, and between advocates of laissez-faire and advocates of more rights for labourers, created a fertile ground for American economics. During this period, there was not a single orthodoxy in American economics; one could work with marginalist ideas, historical ideas, or with both. New School economists such as Ely, Clark, and Seligman, as well as Old School economists such as Taussig and Hadley, all employed both techniques in their work. All that was required to be taken seriously in this world of plural methods was a dedication to the examination of the contemporary economic issues that were arising as America became an industrialized, urbanized nation. That the *American Economic Review*, which was founded in 1911, showed no marked tendency towards any school in its published

articles is strong evidence that there was no single dominant school at this time.

Perhaps the most notable dissent from this détente was the *sui generis* Thorstein Veblen. Like the young economists returning from Germany, Veblen was interested in a more 'scientific' economics. But he was never much interested in large-scale empirical work such as the social survey movement fostered by Ely and Commons at the turn of the century. Instead, Veblen wanted economics to be rebuilt as an evolutionary science based on Darwinian principles of natural selection (Hodgson 2004). Veblen's greatest theoretical advances would come during the first decade of the 20th century, but they were offered as part of a sardonic and biting criticism of Clark's work and so not only alienated Clark and his followers but also the bulk of the profession who saw their toolbox as containing many different techniques, one of which might sometimes be marginal reasoning. Veblen agreed that American economists should examine the newly emerging American capitalism, but he was not interested in pluralism of techniques.

Perhaps the most visible sign of the emerging détente among most American economists at this time was Ely's election to the presidency of the AEA in 1900. Not only had his feared defection been averted, but he had been successfully pulled back into the centre of the organization. Ely and Veblen never shared a close personal or professional relationship and Veblen would later harbour a bitter resentment that his brilliance and originality had been ignored by the AEA in the pivotal years when it would have helped his professional stature. Despite the fact that Ely and Veblen shared an interest in historical analysis, Veblen's style and his self-certainty in the Darwinian method kept him apart from the mainstream.

The year 1900 also marked the beginning of the Progressive Era, a profound shift in American society that would propel economists like Ely, Adams and Commons into the mainstream of American thought. With the emergence of progressivism during the first two decades of the 20th century, the cultural, political, and religious centre of American society would move away from 19th-century ideals of laissez-faire and towards an ethos of active civic engagement in trying to ameliorate the many social dislocations of the new industrialism, emerging urban poverty and concentrations of power in large corporations. The public party of Protestantism was the central force of progressivism in the first decade of the new century and economists like Ely and Adams, who had been subjected to political pressure to mitigate their views in the late 19th century, now found themselves in great demand and at the head of progressive projects such as regulatory commissions and social survey projects (Furner 1975; Bateman 2001). Just as the laissez-faire economists in the 19th century had benefited from the dominant Protestant ethos, so too would the young, so-called 'ethical economists' benefit from the rise of public party Protestantism early in the 20th century. This public recognition, and the university positions that often came with it, undoubtedly made it easier for the ethical economists to ignore Veblen's criticisms. This is not to say that the economics profession as a whole was receiving the attention it believed it deserved. On the contrary, for a significant part of the 20th century, the AEA was actively concerned with measures that would secure it an appropriate public profile (see Bernstein 2001).

The First World War and the End of the Idyll

At the same time that the young economists who had founded the AEA were rising to popular prominence, another generation of marginalist economists was also beginning to emerge. One of the most distinguished of the new generation of marginalists was trained by Ely. Allyn Young received his doctorate working under Ely at Wisconsin and became the first of many to become revisers and co-authors of Ely's best-selling *Principles* text. Some of the new generation of marginalists, such as Frank Fetter, examined the psychological dimensions of utility and sought to more clearly articulate the welfare implications of marginal utility reasoning. Perhaps the most innovative marginalist thinker to emerge after Clark

was Irving Fisher. Fisher's mentors at Yale were William Graham Sumner and Willard Gibbs. Gibbs was one of the most prominent mathematicians of his generation and he influenced Fisher to develop marginal economics in a more technical, mathematical form that would later come to characterize American neoclassical economics.

While Fisher's work was recognized and lauded, it was not, however, representative of the mainstream in the first two decades of the 20th century. Progressivism defined the centre of American political and intellectual life from roughly the turn of the century to the end of the First World War; but during the second decade of the century it shed some of its public rhetoric of *Protestant* reform as it began to pull in Jewish writers like Walter Lippman and Herbert Croly. Progressivism also became more focused on industrial efficiency as many embraced Frederick Taylor's time and motion studies as a means to achieve greater productivity and raise the standard of living. But the moralism of Protestant reform still suffused much of progressive thinking and would lead to the movement's quick demise after the war.

The problem for progressivism after 1918 was that one of its greatest proponents, Woodrow Wilson, had led the nation into war using the rhetoric of reform and democracy. He had been supported by the Protestant clergy, many of whom had used their pulpits to preach the justness and necessity of the war when America had entered the conflict in 1917. Thus, when the war was over, and the atrocity of the trench fighting was driven home to people, there was a quick and sudden turn against progressivism and especially against the moralism and rhetoric of moral improvement that had underpinned it (Danbom 1987). The hope that had supported the progressive movement was now widely seen to have been based on an unrealistic understanding of human nature.

This shift away from progressive moralism had a quick and deep effect upon all the American social sciences that had been pioneered by public party Protestants; sociology, political science, and economics all made a sudden turn to a more 'scientific' and less 'moralistic' basis. Dorothy Ross (1991) has called this 'the advent of scientism'. Of course, all three social sciences had considered themselves scientific before the war (Furner 1975), and all had been interested in empirical survey work of urban and rural populations; but they had all operated on an implicit belief that if this survey work revealed social pathologies such as poverty and child labour, that good Christian women and men would surely act to alleviate them when faced with the evidence. After the war, such a reliance on an idea that people were well motivated and altruistic was abandoned. So, too, was the idea that people who lived in squalid social conditions would experience moral improvement if the conditions were changed.

Thus, in the years immediately before 1920, progressive social science quickly unravelled. Realizing that they had lost the sympathy of the larger populace, and faced with the need for serious soul-searching on their own part, all three social sciences made an explicit effort to eschew the optimistic, moralistic rhetoric of progressivism and embraced a new kind of 'scientific' endeavour. Entwined in their decisions to embrace a more value-neutral approach to social enquiry was a clear understanding that both public and private funding hinged on catching the tone of the times.

The name of this movement in economics was 'institutionalism' (see institutionalism, old; Rutherford 2000; Bateman 2004). The term was coined in 1918, at exactly the moment when the break from liberal Protestantism was happening in all three of these social sciences. The men who formed the nucleus of this emerging group, men like Walton Hamilton and Wesley Mitchell, self-consciously endeavoured to set up an empirical, data-generated research project that would be appealing to funding agencies such as the Carnegie and Rockefeller Foundations. The founding of the National Bureau of Economic Research (NBER), with Mitchell as the director, was one of the signal achievements of the early institutionalists.

One effect of the effort towards a more empirical basis for economic research was that subjects that had been at the centre of American economics for at least four decades, such as poverty and philanthropy, were dropped from the research agenda of almost all institutionalists. Instead,

intense attention was paid to the cost structure of American industry, the business cycle, and the working of the financial system; these seemed to be the proper objects of serious economic scientists. Ultimately, the object of the institutionalists was to find a more scientific basis for 'social control', thus eradicating the need for the moralism of the progressives.

The most notable break with the past, however, was that for the first time since the founding of the AEA, there was now a *group* of American economists who were attempting to establish an institutional and historical approach to economics and who did not want a détente with marginal analysis. The *Methodenstreit* that many people now project back on to the late 19th century was actually beginning to emerge. The institutionalists were interested in developing a behaviouralist basis for individual behaviour and they eschewed the idea that marginal decision-making was the driving force behind most economic activity.

The institutionalists were also the first *group* of American economists to work to establish an explicitly secular economics. This reflected their desire to distance themselves from the moralizing rhetoric of the Christian economists who had founded the AEA and drew from the work of one of the institutionalists' main influences in pragmatic philosophy, John Dewey. While not every individual American economist had been a Protestant before the First World War, the ethos of Protestantism had suffused and stabilized the détente that held for most of the three decades after Ely had been exonerated in his academic trial. Liberal Protestant economists had believed that a kind of moral Darwinism was at work in American society and their common purpose in exploring social questions to determine where and when state intervention might (or might not) be appropriate had been made possible by their shared ethos. Now, however, the men who stepped forward to found institutionalism were making an explicit argument that science alone should inform their work; they staked their future on the idea that they could do better, more empirical science than the a priori marginalists who they believed depended on untested and incorrect assumptions about human behaviour.

The marginalists were fully up to the fight, however, and engaged the institutionalists after 1930 in an increasingly pointed dispute. In the first decade of institutionalism's rise, the détente held reasonably well. And during the 1920s, institutionalism was at least as well represented in the top graduate schools as marginalism. While marginalist thinkers such as Irving Fisher went about their work, the instituionalists controlled two of the top four graduate courses (Columbia and Wisconsin) and produced a large plurality of American Ph.D.s (Bowen 1953; Backhouse 1998; Biddle 1998).

The Great Depression and the New Deal

Contrary to the idea that American economics in the 1920s and 1930s was characterized by a final struggle between some latter-day historicists (that is, institutionalists) against the dominant neoclassicists, the pluralism of this period was much richer and represented a wide range of possibilities regarding the future direction of the profession. It might more correctly be said that no school of thought in American economics was completely dominant at this time. A nice example of this diversity would be to consider Harvard in the 1930s. Frank Taussig was still on the faculty and remained one of the leading authorities on international trade. Joseph Schumpeter would join the department in 1932, bringing a Continental influence, if not exactly an Austrian one. Edward Chamberlin would finish his work on imperfect competition at Harvard during the 1920s, under the twin influences of Marshall and Allyn Young. Young, who died in 1929, had been one of America's top theorists, but as noted above, he was a student of Ely's who had later tacked hard to the marginalist tradition. One could not define this department as simply neoclassical, but neither was it simply institutionalist. The full secularization of American economics, however, had prepared the ground for a more strident return to laissez-faire arguments, much like the ones that had existed before 1885.

For by the 1930s, leading marginalist thinkers such as Frank Knight were prepared to engage in

what Yuval Yonay (1998) has termed 'the fight for the soul of economics'. Despite their control of many of the top graduate courses (Backhouse 1998), and their initial success at fundraising, the institutionalists were hit hard by the kind of criticism that Knight levelled against Sumner Slichter in his 1932 review of Slichter's new institutionalist textbook. Knight did not like Slichter's methods of analysis, but his real bête noire was Slichter's focus on intervention to improve the performance of the economy. By this time, the institutionalists had clearly staked out their position of advocacy for 'controlling' the economy, and the advocates of marginalism, such as Knight and Jacob Viner, strongly disagreed. However, though some took clear sides, there were others who straddled both camps: John Maurice Clark (J.B. Clark's son) drew on many neoclassical tools, such as externalities, but his work on the control of business went further in directions favoured by institutionalists, turned away from the cause in the 1930s, arguing that many of the instituionalists' concerns were best handled by treating them as externalities within the neoclassical model; before his untimely death in 1929, Allyn Young had done his pioneering work in the economies of scale and supervised Chamberlin's doctoral thesis. The divide between institutionalism and neoclassicism was thus still blurred in this period.

Equally destructive of the myth that American economics was essentially a linear narrative of the development of neoclassical economics after 1890 is the rich American tradition of work on money, banking and the business cycle during the inter-war years (Laidler 1999). The work in these areas drew from the pre-First World War work of major figures such as Laughlin, Fisher, and Mitchell, and represented a wide range of theoretical development, as well as a full range of policy options.

Perhaps the most well remembered work from this period is Irving Fisher's work on the quantity theory and the relationships between money growth, the price level, and economic activity. Fisher (1911, 1923, 1925) was not alone in positing a close relationship between money and prices. Carl Snyder (1924) of the Federal Reserve Bank of New York drew on Fisher's work to suggest that the close relationship between money and prices supported a 'money growth guideline' (Laidler 1999).

There was no orthodoxy in monetary and cycle theory, however. For instance, Irving Fisher (1925) came to believe that there was no business cycle, simply fluctuations around the mean values of prices. This outlook contrasted sharply with Mitchell's careful empirical search for the factors that underlie what he believed were the regular oscillations of the economy. Both Mitchell and Alvin Hansen worked in the 1920s to develop versions of the accelerator principle, a concept previously developed by J.M. Clark in 1917, trying to uncover the ways that a growing economy could pick up momentum and how that same pattern of growth could ultimately lead to a downturn.

Likewise, just as some economists did not agree with Fisher's conclusions about the nature of the cycle, there were many who did not agree with his conclusions regarding the relationship between money and prices. In the inter-war years, there developed what David Laidler has termed an American version of the British Banking School. H. Parker Willis, who worked for the Federal Reserve in the second decade of the 20th century and later taught at Columbia's School of Business, and Benjamin Anderson, who had a position at Harvard before becoming the chief economist at Chase Bank, both argued against Fisher's use of the quantity theory. Willis had studied under Laughlin at Chicago and he followed in Laughlin footsteps in denying the necessary connection between money and prices.

In this rich mix of work regarding money and the business cycle, there was, not surprisingly, widespread disagreement about the possibilities for stabilization policy. Economists as diverse as Fisher, Mitchell, and Allyn Young supported different kinds of stabilization policy. Others, like Willis, adhered to a version of the real-bills doctrine, arguing that stability would come only through a prudent effort on the part of the Federal Reserve to limit its lending to those with high-quality, short-term commercial paper. Although

they did not have academic appointments, William Trufant Foster and Waddill Catchings (1923, 1925, 1928) used the Pollak Foundation for Economic Research as an effective platform to publicize a version of underconsumptionist theory. They argued that there was a need for monetary and fiscal expansion to sustain consumption and avoid recession. Paul Douglas (1927) at Chicago, who became a US senator from Illinois, made these ideas popular with his call for large-scale public works.

While there were, thus, many forms of arguments for fiscal and monetary efforts to sustain prosperity, it might seem that the institutionalists' predisposition for controlling the economy would have been popular after 1929, but neither the popular nor the professional tide turned toward the institutionalists after 1929. In retrospect, it was, perhaps, their unique bad luck to have played a role in developing parts of Franklin Delano Roosevelt's initial response to the Great Depression, his first New Deal. Many institutionalists joined Rexford Tugwell from Columbia University in Roosevelt's first term of administration, but they failed to provide a recognizably successful policy for combating the depression. Although the marginalist economists were not offering a popular plan for recovery, the institutionalists' efforts in the New Deal did not provide them with a set of successes upon which to build their legacy (Barber 1996).

It was also the bad luck of the institutionalists that by the end of the 1930s, when John Maynard Keynes's *General Theory of Employment, Money, and Interest* (1936) came to be widely seen as providing a theoretical underpinning for recovery, that the emerging neoclassical theorists like Paul Samuelson were already working to give his theory a neoclassical underpinning. The efforts to cast the *General Theory* in a general equilibrium model effected a remarkable and unexpected transformation in the future prospects for American economics, overshadowing the fact that the foundations of what came to be thought of as Keynesian fiscal policy were laid by institutionalists (Barber 1996; Rutherford and DesRoches 2008). Before the Great Depression, economists of all stripes had argued about possible monetary and fiscal interventions (see Barber 1985; Laidler 1999). Although the pure idea of 'social control' was an institutionalist construction, the potential for using fiscal and monetary policy came from almost every corner of the profession. Keynes's work provided a common analytical framework for examining such macroeconomic interventions, and Samuelson's work linked that analytical framework to marginalist, neoclassical ideas. Likewise, Alvin Hansen's embrace of Keynes's theoretical framework after his initial resistance lent important impetus to an alternative form for analysis of the business cycle, even though he turned to Keynesian economics only when he saw that it could be used to defend ideas he already held. Hansen built his graduate seminar on fiscal policy at Harvard around his reinterpretation of Keynes's *General Theory*, and in the 1940s and early 1950s wrote a series of texts that embodied much of the institutionalist concern with the business cycle and economic stabilization (see Mehrling 1997, chs 7–8). The project that institutionalists had undertaken in the 1920s to provide a new psychological basis for economic behaviour had never led to any substantive advances, which helped to make the mathematical elegance of Samuelson's work all the more attractive. It did not hurt that the tools of neoclassicism, in their Keynesian guise, were now seen to be as amenable to intervention as they were to laissez-faire. Edward Chamberlin's (1933) work analysing imperfect competition in the 1930s also started to lend a new sense of realism and possibility to the emerging neoclassical framework. In the 1930s, much empirical work had been undertaken by institutionalists, but it was Edward Mason and Joe Bain, who, drawing on Chamberlin's theory, developed the framework that was to dominate empirical work on industrial economics in the 1950s. Harold Hotelling's theoretical breakthroughs in formulating mathematical models of resource depletion in the 1930s added even more lustre to neoclassicism. A new, sharp sense of what it meant to be an economic theorist was emerging, and institutionalism did not appear to have a ready response.

The Econometric Movement and the Second World War

The event that symbolized this confidence in more formal theorizing than the institutionalists had generally favoured was the foundation of the Econometric Society in 1930. It was an international society, but Fisher, Schumpeter and other American economists were influential in its formation. Its importance lay in providing a focus for mathematical theories, covering both the cycle and microeconomics, and statistical research. An influential figure was Alfred Cowles, who not only supported the Econometric Society and its new journal, but also established the Cowles Commission. This was an economic research organization, set up in 1932, the heyday of which was from 1939 to 1955 when it was based in Chicago. Located outside the economics department, it provided a focus for the development and propagation of neoclassical theory (with particular emphasis on Walrasian general equilibrium theory) and statistical methods for testing and applying the theory. It was here, in the early 1940s, that Tjalling Koopmans and Trygve Haavelmo developed the methods that led to what has been called the probabilistic revolution in econometrics (Morgan 1990).

Two further developments were important in the transformation of American economics that took place in this period. One was the influx of a large number of émigrés from Europe. The United States had always been a home for such people, and in the 1920s many economists arrived from Russia and Eastern Europe, but this increased dramatically in the 1930s and 1940s. By the mid-1940s, almost half the authors of articles in the *American Economic Review* had been born outside the United States, most of these being affiliated to American universities (Backhouse 1998).

The other influence was the Second World War itself. This was, like no previous war, an economists' war (Bernstein 2001). Economists were recruited en masse into government. Many were employed to tackle what were clearly economic problems relating to domestic economic activity or to the estimation of enemy economic capacity. The most notable outcome of such work was national income analysis. Official estimates of US national income had first been calculated in 1933, in response to the onset of the depression, when Simon Kuznets was seconded to the Bureau of Foreign and Domestic Commerce from the NBER, where he had been working on the problem (elsewhere Clark Warburton, at the Brookings Institution, and Laughlin Currie had been working on similar lines). Under Robert Nathan this work was developed, and monthly figures were produced by 1938. But it was only during the war that these estimates were developed, under Martin Gilbert, into a system of accounts. One reason for this was that national income proved indispensable to the war effort, its main achievement being to calculate what Roosevelt could promise in his Victory Program.

However, the significance of the war went further than this, for economists also became involved in activities not traditionally associated with their subject. Operations Research, initiated in Britain in the 1930s, was taken up by the American armed forces, through the Office of Strategic Services (a forerunner of the Central Intelligence Agency) where economists were employed alongside mathematicians, statisticians and physicists to solve problems related to military strategy and tactics. Out of this arose techniques that later proved influential, such as linear programming, with which members of the Cowles Commission (Koopmans and George Dantzig) were heavily involved. Economists achieved a high reputation as general problem-solvers. Most important, however, was the effect on the way economics was conceived. Much of this work was focused on optimization and was highly technical. Economics came to be seen as akin to engineering. In the 1920s and for much of the 1930s it had been institutionalism that was associated with quantitative work; statistical work related to neoclassical theory did exist (for example work on measuring demand functions) but there was no parallel with the work being done at the NBER and by the institutionalists. In contrast, by the 1940s, there was in place a serious research programme, with techniques that were perceived to rival those of the hard sciences, in which theory and data interacted in a way that was different

from that found in inter-war institutionalism. The transition was a very slow process: for instance, when Kenneth Arrow entered Columbia as an undergraduate in the 1940s, he was still not taught modern price theory (Colander et al. 2004, ch. 10), but rather the institutional economics of the 1920s and 1930s. The scene was set for the disputes that were, in the late 1940s and early 1950s, to determine the way economics was to evolve after the Second World War.

Conclusions

At the beginning of the 20th century, America's economic heritage was still tied up with the cultural influence of Protestantism. By the 1930s, that legacy had disappeared. The significance of institutionalism was not that it continued the earlier historical economics acquired by Ely, J.B. Clark and their contemporaries in Germany but that it was an organized movement for a purely secular and scientific economics (Rutherford 1999). Some 50 years after the founding of the AEA, the profession was still split deeply on the proper role of the state in the economy, but everyone in the discussion now believed that the role of the state was a scientific question and was actively engaged in the development of the tools to answer the question. Neoclassical methods, the forerunners of those that dominated the profession in the post-war era, were being developed, but there was an immense variety, which the labels neoclassical and institutionalist fail to capture. It was a period of genuine pluralism in economics (Morgan and Rutherford 1998).

In retrospect, it is possible to discern the advance of neoclassical and more technical economics on a broad front. Young, Chamberlin, Hotelling, Samuelson and others were establishing a theoretical framework that could animate both microeconomic and macroeconomic work. However, rather than see this process as inevitable it is important to see the importance of external factors in determining the outcome of inter-war pluralism. The Great Depression exerted a profound effect; institutionalist planning was tainted by the failure of Roosevelt's first New Deal, in which planners such as Rexford Tugwell were heavily involved. The Second World War was perhaps even more important in helping change perceptions of what economics was and ideas about the position of economists in society. The rise of the Nazi party and their policies in Europe not only removed a major rival to the supremacy of Anglo-American economics, but caused an influx of economists into the United States who proved highly influential.

In addition, it is important not to exaggerate the extent of any neoclassical victory. Institutionalists remained strong in many fields. The NBER's work in establishing a statistical basis on which empirical analysis could be based was unrivalled. Furthermore, even where there would appear to be evidence for the conversion of institutionalists to other approaches, significant elements of institutionalism remained, as in J.M. Clark's work on the control of business or Hansen's use of Keynesian methods for analysing the business cycle. The legacy of institutionalism was widespread. There was a swing away from institutionalism, which can be documented in many ways (see, for example, Backhouse 1998; Biddle 1998) but the story was not linear, and in the 1930s and 1940s it was highly dependent on external events.

See Also

- ▶ American Economic Association
- ▶ Clark, John Bates (1847–1938)
- ▶ Ely, Richard Theodore (1854–1943)
- ▶ Institutionalism, Old
- ▶ Neoclassical Synthesis
- ▶ Young, Allyn Abbott (1876–1929)

U

Bibliography

Backhouse, R.E. 1998. The transformation of US economics 1920–1960: Viewed through a survey of journal articles. *History of Political Economy* 30(Suppl.): 85–107. In *Interwar pluralism to postwar neoclassicism*, ed. M.S. Morgan, and M. Rutherford. Durham: Duke University Press.

Barber, W.J. 1985. *From new era to new deal: Herbert Hoover, the economists, and American economic*

policy, 1921–1933. Cambridge: Cambridge University Press.

Barber, W.J. 1988. *Breaking the academic mould: Economists and American higher learning in the nineteenth century*. Middletown: Wesleyan University Press.

Barber, W.J. 1996. *Designs within disorder: Franklin D. Roosevelt, the economists, and the shaping of American economic policy, 1933–1945*. Cambridge: Cambridge University Press.

Barber, W.J. 2003. American economics to 1900. In *A companion to the history of economic thought*, ed. W.J. Samuels, J.E. Biddle, and J.B. Davis. Oxford: Blackwell.

Bateman, B.W. 1998. Clearing the ground: The demise of the social gospel movement and the rise of neoclassicism in American economics. *History of Political Economy* 30(Suppl.): 29–52. In *Interwar pluralism to postwar neoclassicism*, ed. M.S. Morgan, and M. Rutherford. Durham: Duke University Press.

Bateman, B.W. 2001. Make a righteous number: Social surveys, the men and religion forward movement, and quantification in American economics. *History of Political Economy* 33(Suppl.): 57–85. In *The age of economic measurement*, ed. J.L. Klein, and M.S. Morgan. Durham: Duke University Press.

Bateman, B.W. 2004. Why institutional economics matters as a category of historical analysis. *Research in the History of Economic Thought and Methodology* 22A: 219–256.

Bateman, B.W. 2005. Bringing in the state?: The life and times of laissez-faire in the 19th century United States. *History of Political Economy* 37(Suppl.). In *The role of government in the history of political economy*, ed. Steven Medema and Peter Boettke. Durham: Duke University Press.

Bateman, B.W., and E. Kapstein. 1999. Retrospectives: Between god and the market: The religious roots of the American economic association. *Journal of Economic Perspectives* 13(4): 249–258.

Bernstein, M. 2001. *A perilous progress: Economists and public purpose in twentieth century America*. Princeton: Princeton University Press.

Biddle, J. 1998. Institutional economics: A case of reproductive failure? *History of Political Economy* 30(Suppl.): 108–133. In *Interwar pluralism to postwar neoclassicism*, ed. M.S. Morgan, and M. Rutherford. Durham: Duke University Press.

Bowen, H.R. 1953. Graduate education in economics. *American Economic Review* 43: 1–223.

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

Coats, A.W. 1988. The educational revolution and the professionalization of American economics. In *Breaking the academic mould: Economists and American higher learning in the nineteenth century*, ed. William J. Barber. Middletown: Wesleyan University Press.

Coats, A.W. 1992a. *British and American economic essays*, *On the history of economic thought*. Vol. 1. London: Routledge.

Coats, A.W. 1992b. *British and American economic essays*, *The sociology and professionalization of economics*. Vol. 2. London: Routledge.

Cohen, N. 2002. *The reconstruction of American liberalism, 1865–1914*. Chapel Hill: University of North Carolina Press.

Colander, D., R. Holt, and J.B. Rosser, ed. 2004. *The changing face of economics: Conversations with cutting edge economists*. Ann Arbor: University of Michigan Press.

Danbom, D. 1987. *The world of hope: Progressives and the struggle for an ethical public life*. Philadelphia: Temple University Press.

Dorfman, J. 1949. *The economic mind in American civilization*. Vol. 3, 1865–1918. New York: Viking Press.

Douglas, P.H. 1927. The modern technique of mass production and its relation to wages. *Proceedings of the Academy of Political Science* 12: 17–42.

Fisher, I. 1911. *The purchasing power of money*. New York: Macmillan.

Fisher, I. 1923. The business cycle largely 'a dance of the dollar'. *Journal of the American Statistical Association* 18: 1024–1028.

Fisher, I. 1925. Our unstable dollar and the so-called business cycle. *Journal of the American Statistical Association* 20: 179–202.

Foster, W.T., and W. Catchings. 1923. *Money*. Boston: Houghton Mifflin.

Foster, W.T., and W. Catchings. 1925. *Profits*. Boston: Houghton Mifflin.

Foster, W.T., and W. Catchings. 1928. *The road to plenty*. Boston: Houghton Mifflin.

Furner, M.O. 1975. *Advocacy and objectivity: A crisis in the professionalization of American social sciences, 1865–1905*. Lexington: University of Kentucky Press.

Hadley, A. 1885. *Railroad transportation*. New York/London: G.P. Putnam.

Herbst, J. 1965. *The German historical school in American scholarship: A study in the transfer of scholarship*. Ithaca: Cornell University Press.

Hodgson, G. 2004. *The evolution of institutional economics: Agency, structure and darwinism in American institutionalism*. London/New York: Routledge.

Laidler, D.W. 1999. *Fabricating the Keynesian revolution: Studies of the inter-war literature on money, the cycle, and unemployment*. Cambridge: Cambridge University Press.

Laughlin, J.L. 1885. *The history of bimetallism in the United States*. 4th ed, 1898. New York: Appleton.

Marty, M. 1986. *Protestantism in the United States: Righteous empire*. New York: Scribners.

Mehrling, P. 1997. *The money interest and the public interest: American monetary thought, 1920–1970*. Cambridge, MA: Harvard University Press.

Morgan, M.S. 1990. *A history of econometric analysis*. Cambridge: Cambridge University Press.

Morgan, M.S., and M. Rutherford. 1998. *Interwar pluralism to postwar neoclassicism*. Durham: Duke University Press.

Ross, D. 1991. *The origins of American social science*. Cambridge: Cambridge University Press.
Rutherford, M. 1999. Institutionalism as 'scientific' economics. In *From classical economics to the theory of the firm: Essays in honour of D.P. O'Brien*, ed. R.E. Backhouse and J. Creedy. Cheltenham: Edward Elgar.
Rutherford, M. 2000. Understanding institutional economics: 1918–1929. *Journal of the History of Economic Thought* 22: 277–308.
Rutherford, M., and C.T. DesRoches. 2008. The institutionalist reaction to Keynesian economics. *Journal of the History of Economic Thought*.
Snyder, C. 1924. New measures in the equation of exchange. *American Economic Review* 14: 699–713.
Taussig, F. 1888. *Tariff history of the United States*. New York/London: G.P. Putnam & Sons.
Yonay, Y. 1998. *The struggle for the soul of economics: Institutionalist and neoclassical economists in America between the wars*. Princeton: Princeton University Press.

United States, Economics in (1945 to Present)

Roger E. Backhouse

Abstract

After 1945, American economics was transformed as radically as in the previous half century. Economists' involvement in the war effort compounded changes that originated in the 1930s to produce profound effects on the profession, and many of these were continued through institutions that developed during the Cold War. This article traces the way the institutions of the profession interacted with the content of economics to produce the technical economics centred on a core of economic theory and econometric methods that dominate it today. Attention is also drawn to the broader role of American profession in economics outside the United States.

Keywords

Allied Social Science Association; American Economic Association; American Enterprise Institute; American Finance Association; Austrian economics; Axiomatics; Banking School; Behavioural economics; Bounded rationality; Brookings Institution; Chicago School; Complexity theory; Council of Economic Advisers (USA); Cowles Commission; Currency School; Development economics; Econometric Society; Econometrics; Evolutionary game theory; Experimental economics; Federal Reserve System; Formalism; Foundation for Economic Education; Friedman, M.; Game theory; General equilibrium; Harvard University; Heritage Foundation; Heterodox economics; History of economic thought; Industrialism; Institutional economics; International Monetary Fund; IS–LM model; Keynesian revolution; Keynesianism; Koopmans, T. C.; Liberty Fund; Macroeconometrics; Macroeconomics, origins and history of; Marginalist controversy; Massachusetts Institute of Technology; Mathematical economics; Mathematics and economics; Microeconomics; Microfoundations; Models; Monopolistic competition; Mont Pèlerin Society; Nash, J.; National accounting; Old institutionalism; Operations research; Perfect competition; Positive economics; Post Keynesian economics; Preference reversals; Probability distributions; Public choice; Radical economics; RAND Corporation; Rational choice; Rockefeller Foundation; Schultz, H.; Simultaneous equations; Statistics and economics; Stiglitz, J.; Systems analysis; Union of Radical Political Economy; United States, economics in; Vining, R.; Von Neumann, J.; World Bank

JEL Classifications

B1

The Effect of the Second World War

The Second World War is more than a conventional dividing line, for it profoundly affected the course of economics in the United States. The interwar period had been one of pluralism within

economics: there was a variety of competing approaches to the subject, none of which was dominant. It was also comparatively easy to discern distinctively 'American' trends in economics, which could be related either to the intellectual environment (notably the pragmatism of C.S. Peirce, William James and John Dewey) or to economic circumstances (such as the recent establishment of the Federal Reserve System). Within a decade of the Second World War, if not earlier, this had changed dramatically. Economics was becoming more technical, the foundations of an orthodoxy were being laid, and the position of the United States in relation to other countries was changing. The conventional explanation is the Keynesian revolution, reinforced by the rise of mathematical economics, but there is much more to the story than that.

The key to this picture is the so-called 'old' institutionalism. In the interwar period, Institutionalism was a very broad movement aimed at making economics more scientific through placing it on firmer empirical foundations. Though its boundaries were very blurred, it is reasonable to see it as covering economists as diverse in their empirical work as Wesley Mitchell, Simon Kuznets, Gardner Means, John Commons and John Maurice Clark. Though they had connections with economists in Europe, it was a distinctively American movement. Mathematical economics is inherently less culturally specific, but even here there were distinctive American approaches to the subject: Paul Samuelson's early work, under E.B. Wilson, drew on a type of mathematics very different from that used by Europeans. The same could be said about the early econometricians, from Henry Ludwell Moore to Henry Schultz: there were important European parallels, but they were pursuing research in a way that was distinctive. Monetary economics illustrates both the distinctiveness of American economics and the blurred boundaries between different approaches to the subject. Because the Federal Reserve System had been established much later than the major European central banks, there was much more lively debate over the principles on which it should be run. The result was a rich mixture of arguments spanning the divides between neoclassical and institutionalist, Harvard and Chicago, Banking School and Currency School.

The Second World War was important for several reasons. First, economics became tied up with the war effort. Economists were clearly involved in places such as the Office of Price Administration, the Treasury or the War Production Board. It was in the last of these that national income statistics, first calculated in 1933, were developed into a system of national accounts, providing the basis for planning the massive shift of resources from civilian to military production that took place after 1941. However, perhaps more significantly, economists became involved in fighting the war, primarily through the Office of Strategic Services (OSS), forerunner of the Central Intelligence Agency which employed around 50 economists under Edward Mason in its Research and Analysis division (Leonard 1991; Katz 1989). In the OSS, economists and other social scientists were employed alongside physicists and other scientists in tasks where economics shaded imperceptibly into statistics and engineering. They analysed intelligence, and became intimately involved in questions of military strategy and tactics, emerging from the war with an enhanced reputation: not only was economic analysis important, but many economists had proved themselves useful as general problem solvers. Operations research, a set of techniques centered on optimization subject to constraints, came to be much more central to economics.

At the end of the war, the Servicemen's Readjustment Act (1944), the so-called G.I. Bill, offered financial support to US ex-servicemen who wanted to continue their education. This fuelled a dramatic increase in the university system. The number of bachelor's degrees awarded in US higher education institutions, which had never risen above 187,000 before the war, rose to 271,000 in 1947–8 (317,00 if higher degrees are included too) and 432,000 in 1949–50, many of these choosing to study economics. This accelerated the generational shift that was taking place, providing academic openings for economists returning from government service to civilian life, many of these being in institutions that had

not been prominent before the war. While some economics departments continued as before, there was a shift in the profession's center of gravity away from places such as Wisconsin (the leading centre for Institutionalism) towards ones like MIT, Berkeley and Stanford – other places, such as Columbia, Harvard, Chicago and Yale were important before and after the war (Barber 1997; Backhouse 1998). The subject began to be taught using textbooks written by young economists (Kenneth Boulding, Lorie Tarshis and Paul Samuelson) during the 1940s, in place of ones that had their origins nearer the turn of the century.

The Cowles Commission and RAND

A particularly important center for quantitative work in the 1940s was the Cowles Commission, which had moved to Chicago in 1938, and of which Jacob Marshak became Research Director in 1943. He laid out a programme of research focusing on the development of new methods to take account of the specific features of economic data, perceived to be simultaneity, the importance of random disturbances and the prevalence of aggregate time-series data. This programme proved to be one that attracted American economists, many of whom were involved in the war effort, and many of the highly technical European emigrés such as Trygve Haavelmo, Tjalling Koopmans and Abraham Wald. Not surprisingly, the operations research side of economics was dominant here, not simply in obvious ways, such as the work by Koopmans and George Dantzig on linear programming and the simplex method, but in the broader conception of economics as engineering. This was not confined to the Cowles Commission (one can see such an influence at other places such as Massachusetts Institute of Technology, MIT) but such work clearly centred on Cowles.

The important idea that emerged from this phase of the Cowles Commission's work was that the economic system could be analysed as a probability distribution, the task of economics being to identify the properties of that distribution. General equilibrium theory, embracing individual optimization within a system of simultaneous equations, provided an account of the structural relationships. Statistical methods, pioneered by Haavelmo and Koopmans, provided the means for relating that theory to data that exhibited random shocks in addition to any systematic relationships between the data, not just estimating coefficients but also testing the theory. This has been called a probabilistic revolution (Morgan 1991). Controversy over these new methods erupted in 1947 when Koopmans (1947) challenged, head on, what had previously been considered the scientific way to do empirical economics – the National Bureau of Economic Research's (NBER) meticulous data-gathering and comparatively informal data-analysis – represented by *Measuring Business Cycles,* by Burns and Mitchell (1946). As Rutledge Vining (1949), replying for the NBER, justly claimed, Koopmans had written a manifesto for the Cowles Commission's new methods, privileging the testing of theory over the search for hypotheses. In addition to the techniques mentioned above, its fruits ranged from the monetary theory that formed the heart of Don Patinkin's *Money, Interest and Prices* Patinkin (1956), the leading exposition of macroeconomic theory till the 1970s, and Lawrence Klein's models of the US economy, from which developed much of macroeconometrics.

In the late 1940s, the RAND Corporation, in Santa Monica, California, emerged as a new focus for technical economic analysis. RAND was initially a division of the Douglas Aircraft Company, but from 1948 it became a non-profit organization, funded at first by the US Air Force, and later by other bodies, of which the Ford Foundation was the most important. It was an interdisciplinary environment where economists worked alongside scientists, mathematicians, engineers and other social scientists. It was established by senior figures in the US Air Force and was motivated by the Soviet threat and the emerging Cold War, and embodied lessons learned in the Second World War. Through H. Rowan Gaither, Chairman of RAND's Board of Trustees and, from 1953, President of the Ford Foundation, RAND became closely linked to Ford: its overall product was 'systems analysis', a broad umbrella under

which a range of mathematical work could be sponsored, linked primarily by certain sets of mathematical techniques and a vision of economics centered on rational choice. Though it was associated with much else, from Kenneth Arrow's *Social Choice and Individual Values* Arrow (1951) to *Linear Programming and Economic Analysis* by Dorfman et al. (1958), its main significance was in game theory, bringing together economists from Cowles (such as Arrow) with economists and mathematicians from Princeton (which included John Nash), the major academic centre of research into game theory during the 1950s.

The precise significance of the military involvement in economics is not yet clear. The Office of Naval Research (ONR) provided much funding, especially for game theory research, and the US Air Force was behind RAND. Clearly some projects were directly driven by military imperatives, such as working out a strategy for responding to (or anticipating) a Soviet nuclear strike. There are also clear links from systems analysis/operations research, and the techniques associated with these, to military requirements. Against this, those involved emphasize that researchers at RAND were given great freedom and military sponsorship had little or no effect on what they did (see Mirowski 2002). However, even if researchers did have a high degree of freedom, there was certainly selection bias in the types of projects and researchers who received support from these sources, and the scale of such funding makes it plausible to argue that may have had a significant effect on the way the profession developed.

Mathematics, Technique and the 'Core'

To say that economics has become more mathematical in the post-war period is too obvious to need justification. However, the significance of this process and the way it came about are far less obvious and need disentangling.

Mark Blaug (1999) has labelled what happened to economics after the 1950s 'the formalist revolution'. However, within this lie a number of very different developments. One is the incursion into economics of formalist mathematics. At the outset of *The Theory of Value,* Debreu (1959) wrote that he was approaching his subject with the degree of rigour associated with the contemporary formalist school of mathematics. His work formed part of a broader movement towards placing economic theory on an axiomatic foundation, and comprising the literature on existence, uniqueness and stability of general equilibrium (see Weintraub 2002). Even here, however, it is possible to discern strands that on closer inspection are very different. Ingrao and Israel (1990) distinguished the formal and interpretive branches, associating the former with Debreu and the latter with Arrow. Others have traced differences to disputes over formalism in mathematics (Weintraub 2002). The most eminent mathematician to engage with economics, John von Neumann, was not only a critic of formalism (in the sense of Hilbert): his interest in economics stemmed from a broader concern with artificial intelligence that, Mirowski (2002) has argued, differentiated his views sharply from economists at the Cowles Commission and others pursuing general equilibrium analysis.

More significant than this is the fact that most economics, as Solow (1997) has observed, is not formalistic in this sense. Rather, what has happened is that economics has become more 'technical': he was probably right to argue that axiomatics was of no interest to most economists. Perhaps the most influential exponent of mathematics in economics, Paul Samuelson, whose *Foundations of Economic* Analysis Samuelson (1947), written at Harvard and the basis for the style of economics he established at MIT, amounted to a manifesto for mathematical economics. His work, which arose from a mathematical tradition very different from the European traditions out of which von Neumann and Debreu came, sought to be rigorous without being based on axiomatization. Even further from formalism, but equally influential, was the Chicago School, dominated from the 1940s to the 1970s by Milton Friedman. Friedman favoured simpler models and was more sceptical about complex mathematical reasoning. Thus Hands and Mirowski (1998) have distinguished three schools in post-war neoclassical price theory – Stanford (Arrow), MIT

(Samuelson) and Chicago (Friedman). Whether or not one accepts the claim made by Hands and Mirowski that these represent three responses to the failure of Henry Schultz's attempt to quantify demand theory before his death in 1938, this provides a useful way to represent the variety of ways in which a common theoretical core was developed.

However, becoming more technical is not synonymous with using mathematics. Another dimension is the separation of theory and application. Though the distinction between theory and applied work is taken for granted by most contemporary economists, the situation was very different before the Second World War. There was much work where it is impossible to draw any distinction between statements that are intended to describe the world and ones that are at the level of theory. In what we would now consider applied work, the practice of clearly separating theory and application is something that emerged only after the Second World War (see Backhouse 1998). This change is reflected in the language economists began to use: they began to talk in terms of models. Though the idea of a model has deeper roots, talking in terms of models took off only from 1939, having been very rare before that.

As the discipline changed, so did the curriculum, something in which the American Economic Association (AEA) became involved. During the 1940s, partly in response to demands of wartime, and partly because of broader uncertainty about how economics should be taught, the AEA established committees on undergraduate education, the main outcome being a report in 1950. Out of this rose the suggestion to review graduate education, resulting in a report by Howard Bowen, sponsored by the AEA and funded by Rockefeller, which appeared in Bowen 1953. On the grounds that 'technical knowledge' of economics would be useful for those working as economists in government, business and education, this argued that 'there should be a "common core" for all students who are to be awarded advanced degrees in economics' (Bowen 1953, p. 2). This core consisted 'primarily of economic theory including value, distribution, money, employment, and at least a nodding acquaintance with some of the more esoteric subjects such as dynamics, theory of games and mathematical economics'. No one, it was argued, had claim to an economics Ph.D. without 'rigorous initiation' into these and economic history, history of economic thought, statistics and research methods (Bowen 1953, p. 43). Interestingly, mathematics was placed alongside Russian, German and Chinese: it was important to have some economists with knowledge of it, but it was not necessary for all to do so.

In Bowen's report, the core was still very broad – a statement of the range of knowledge – that economics Ph.D's should be expected to have. Over the following two decades it came to be used more narrowly. For example, Richard Ruggles (1962, p. 487) wrote of the function of graduate training being 'to provide a common core of basic economic theory' that would be used elsewhere in the programme, and observed that 'at a great many universities' training in mathematics was required. However, this was still discussed alongside language requirements. Though such questions had been raised as early as the Bowen Report, it was in the 1960s that the AEA meetings hosted debates over the role of economic history and history of economic thought in the graduate curriculum. Gordon (1965) conducted a survey implying, as Bowen had found a decade earlier, that though most graduate schools still offered the subject, history of economic thought was declining, and that there was pressure for it to decline further, particularly from younger faculty. In the survey by Nancy Ruggles (1970), the subject was defined in the now familiar way of a unifying core of micro and macro theory, quantitative methods (interestingly, econometrics, simulation, survey methods and operations research) and a range of applied fields that did not include any history.

These trends continued to the end of the century. They are best summed up by saying that economists were increasingly being trained, at Ph.D. level, as technicians rather than as scholars in the traditional sense of the term. In the 1940s, when concerns were raised about this in AEA meetings, it was still plausible to respond the demands of scholarship, and breadth of education,

were compatible with mastering the necessary technical skills; but by the 1970s this was becoming more and more difficult. The demands of technique were pushing courses that provided breadth, symbolized by history of economic thought, out of the curriculum. By the end of the 1980s, this had gone so far that some Liberal Arts professors claimed that Ph.D.'s from the leading graduate schools were no longer equipped to teach at undergraduate level: not only did they know too little of the past and present literature on economics, but they did not know enough about the institutions of contemporary market economies. There were even signs that Ph.D. students themselves were sceptical about the value of the hurdles over which they were jumping (Colander and Klamer 1987, 1990). In response to these concerns, the AEA established a Commission on Graduate Education in Economics, which reported in 1991 (Krueger 1991; Hansen 1991; see Coats 1992a, for a comparison of this and Bowen's report). Though some changes were recommended, these were minor and had little effect (Colander 1992). When Colander (2005) repeated his earlier survey a decade and a half later, he found much lower levels of dissatisfaction, though concluded that the students had adjusted to the more technical syllabus, not the other way round.

Economic Analysis

The way the use of mathematics spread within economics was inextricably linked with developments in economic analysis. It was not simply a matter of making earlier, less rigorous, analysis more precise. To be able to use the mathematical techniques in the way they did, the basis on which economics rested had to change. For the institutionalists, very broadly interpreted, being scientific meant basing economics firmly on evidence about how the world worked. It was because he believed that empirical research would establish accounts of human behavior that were more complex than those offered by economic theorists that Mitchell (1925) had predicted that economists would lose interest in an abstract, artificial man. The result was that the 1930s and 1940s saw a wealth of empirical work on industrial organization, pricing, labour markets and many other aspects of economic behaviour. But mathematical theory, given the techniques then available to economists, necessitated working with simpler assumptions, in which agents were maximizers operating in markets where competitive structures were precisely defined, and if possible were perfectly competitive. It is perhaps because of the strong institutionalist element in American economics that the debates through which these simplifying assumptions were established were dominated by American economists.

The clash between institutionalism and new, technical approaches explicitly came to the surface in Koopmans's review Koopmans (1947) of *Measuring Business Cycles, by* Burns and Mitchell (1946). Koopmans presented his approach as building on the work of those like Burns and Mitchell who simply measured: it was necessary to pass beyond that to the 'Newton stage' in economic theorizing, where theory and data analysis informed each other. Data would be used to test theory. Representing the old view, Rutledge Vining (1949) pointed out that the Cowles Commission methods involved more than this: that they presumed a specific type of theory and empirical methods. If one did not know what theory was suitable, different empirical methods were required. Burns and Mitchell, Vining argued, were concerned with discovery as much as with testing, for, in the absence of empirical work, economists did not know what form theory should take.

Substantially the same issue arose in the so-called 'marginalist' controversy, provoked by Richard Lester's article in the *American Economic Review* Lester (1946). Though Lester was portrayed by critics as drawing naive conclusions from surveys, and as presenting a radical challenge to profit maximization, he is better seen as arguing that economics needed to be based firmly on the mass of evidence that had been accumulated during the previous decade or more on how firms behaved and on how labour markets worked. Controversy here was more prolonged and more complex, spilling over into discussion of industrial organization, where Lester's critics included economists on both sides of the divide

between Harvard (dominated by Mason and Chamberlin) and Chicago (where Friedman and Stigler were; see Lee 1984; Mongin 1992). Fritz Machlup changed the debate into one about marginal analysis and, together with Friedman, established the principle that economics was about explaining behavior, not explaining how decisions were made. Though he did not intend it that way, Friedman's (1953) methodology of positive economics, with its emphasis on testing predictions, not assumptions, could be taken as vindicating economic theory's neglect of its empirical foundations.

It is no coincidence that these controversies took place in the pages of US journals, as did a less prominent one a few years later on the role of mathematics in economics (Novick 1954, and ensuing discussion; see Mirowski 2002, pp. 402–5). There were parallels in other countries, but it was in the United States, where in the 1930s institutionalists and neoclassicals had vied with each other, that the most marked cleansing of approaches that could not be formalized so easily was taking place. In the 1950s and 1960s, formal modelling based on maximization and, increasingly, competitive markets spread throughout the discipline. Price theory became more formal and increasingly dominated what came to be known as microeconomics. Keynes's behavioural microfoundations (based on 'propensities' and imprecise generalizations from observed behaviour such as 'animal spirits') were replaced with optimizing ones and macroeconomics came to be seen through the lens of utility maximizing agents, as in Don Patinkin's *Money, Interest and Prices* Patinkin (1956). General equilibrium analysis, summed up by Debreu's *Theory of* Value Debreu (1959), though never more than a minority activity, came to be seen as the fundamental theory on which more workaday theorizing rested.

During this period, however, there were limits to the application of formal theory. Though formal microfoundations could be provided for many of the functions, macroeconomics was seen as separate, not entirely reducible to a single, consistent microeconomic theory. In microeconomics, strategy and industrial structure remained outside the purview of formal theory, empirical work dominating work on industrial organization. Development economics offers another example of a field that stood apart from other fields, reflecting the assumption, held widely though not universally, that people in different societies behaved in different ways. Thus, although economists later came to see the rise of formal theory and mathematical methods as the key development during the period, its progress was slow, and it was anything but pervasive as late as the 1960s. The way in which less formal approaches, based on assumptions that ran directly counter to those underlying what later became dominant, is nicely illustrated by a project entitled 'The Inter-University Study of Labor Problems and Economic Development', undertaken by John Dunlop, Clark Kerr, Frederick Harbison and Charles Myers (see Cochrane 1979). Its thesis, that industrialism required the development of a new type of man, ensured that, as the assumptions underlying modern theory became established, it came to be seen as a quaint relic of the past. However, its significance rests in its being a large project, lasting over 20 years, receiving $855,000 from the Ford Foundation and $200,000 from Carnegie, producing around 40 books and, in Cochrane's view, helping to define labour economics as that field existed in the late 1970s. Though its final report came as late as 1975, its objectives were framed against the background of thinking on labour questions that the older institutionalist economists would have understood; its analysis drew on sociology and industrial relations as well as on what would now be recognized as properly economic analysis.

However, from the 1970s things changed. Formal methods, based on individual optimization, were used to analyse problems of uncertainty and information, the most prominent exponent of this being Joseph Stiglitz. These ideas were applied to labour markets, finance, and many other fields. Macroeconomics turned away from what Lucas called 'free parameter' models – ones containing parameters that were not based on optimizing behaviour. Public choice theory, which emerged at the boundaries of economics and political science, brought government and much organizational behaviour within the scope of rational choice (see Medema 2000). Initially, this was not

widely considered to be economics, with the result that public choice scholars found it hard to publish in the major journals, leading James Buchanan, Gordon Tullock and their associates to establish the Public Choice Society, and to develop their own journals. However, fairly soon the main economics journals opened up to such work. Methods were found to build models of general equilibrium with monopolistic competition, these enabling trade theory to move away from assumptions of perfect competition that were thought unrealistic in many contexts. Game theory was introduced to analyse problems of strategy, first transforming industrial organization and later being extended to almost every other field of economics. Development economics, like macroeconomics, ceased to be considered as resting on principles different from those in the rest of the discipline. Rational choice methods were even applied by economists with clear radical sympathies, such as in the rational choice Marxism of John Roemer and Jon Elster.

Most of these developments were international in their reach, but all were centred, squarely, in the United States. In none of the cases just mentioned would it be conceivable to write the story without discussing the role of American economists and economists based in the United States, whereas it would be possible (if not always the full picture) to do so without mentioning work in the rest of the world. The collective effect of these developments was to transform a discipline in which, though rational, maximizing behaviour was central, numerous exceptions and special cases existed, to one where it could plausibly be argued that economic theory was simply working out the implications of maximizing behaviour. Economic theory could be seen as resting on a single behavioural postulate.

Faced with this scenario, in which economic theory in the United States became, methodologically, narrower, some economists rebelled. Radical economists, stimulated by the Vietnam War, began to argue in the late 1960s, that economists were systematically ignoring questions such as power, class, and income distribution. Frustrated by their inability to persuade their more orthodox colleagues to take their ideas sufficiently seriously, they formed the Union of Radical Political Economy, setting up networks, conferences and a journal (Coats 1992b, 2001). Shortly afterwards, inspired by Joan Robinson's Ely Lecture at the AEA in 1971, Alfred Eichner, Jan Kregel and others organized what developed into the grouping known as Post Keynesian economics (Lee 2000). 'Austrian economists', influenced by the work of Ludwig von Mises, Friedrich Hayek, and Ludwig Lachmann, encouraged by Hayek's being awarded the Nobel Memorial Prize in 1974, also began to organize themselves. In all cases, the organization of these groups was motivated by the sense of exclusion they felt from the mainstream, represented by the meetings of the American Economic Association and the leading economics journals, who regarded their work as generally of low quality. These movements remained small, with strengths in particular institutions (Austrians at New York and Auburn; Post Keynesians at Rutgers and Tennessee; Public Choice in Virginia; Radicals at Amherst and the New School).

These self-consciously 'heterodox' groups were but part of a wider fragmentation of the discipline. Technological changes meant that economic print runs for books and journals fell during the period, and the costs of travel and communications fell. Together with the increased size of the economics profession, these developments made it easier for sub-fields of economics to organize, represented most clearly by the rapid rise in the number of specialist journals. The changing character of the profession was reflected in the Allied Social Science Association (ASSA), the main professional meeting of American economists, organized by the AEA. By 1998, though the AEA, the American Finance Association and the Econometric Society dominated the meetings, there were 52 societies affiliated with the ASSA (compared with 34 in 1980), and the AEA was having to restrict the number of sessions these societies were organizing, something it had not done two decades before.

The 1970s and 1980s were arguably years of integration, when the American economics became more homogeneous, the core of microeconomics based on individual optimizing behaviour being applied to more and more. More than ever

before, and in dramatic contrast with the situation before the Second World War, there was an orthodoxy. However, this was questioned and developed at both empirical and theoretical levels. One reason was that developments in data collection and in computing meant that economists were able to analyse the behaviour of real individuals in a way that economists of earlier generations (see, for example, Mitchell 1925) could do little more than dream about: 'microeconometrics' was recognized with the 2000 Nobel Memorial Prize for James Heckman and Daniel McFadden. It became possible to engage in quantitative analysis of microeconomics as never before. Another reason was that economists turned to experimentation as a source of data: experimental economics, considered esoteric as late as the 1970s, was given respectability with the debates over preference reversals in the pages of the *American Economic Review* and the award of the 2002 Nobel Memorial Prize to Daniel Kahneman and Vernon Smith. This rapidly spread throughout the profession. When 'behavioural' economics started being taken seriously in finance, a field where predictive power was always paramount, it was a sign that alternatives to conventional views of rationality were being taken very seriously. Bounded rationality, on which Herbert Simon had been working since the 1950s at Carnegie Mellon, and for which he got the Nobel Memorial Prize in 1978, moved from being something idiosyncratic, if respected, to being a mainstream technique. Evolutionary game theory and complexity theory offered new ways to think about economic change that expanded the boundaries of what was accepted in the subject. By the end of the century, though rational choice models remained immensely strong, it became much harder to describe economics as dominated by an orthodoxy. Once again, though these developments were international in their scope, they were centred on the United States, just as were the developments of the 1970s. Ideas whose main supporters were European, such as the competing views of consumer theory associated with Werner Hildenbrand (who derived demand functions from assumptions about distributions of characteristics across individual consumers), had far less influence.

Economists, Ideology and Policy

Ideology was never far from the surface. In the 1940s concern with 'Reds' was common in the United States, though economists might consider the problem only to dismiss it. After 1945, as the Cold War developed, these concerns with Communism grew, reaching their peak with Joseph McCarthy's search for Communist sympathizers. Economists had frequently been viewed with suspicion amongst businessmen, some of whom were important patrons of higher education, but the stakes were raised. Planning was suspect, a legacy from the days of the New Deal, and Keynes provided a convenient focus, for he was a more real threat than Marx: according to the *Chicago Tribune,* he was the Englishman who ruled America (see C.D.W. Goodwin 1998, on these episodes). Influential figures argued that Keynesianism was tantamount to Communism. Textbooks, such as those of Lorie Tarshis and Samuelson, that discussed Keynesian theory were attacked and sometimes removed from syllabi under pressure from aggrieved sponsors (Colander and Landreth 1996, 1998).

The cases where economists were forced out of academic positions because of real or alleged Communist sympathies are comparatively easy to document (Goodwin 1998; Lee 2004). What is much harder to prove is the effect this had on how economics pursued their work. There were certainly great pressures to be technical, for arcane communications between specialists were much less likely to be considered suspect than ideas that reached out beyond academia. Using an evolutionary model, Goodwin (1998, p. 79) distinguishes between 'conceptual variation' and 'intellectual selection' , arguing that the attitudes of economists' patrons must have influenced the latter. However, doubts about its closeness to communism did not prevent Keynesianism from becoming widely accepted in academia, though that may have contributed to its being expressed in more careful, technical language than might otherwise have been the case (see Colander and Landreth 1996, p. 172).

This bias towards becoming more technical chimed with another pressure – to be seen as doing science. When the National Science Founda-

tion was established in 1950, the inclusion of social science was controversial and did not take place till 1956. If economists were to obtain support, they had to ensure that their work was seen as scientific. Given prevalent beliefs about science at the time, this favoured narrower, more technical work, and worked against the pluralistic interdisciplinarity that had been more common before the war (Goodwin 1998, pp. 65–7). Similar issues arose in the context of support by philanthropic foundations, of which Sloan, Russell Sage, Rockefeller and Ford were the most important. Here, concern with being rigorous was intertwined with suspicion of planning and doubts about Keynesianism.

Similar considerations affected the body that brought economists into the heart of the US government, the Council of Economic Advisers (CEA) established by the Employment Act of 1946. This was intended to be a conservative institution, providing expert advice with minimal government interference. Its first chair, Edwin G. Nourse, shared this view: unlike his colleagues on the CEA, he was careful to avoid being seen as an advocate for White House policy (Bernstein 2001, pp. 110–11). Unlike his successor, Leon Keyserling, he viewed economics as providing technical, disinterested expertise. Despite criticism, the CEA survived, achieving its greatest influence in the Kennedy administration, when Walter Heller, Kermit Gordon and James Tobin applied Keynesian demand-management policy to the problem of reducing unemployment.

Given that President Lyndon Johnson would not let it compromise his Great Society program, the escalating war in Vietnam led to rising federal deficits. CEA members warned the President about the consequences of this, but the CEA's Keynesian policies were nonetheless blamed for the inflation and dislocation that followed during the 1970s. After 1979, alongside the decline of Keynesianism in academia, influence on stabilization policy rapidly shifted to the Federal Reserve under Paul Volcker and later Alan Greenspan. This shift from the CEA to the Fed marked not a decline in the influence exerted by economists, but a change in its structure: there was a convergence between research done in academia and in central banks and other agencies (see McCallum 2000, p. 123), and a shift of emphasis towards microeconomic policy. Economists increasingly saw their role, not as engineers advising on how to operate fiscal and monetary levers, but as designers of institutions and of systems that would achieve desired outcomes in a world where policymakers were seen as part of the system rather than outsiders manipulating it. Frequently this involved creating new markets, or 'reinventing the bazaar' (McMillan 2002).

There were also more conscious attempts to impose an ideological agenda on economics. RAND, the most influential think tank in the 1950s, became closely involved with the Ford Foundation (these arguments are developed in Amadae 2003). It was explicitly a non-political organization, directed towards impartial research. However, under the chair of its board of trustees, H. Rowan Gaither, also president of the Ford Foundation after 1953, RAND focused on 'systems analysis', based on principles of rational action. Rational choice, central to the work of RAND since its inception, could be seen as providing, though its focus on the independent individual, a justification for a free society, and an alternative to Communist collectivism (see Amadae 2003). RAND's ideology, like that of Ford, was one of technocratic management, by experts using rigorous quantitative techniques. This ideology became prominent in government in the 1960s when applied by Robert McNamara, who came from the Ford Motor Company, as Secretary of Defense.

Others had a more explicit ideological agenda. The American Enterprise Institute (established 1943), The Foundation for Economic Education (1946) and the Liberty Fund (1960) were established specifically to propagate free-market ideas. There were followed in the 1970s by a series of think tanks specifically to develop and apply such ideas to policy. The Heritage Foundation (1973) was specifically seen as providing a counterweight to the Brookings Institution (established 1927), which had come to be seen as part of finely tuned liberal policymaking machine (liberal being understood in the American sense). The aim of its president, Edwin Feulner, was to create 'a new conservative coalition that would replace the New

Deal coalition which had dominated American politics for half a century' (L. Edwards 2005, p. 371). When Ronald Reagan took office, the Heritage Fund provided policy ideas ready to put before the new administration. Hayek, who had moved to Chicago in 1950, played a particularly influential role in stimulating such organizations, within the United States and elsewhere, being part of an influential network centred on the Mont Pelerin Society, an international group of libertarian thinkers established in 1947, whose founders included four Chicago economists and representatives from the Foundation for Economic Education.

Businessmen and conservative foundations also sought to stimulate free market thinking within economics, many of them effectively targeting specific institutions and programs. Though tiny compared with the big foundations such as Rockefeller and Ford, the Volker Fund (which supported Hayek and Mises), the Earhart Foundation (with a programme of one-year fellowships), the Scaife, Bradley and Olin foundations (which between them targeted support at, *inter alia,* Chicago's Law and Economics programme, and various centres of public choice theory in Virginia) managed to achieve influence out of proportion to their size (see Backhouse 2005).

The International Dimension

After the Second World War, the United States dominated the economics profession. Not only had German economics been devastated by the Nazi Party, but the resulting emigration contributed enormously to the expansion of American economics. The United States was not the only home for German and other European exiles, many moving to Britain, but it received more than any other country. Britain experienced no such loss, but its university system was too small for it to be a serious rival. The result was that many ideas that had originated in Europe rapidly came to be associated with the United States. The clearest examples of this are general equilibrium theory and econometrics, where European emigrés, led by Jacob Marshak, were instrumental in developing ideas that rapidly lost any close connection with their European origins. For example, in the 1930s, general equilibrium analysis had been an almost exclusively Viennese phenomenon (in Karl Menger's seminar), whereas by the 1950s, not only had those who worked on it there (Wald and von Neumann) moved to the United States, but its leading practitioners were an American (Arrow) and a Frenchman (Debreu), but both based in the United States.

Another clear example of this process is Keynesian economics, central to the evolution of American economics from the 1940s to at least the 1980s. This clearly originated in Britain, and British economists such as John Hicks and James Meade played important parts in the subsequent Keynesian revolution. However, Keynesianism was rapidly Americanized. The key figure here was Alvin Hansen, the force behind Harvard's fiscal policy seminar, and later author of the influential *A Guide to Keynes* Hansen (1953). As has been argued by Mehrling (1998), Hansen's 'conversion' did not involve a rejection of his earlier ideas; rather, Keynesianism provided a vehicle through which his ideas on policy, rooted in the American institutionalist tradition, could be developed. Lawrence Klein (1947) provided another interpretation of Keynesian economics, relating it to with the econometric approach emanating from the Cowles Commission. Samuelson (1948) integrated Keynesian ideas into a textbook aimed at American students. During the 1950s and 1960s, the most influential work on macroeconomics was, with few exceptions, undertaken in the United States. Friedman's work on the consumption function provides another example of Keynesian ideas being assimilated into an American tradition (the empirical studies of the NBER).

What was happening here is that economics was becoming more international, but centred on the United States, a development made possible by the openness of the American system at a time when the profession was expanding and opportunities for immigrants were great. The United States dominated, not simply because of its size, but because of its resources. During the interwar period, the Rockefeller Foundation had been instrumental in building up economics in key European institutions in Britain, Scandinavia and

many other countries (cf. Goodwin 1998). After 1945, given the close American involvement in Europe that resulted from the war and reconstruction, this influence increased, accelerated by the reduced cost of international travel. In country after country, the economics profession changed in several ways. Academic systems became more open and competitive, with increased emphasis on publication in journals. There was a movement away from publication in the native language towards publication in English. Journals moved away from being national organs to ones that published articles by economists from a wide range of countries. Graduate education moved towards the American model, away from the traditional European model of a major thesis, publishable as a book, towards a Ph.D. comprising advanced coursework and a short thesis that could be the basis for three journal articles. The mathematical demands made of students rose progressively. Many economists either undertook postgraduate study in the United States or spent sabbaticals in US universities.

The speed and extent of these changes varied enormously (see the case studies in Coats 1997). For example, in the UK, the proportion of staff with a degree from an American university rose steadily from 1950 to the 1990s. The highest proportion was at the London School of Economics, where it reached 45 per cent by the mid 1990s, whereas in other universities it was only five per cent. In Belgium, CORE at the Université Catholique de Louvain was an important centre for economists with strong US connections. Similarly, there was variability in the speed with which Ph.D. requirements changed, some British universities adopting the American model in the 1950s and 1960s, while others did not require any coursework beyond undergraduate level till the 1990s. In Continental Europe, there was the complication of language, and in many cases of academic systems that were much more rigid and less rapid to change, but many of these changes still took place. Outside Europe, there was the further factor of decolonization. At the end of the Second World War, many countries were still closely linked to former colonial powers, and the changes involved a switch from those to the United States.

There is dispute over whether this process should be labelled 'Americanization' or simply 'internationalization' (see Coats 1997, pp. 395–9). The process certainly did involve internationalization, and it was arguable that many changes (such as the move towards advanced coursework) were necessitated by the rising technical demands made by the subject. As has already been explained, many of the ideas on which the period's economics was based were European, not American in origin. Arguably, the United States appeared to dominate what was primarily an international system simply because of its size. However, there are strong reasons for considering the process as involving Americanization. In many cases, in Europe and elsewhere, the United States provided an example that was deliberately copied. In other cases, changes were brought about through connections with US universities. Harry Johnson, Canadian, but a professor at Chicago and Geneva, was important in bringing about changes at LSE where he also held a chair in the 1960s. Chicago economists developed close links with Latin American countries, consciously exporting Chicago economics to Chile: Chilean students studied in Chicago, and Chicago staff taught at the Catholic University of Chile (see A. Harberger 1997; Valdes 1995). Similar developments took place in Brazil, though involving a much wider range of universities: Chicago, Berkeley, Harvard, Yale, Michigan, Illinois and Vanderbilt (see M.R. Loureiro 1997). The US Agency for International Development and the Ford Foundation provided a significant role in funding several of these inter-university agreements.

Similar remarks could be made about the US role in the international organizations that emerged after 1945, notably the International Monetary Fund (IMF) and the World Bank: they were vehicles for the internationalization of economics, along a model dominated by the United States.

Conclusions

Since the Second World War, economics in the United States has been transformed as dramatically as in its previous half-century. It is inevitable that economists looking at these changes focus on the economic ideas themselves, usually telling the story

as one of progress. However, this transformation involved changes in the structure of the profession (notably in the nature of graduate education) and its place in society as well as changes in economic analysis. It is natural for historians to focus on connections between economic ideas and the institutions out of which they arose. To do this raises the question of whether these external factors influenced the course of economic ideas: of whether things could have been different. The difficulty here is that it is hard to construct a plausible account of how things might have been different because, even if it was the result of adaptation to chance events, what actually happened generally looks inevitable in retrospect. However, what can be done is to sketch some of the possible routes that could have been taken but were not.

Different paths were open for American economics in the 1930s. During the New Deal, economists such as Gardner Means had built up an enormous body of statistical data on how product and labour markets operated. From this starting point, economists could have chosen to build models that were less general but more securely rooted in specific institutional detail than was Walrasian general equilibrium theory. The route the discipline actually took was determined, *inter alia,* by the Second World War and the Cold War and the encouragement it gave to certain types of theorizing and certain types of empirical work. Macroeconomics offers a second account of alternative paths that might have been taken. The interwar literature contained discussions of rational expectations, dynamics, intertemporal equilibrium and credibility of policy regimes (see Backhouse and Laidler 2004). However economists did not pursue such ideas but developed a macroeconomics centred on a static equilibrium framework (the IS–LM model); much of the dynamics that had been lost was then 'rediscovered' in the 1970s. Had they developed a different set of ideas from the interwar literature (even from Keynes's own work), macroeconomics could not have developed as it did.

To argue in this way is not to claim that one path was right and the other wrong, or even that economists were aware of the directions in which their own theoretical choices (aimed at solving specific problems) would lead. Instead, historical accounts generally rest on two pillars. First, the standards by which economists judge their work – their standards of scientific rationality – reflect the intellectual climate in which they are working. In some cases, likely factors can be identified. For example, Weintraub (1998) has argued that conceptions of what it meant to be rigorous changed dramatically as a result of developments in quantum mechanics. But many of the influences on the criteria economists use to assess their work are harder to identify and have to be disentangled, cautiously, out of the historical record. Second, evolution requires not simply a mechanism for generating new ideas but also a selection process. To understand the way economic ideas have developed since the Second World War, it is necessary to consider the demand for economic ideas as well as the supply: thus, even if the economists are resolutely impartial, applying high scientific standards to their work, the identities and view of their patrons may serve, through favouring some types of inquiry rather than others, to affect the evolution of the subject. We may be too close properly to understand many of the connections, but economics during this period, just as much as the economics of earlier centuries, cannot be divorced from the institutional setting in which it developed.

See Also

- ▶ United States, Economics in (1776–1885)
- ▶ United States, Economics in (1885–1945)

Bibliography

Amadae, S.M. 2003. *Rationalizing capitalist democracy.* Chicago: Chicago University Press.

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

Backhouse, R.E. 1998. The transformation of U.S. economics, 1920–1960, viewed through a survey of journal articles. In *From interwar pluralism to postwar neoclassicism,* eds. Morgan, M. and M. Rutherford. 85–107. Durham: Duke University Press. *History of Political Economy 30* (Annual Supplement).

Backhouse, R.E. 2005. The rise of free-market economics: Economists and the role of the state since 1970. In *The*

role of government in the history of economic thought, ed. S.G. Medema and P. Boettke. Durham: Duke University Press. *History of Political Economy* 37 (Annual Supplement).

Backhouse, R.E. and Laidler, D.E.W. 2004. What was lost with ISLM. In *The IS–LM model: Its rise, fall, and strange persistence,* ed. K.D. Hoover and M. De Vroey. Durham: Duke University Press. *History of Political Economy* 36 (Annual Supplement).

Barber, W.J. 1997. Postwar changes in American graduate education in economics. In *The post-1945 internationalization of economics,* ed. Coats, A.W.. Durham: Duke University Press. *History of Political Economy* 28 (Annual Supplement).

Bernstein, M. 2001. *A perilous progress: Economists and public purpose in twentieth century America*. Princeton: Princeton University Press.

Blaug, M. 1999. The formalist revolution or what happened to orthodox economics after World War II? In *From classical economics to the theory of the firm: Essays in honour of D.P.O'Brien*, ed. R.E. Backhouse and J. Creedy. Cheltenham: Edward Elgar.

Bowen, H.R. 1953. Graduate education in economics. *American Economic Review* 43(4, part 2): 1–223.

Burns, A., and W.C. Mitchell. 1946. *Measuring business cycles*. New York: NBER.

Coats, A.W. 1992a. Changing perceptions of American graduate education in economics, 1953–1991. *Journal of Economic Education* 23: 341–352.

Coats, A.W. 1992b. Economics in the United States, 1920–1970. In *On the history of economic thought: British and American economic essays*, ed. A.W. Coats, Vol. 1. London: Routledge.

Coats, A.W. 1997. *The post-1945 internationalization of economics*. Durham: Duke University Press *History of Political Economy* 28 (Annual Supplement).

Coats, A.W. 2001. The AEA and the radical challenge to social science. In *Economics broadly considered: Essays in honor of Warren J. Samuels*, ed. J.E. Biddle, J.B. Davis, and S.G. Medema. London: Routledge.

Cochrane, J.L. 1979. *Industrialism and industrial man in retrospect: A critical review of the ford foundation's support for the inter-university study of labor*. Ann Arbor: Ford Foundation, distributed by University Microfilms International.

Colander, D. 1992. The sounds of silence: The profession's response to the COGEE report. *American Journal of Agricultural Economics* 80: 600–607.

Colander, D. 2005. The making of an economist redux. *Journal of Economic Perspectives* 19(1): 175–198.

Colander, D., and A. Klamer. 1987. The making of an economist. *Journal of Economic Perspectives* 12(3): 95–111.

Colander, D., and A. Klamer. 1990. *The making of an economist*. Boulder: Westview Press.

Colander, D., and H. Landreth. 1996. *The coming of Keynesianism to America: Conversations with the founders of Keynesian economics*. Cheltenham: Edward Elgar.

Colander, D., and H. Landreth. 1998. Political influence on the textbook Keynesian revolution: God, man and Lorie Tarshis at Yale. In *Keynesianism and the Keynesia revolution in America: A memorial volume in honour of Lorie Tarshis*, ed. O.F. Hamouda and B.B. Price. Cheltenham: Edward Elgar.

Debreu, G. 1959. *The theory of value*. London: Wiley.

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

Edwards, L. 2005. *The power of ideas: The heritage foundation at twenty-five years*. Ottawa: Jameson Books.

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

Goodwin, C.D.W. 1998. The patrons of economics in a time of transformation. In *From interwar pluralism to postwar neoclassicism*, ed. M. Morgan and M. Rutherford. Durham: Duke University Press *History of Political Economy 30* (Annual Supplement).

Gordon, D.F. 1965. The role of the history of economic thought in the understanding of modern economic theory. *American Economic Review* 55: 119–127.

Hands, D.W., and P. Mirowski. 1998. A paradox of budgets: The postwar stabilization of American neoclassical demand theory. In *From interwar pluralism to postwar neoclassicism*, ed. M. Morgan and M. Rutherford. Durham: Duke University Press *History of Political Economy 30* (Annual Supplement).

Hansen, A. 1953. *A guide to Keynes*. London: McGraw Hill.

Hansen, W.L. 1991. The education and training of economics doctorates: Major findings of the executive secretary of the American economic association's commission on graduate education in economics. *Journal of Economic Literature* 29: 1054–1087.

Harberger, A. 1997. Good economics comes to Latin America, 1955–95. In *The post-1945 internationalization of economics*, ed. A.W. Coats. Durham: Duke University Press *History of Political Economy* 28 (Annual Supplement).

Ingrao, B. and Israel, G. 1990. *The invisible hand: Economic equilibrium in the history of science.* Trans. I. McGilvray. Cambridge, MA: MIT Press.

Katz, B. 1989. *Foreign intelligence*. Cambridge, MA: Harvard University Press.

Klein, L.R. 1947. *The Keynesian revolution*. 2nd ed, 1968. London: Macmillan.

Koopmans, T.C. 1947. Measurement without theory. *Review of Economics and Statistics* 29(3): 161–172.

Krueger, A.O. 1991. Report of the commission on graduate education in economics. *Journal of Economic Literature* 29: 1035–1053.

Lee, F.S. 1984. The marginalist controversy and the demise of full-cost pricing. *Journal of Economic Issues* 18(4): 1107–1132.

Lee, F.S. 2000. On the genesis of post Keynesian economics: Alfred S. Eichner, Joan Robinson and the founding of post Keynesian economics. *Research in the History of Economic Thought and Methodology* 18-C: 3–258.

Lee, F.S. 2004. To be a heterodox economist: The contested landscape of American economics, 1960s and 1970s. *Journal of Economic Issues* 38: 747–763.

Lester, R.A. 1946. Shortcomings of marginal analysis for wage-employment policies. *American Economic Review* 36: 63–82.
Leonard, R. 1991. War as a 'simple economic problem': The rise of an economics of defense. In *Economics and national security: A history of their interaction*, ed. C.D. Goodwin. Durham: Duke University Press *History of Political Economy 23* (Annual Supplement).
Loureiro, M.R. 1997. The professional and political impacts of the internationalization of economics in Brazil. In *The post-1945 internationalization of economics*, ed. A.W. Coats. Durham: Duke University Press *History of Political Economy* 28 (Annual Supplement).
McCallum, B.T. 2000. Recent developments in monetary policy analysis: The roles of theory and evidence. In *Macroeconomics and the real world. volume 1: Econometric techniques and macroeconomics*, ed. R.E. Backhouse and A. Salanti. Oxford: Oxford University Press.
McMillan, J. 2002. *Reinventing the bazaar: A natural history of markets*. New York: W.W. Norton.
Medema, S.G. 2000. 'Related disciplines': The professionalization of public choice analysis. In *Toward a history of applied economics*, ed. R.E. Backhouse and J. Biddle. Durham: Duke University Press *History of Political Economy 32* (Annual Supplement).
Mehrling, P.G. 1998. *The money interest and the public interest: American monetary thought, 1920–1970*. Cambridge, MA: Harvard University Press.
Mirowski, P. 2002. *Machine dreams: Economics becomes a cyborg science*. Cambridge, MA: Cambridge University Press.
Mitchell, W.C. 1925. Quantitative analysis in economic theory. *American Economic Review* 15(1): 1–12.
Mongin, P. 1992. The 'full-cost' controversy of the 1940s and 1950s: A methodological assessment. *History of Political Economy* 24: 311–356.
Morgan, M. 1991. *A history of econometric ideas*. Cambridge, MA: Cambridge University Press.
Morgan, M., and M. Rutherford. 1998. *From interwar pluralism to postwar neoclassicism*. Durham: Duke University Press *History of Political Economy 30* (Annual Supplement).
Novick, D. 1954. Mathematics: Logic, quantity and method. *Review of Economics and Statistics* 36: 357–358.
Patinkin, D. 1956. *Money, interest and prices*. Evanston: Row, Peterson.
Ruggles, N., ed. 1970. *Economics*. Englewood Cliffs: Prentice Hall.
Ruggles, R. 1962. Relation of the undergraduate major to graduate economics. *American Economic Review* 52: 483–489.
Samuelson, P.A. 1947. *Foundations of economic analysis*. Cambridge, MA: Harvard University Press.
Samuelson, P.A. 1948. *Economics: An introductory analysis*. New York: McGraw-Hill.
Solow, R.M. 1997. How did economics get that way, and what way did it get? *Daedalus* 126: 39–58.
Valdes, J.G. 1995. *Pinochet's economists: The chicago school in chile*. Cambridge, MA: Cambridge University Press.
Vining, R. 1949. Koopmans on the choice of variables to be studied and of methods of measurement. *Review of Economics and Statistics* 31(2): 77–86.
Weintraub, E.R. 1998. Axiomatisches missverstaendniss. *Economic Journal* 108: 1837–1847.
Weintraub, E.R. 2002. *How economics became a mathematical science*. Durham: Duke University Press.

Uno, Kozo (1897–1977)

T. Sekine

Keywords

Marxian value analysis; Stages theory of development; Uno, K.; Use value

JEL Classifications

B31

A prominent Japanese Marxian economist known especially for his rigorous and systematic reformulation of Marx's *Capital*. Born in Kurashiki in western Japan in a year of intense social unrest, Uno early took an interest in anarcho-syndicalism and Marxism. Not being of an activist temperament, however, he strictly disciplined himself to remain, throughout his life, within the bounds of independent academic work. For this deliberate separation of theory (science) from practice (ideology) he was frequently criticized. After studying in Tokyo and Berlin in the early 1920s, Uno taught at Tohoku University (1924–38), the University of Tokyo (1947–58) and Hosei University (1958–68). During most of the war years he kept away from academic institutions. He authored many controversial books, especially after the war. His eleven-volume *Collected Works* were published by Iwanami-Shoten in 1973–4.

The problem with Marx's *Capital*, according to Uno, is that it mixes the theory and history of

capitalism in a haphazard fashion (described as 'chemical' by Schumpeter) without cogently establishing their interrelation. Uno's methodological innovation lies in propounding a stages theory of capitalist development (referring to the stages of mercantilism, liberalism, and imperialism) and using it as a mediation between the two.

Capitalism is a global market economy in which all socially needed commodities tend to be produced as value (that is, indifferently to their use-values) by capital. This tendency is never consummated since many use-values in fact fail to conform to this requirement. Only in theory, which synthesizes 'pure' capitalism, can one legitimately envision a complete triumph of value over use-values. The inevitable gap between history, in which use-values appear in their raw forms, and pure theory, in which they are already idealized as merely distinct objects for use, must be bridged by stages theory, which structures itself around use-values of given types (as 'wool', 'cotton', and 'steel' respectively typify the use-values of the three stages).

Uno's emphasis on 'pure' capitalism as the theoretical object has invited many uniformed criticisms. His synthesis of a purely capitalist society as a self-contained logical system follows the genuine tradition of the Hegelian dialectic, and is quite different from axiomatically contrived neoclassical 'pure' theory. Unlike the latter which takes the capitalist market for granted, Uno's theory logically generates it by step-by-step syntheses of the ever-present contradiction between value and usevalues. The pure theory of capitalism is thus divided into the three doctrines of circulation, production, and distribution according to the way in which this contradiction is settled. By specifically articulating the abiding dialectic of value and use-values, already present in *Capital*, Uno has given Marxian economic theory its most systematic formulation, a formulation which militates against the two commonest Marxist errors known as voluntarism and economism.

Uno's approach is not dissimilar to Karl Polanyi's in appreciating the tension between the substantive (use-value) and the formal (value) aspect of the capitalist economy. Unlike Polanyi, however, Uno ascribes more than relative importance to capitalism, in the full comprehension of which he sees the key to the clarification of both pre-capitalist and post-capitalist societies. Thus Uno's approach reaffirms and exemplifies the teaching of Hegel (and Marx) that one should 'learn the general through the particular', and not the other way round.

Bibliography

Albritton, R. 1984. The dialectic of capital: A Japanese contribution. *Capital and Class* 22: 157–176.

Albritton, R. 1985. *A Japanese reconstruction of marxist theory*. London: Macmillan.

Itoh, M. 1980. *Value and crisis: Essays in marxian economics in Japan*. New York: Monthly Review Press.

Sekine, T.T. 1975. Uno-Riron: A Japanese contribution to Marxian political economy. *Journal of Economic Literature* 13: 847–877.

Sekine, T.T. 1984. *The dialectic of capital: A study of the inner logic of capitalism*. Tokyo: Toshindo Press.

Uno, K. 1980. *Principles of political economy: Theory of a purely capitalist society*. Trans. from the Japanese by T. T., Sekine. Brighton: Harvester Press.

Urban Agglomeration

William C. Strange

Abstract

Urban agglomeration is the spatial concentration of economic activity in cities. It can also take the form of concentration in industry clusters or in employment centres within a city. One reason that agglomeration takes place is that there exist external increasing returns, also known as agglomeration economies. Evidence indicates that there exist both urbanization economies, associated with city size, and localization economies, associated with the clustering of industry. Both effects attenuate geographically. Theoretical research has identified many sources of agglomeration economies, including labour market pooling, input sharing, and knowledge spillovers. Empirical

research has offered evidence consistent with each of these.

Keywords

Input sharing; Knowledge spillovers; Labour market pooling; Localization economies; Migration; New economic geography; Production functions; Productivity; Rent seeking; Systems of cities; Urban agglomeration; Urban wage premium; Urbanization economies

JEL Classification

R12

Urban agglomeration is the spatial concentration of economic activity in cities. It can also take the form of concentration in industry clusters or in centres of employment within a city.

That both kinds of concentration exist is not debatable. Cities contain roughly 80% of the US population, and urban population densities are approximately four times the national average. It is not just aggregate activity that is agglomerated; individual industries are concentrated too. There are many examples. Computer software is well-known for its spatial concentration, especially in the Silicon Valley. Automobile manufacturing, finance, business services, and the production of films and television programmes are other notable examples of industrial clustering. Agglomeration also takes place within cities in the form of densely developed downtowns and sub-centres. These patterns are not unique to the United States. Capital and labour are highly agglomerated in every developed country, and they are increasingly agglomerated in the developing world.

Localization and Urbanization Economies

There are two sorts of agglomeration economies. Urbanization economies are associated with city size or diversity. Localization economies are associated with the concentration of particular industries. The idea that a city's size or diversity contributes to agglomeration economies is often attributed to Jacobs (1969), while the idea that industrial localization increases productivity goes back to Marshall (1890).

Economists have looked for evidence of these effects in a number of ways. Since agglomeration economies by definition enhance productivity, one natural approach is to estimate a production function. Estimating the production function requires establishment level data on inputs, including employment, land, capital, and materials. Data on labour is the easiest to obtain, although even the most detailed data-sets are incomplete, for instance by omitting experience in a particular occupation. Although data on purchased materials are sometimes available, data on internally sourced inputs typically are not. Measuring a firm's capital presents serious problems, including accounting for depreciation. Finally, even with good input data it is necessary to control for endogeneity of input use.

Despite the difficulties inherent in estimating a production function, a substantial body of research has estimated the impact of agglomeration on productivity. The very rough consensus is that doubling city size increases productivity by an amount that ranges from 2% to 8%. Some estimates are lower. The diversity of the local environment, another aspect of urbanization, has also been shown to be positively related to productivity. In addition, there is evidence of localization economies of roughly similar magnitude (see Rosenthal and Strange 2004, for a review).

There are other ways to look for evidence of localization and urbanization economies. Glaeser and Mare (2001) identify the existence of an urban wage premium, with workers in cities of over a million residents earning roughly a third more in nominal wages than workers in cities of fewer than 100,000. Even after controlling for the selection of highly productive workers into cities, a significant premium remains. This is evidence of agglomeration economies because firms would not be willing to pay such premium in the absence of a corresponding productivity advantage. Rosenthal and Strange (2003) consider the arrival of new business establishments and find that diversity encourages arrivals and that

localization economies are more important than urbanization economies for the industries examined. Finally, Henderson et al. (1995) analyse employment growth in the United States during the 1970–87 period. For mature industries, the specialization of employment at the metropolitan level is positively associated with growth. These papers are fairly typical of those that have measured the existence of agglomeration economies. The broad conclusion is that both urbanization and localization economies are present.

The Sources of Agglomeration Economies: Why Do Cities and Clusters Exist?

Marshall (1890) identifies three forces that can explain industry clustering: input sharing, labour market pooling, and knowledge spillovers. Input sharing exists when, for example, a clothing manufacturer in New York is able to purchase a great variety of relatively inexpensive buttons from a nearby company that specializes in button manufacturing. Correspondingly, the button manufacturer also benefits because there will be many nearby clothing manufacturers to whom it can sell buttons. The process is, therefore, a circular one. Labour market pooling exists when a film production company in the Los Angeles area can quickly fill a position by hiring one of the many specialized production workers already present locally. Similarly, a specialized worker in Los Angeles can more easily find a new position without having to relocate. In both instances, labour pooling reduces search costs and improves match quality, providing valuable benefits for employers and workers. Knowledge spillovers exist when industrial engineers can learn the tricks of the trade from random interactions with other programmers in the same location. Any of these forces can explain industry clustering. They can also give rise to cities. The sharing of business service inputs can, for example, lead firms in very different industries to benefit from locating in close proximity to each other. Similar sorts of stories can be told about labour market pooling and knowledge spillovers.

Marshall's list is, of course, incomplete. Many other forces can lead to agglomeration. First, there is greater availability of consumer amenities in large cities. A major league sports franchise, for instance, requires a significant fan base in order to be economical. Second, natural advantage can explain both urbanization and localization. For instance, heavy manufacturing has historically developed near sources of minerals and where water transportation was possible. Third, internal economies of scale coupled with transactions costs can lead to self-reinforcing agglomeration. This explanation is the heart of the New Economic Geography (NEG, Fujita et al. 1999), and it has received much attention in recent years.

Various approaches have been adopted in modelling agglomeration economies. Perhaps the simplest is to assume that there is some sort of public good that can be shared more economically in a larger city or cluster (Arnott 1979). This force operates on both the production and consumption sides. Productivity is enhanced by infrastructure, and utility is increasing in public goods. An alternative is to assume that there are local externalities, with agents directly making their neighbours better off (that is, more knowledgeable). Agglomeration economies can also arise from thick market effects in search or matching (Helsley and Strange 1990). The important common element from all these explanations is that agglomeration is associated with situations where market outcomes are not guaranteed to be efficient. Duranton and Puga (2004) provide a more detailed survey of the sharing, matching, and spillover micro-foundations of agglomeration.

The discussion thus far has suggested that agglomeration is always a positive outcome, at least as a second-best solution to market failures. This need not always be the case. Another reason to agglomerate is rent seeking. Ades and Glaeser (1995) show that there are many situations where urbanization can allow a city's residents to claim the output of other agents. They argue that imperial Rome supported a population of more than one million at least in part because the rewards of empire were distributed to residents of Rome as 'bread and circuses' in order to preserve domestic order. In this case, a city may exist,

not because it adds to productivity, but because it allows redistribution.

Empirical work has provided evidence of the presence of Marshall's forces. Jaffe et al. (1993) provide direct evidence, showing that patent citations are geographically localized. Holmes (1999) considers local input sharing. He shows that more concentrated industries have a higher value of purchased input intensity, equal to purchased inputs divided by sales. This is consistent with the presence of input sharing. Costa and Kahn (2001) consider one aspect of labour pooling: matching between workers and employers. The key result is that 'power couples', where both partners have at least a college degree, have disproportionately and increasingly located in large metropolitan areas. After considering other possible explanations, they conclude that power couples have become increasingly urbanized at least in part because it easier for both individuals to find good matches for their specialized skills. There is also evidence of the non-Marshallian agglomeration economies. Glaeser et al. (2001) provide evidence of consumption effects. Kim (1995) documents the importance of natural advantage. Evidence of effects predicted by NEG is reviewed by Head and Mayer (2004). Looking across industries at the sorts of industry characteristics associated with agglomeration, Rosenthal and Strange (2001) find that all of the factors discussed above contribute to industrial agglomeration. The evidence is strongest for labour market pooling.

The Geography of Agglomeration Economies: Cities and Neighbourhoods

It is common to consider agglomeration economies at the city level. This is because it is significantly easier to carry out estimation of production functions, wage premiums, births, or growth in that sort of aggregate analysis. In the previous section's analysis, however, it is clear that agglomeration economies depend on the distance between agents rather than on the political boundaries of cities. Thus, it makes sense to consider the degree to which agglomeration economies are at work at different levels of geography. Are they a neighbourhood effect or do they operate at the city level? In a sense, this question about the boundary of an agglomeration is parallel to asking about the boundary of a firm. The canonical question in the theory of the firm literature is: make or buy? Should an activity be carried out internally or through market transactions? The parallel agglomeration question is: near or far? Should one activity take place in close proximity to another or at a great distance?

Rosenthal and Strange (2003) address this issue by using geo-coding software to measure total employment and own-industry employment within a certain distance of an employer. Using these measures, the paper calculates the effects of the local environment on the number of firm births and on these new firms' employment levels for six industries (computer software, apparel, food processing, printing and publishing, machinery, and fabricated metals). The key result is that agglomeration economies attenuate with distance. The effect of additional employment beyond five miles is shown to be roughly one-quarter to one-half of the effect of additional employment within a firm's own zipcode (postal code). This result is consistent with both the concentration of employment downtown and in sub-centres.

Agglomeration Economies in a System of Cities or Regions

There has been considerable theoretical work on the implications of agglomeration economies for a system of cities or regions. Fujita and Thisse (2002) review and synthesize the literature. With apologies for oversimplification, the analysis in the literature proceeds as follows. First, it is assumed that there exists some sort of agglomeration economy. The specification may be of a reduced form shifting of the production function or of a particular agglomerative force. In the NEG literature (Fujita et al. 1999), for instance, agglomeration arises from backward linkages between firms and input suppliers and forward linkages between firms and consumers. Second, equilibrium in the system is characterized.

The equilibration always involves the individual location decisions taken by firms and households. It may also involve some coordination by a large agent, either a local government or a profit-maximizing city developer. Finally, the dynamic properties of the equilibrium are considered.

The systems of cities literature obtains several results. First, the characteristics of cities in an equilibrium system depend crucially on the sorts of agglomeration economies at work. If there are only localization economies, then the system will feature cities that specialize by industry. If there are in addition, or instead, urbanization economies, then diverse cities can arise. Second, there can be multiple equilibria. This means, for example, that there is no guarantee that the largest city in the middle of North America would be where Chicago is. Third, history matters. Agglomeration economies can be a conservative force in that they make it difficult for firms and workers to change their locations. Fourth, there is potential for catastrophic change, with a small change in parameters inducing a large change in outcomes. The attraction of other firms can cause an agglomeration to persist beyond the point at which it would have arisen from *de novo* location decisions. Eventually, however, when the attractiveness of other locations becomes sufficiently great, the agglomeration collapses suddenly. Fifth, equilibrium is not likely to be efficient. This result arises most starkly in a model where city sizes are determined solely by individual migration decisions. This ignores the existence of private developers and governments, which are both rewarded from realizing more efficient cities. However, in order to realize an efficient allocation, a developer would require unlimited control of city formation, a condition that is unlikely to obtain (Helsley and Strange 1997).

See Also

- ▶ New Economic Geography
- ▶ Spatial Economics
- ▶ Systems of Cities
- ▶ Urbanization

Bibliography

Ades, A., and E. Glaeser. 1995. Trade and circuses: Explaining urban giants. *Quarterly Journal of Economics* 110: 195–227.

Arnott, R. 1979. Optimal city size in a spatial economy. *Journal of Urban Economics* 61: 65–89.

Costa, D., and M. Kahn. 2001. Power couples. *Quarterly Journal of Economics* 116: 1287–1315.

Duranton, G., and D. Puga. 2004. Micro-foundations of urban agglomeration economies. In *Handbook of urban and regional economics*, vol. 4, ed. J. Henderson and J.-F. Thisse, 2063–2118. Amsterdam: North-Holland.

Fujita, M., and J. Thisse. 2002. *The economics of agglomeration*. Cambridge: Cambridge University Press.

Fujita, M., P. Krugman, and A. Venables. 1999. *The spatial economy: Cities, regions, and international trade*. Cambridge: Cambridge University Press.

Glaeser, E., and D. Mare. 2001. Cities and skills. *Journal of Labor Economics* 192: 316–342.

Glaeser, E., J. Kolko, and A. Saiz. 2001. Consumer city. *Journal of Economic Geography* 1: 27–50.

Head, K., and T. Mayer. 2004. The empirics of agglomeration and trade. In *Handbook of urban and regional economics*, vol. 4, ed. J. Henderson and J.-F. Thisse, 2609–2670. Amsterdam: Elsevier.

Helsley, R., and W. Strange. 1990. Agglomeration economies and matching in a system of cities. *Regional Science and Urban Economics* 20: 189–212.

Helsley, R., and W. Strange. 1997. Limited developers. *Canadian Journal of Economics* 30: 329–348.

Henderson, J., A. Kuncoro, and M. Turner. 1995. Industrial development in cities. *Journal of Political Economy* 103: 1067–1085.

Holmes, T. 1999. Localization of industry and vertical disintegration. *Review of Economics and Statistics* 81: 314–325.

Jacobs, J. 1969. *The economy of cities*. New York: Vintage.

Jaffe, A., M. Trajtenberg, and R. Henderson. 1993. Geographic localization of knowledge spillovers as evidenced by patent citations. *Quarterly Journal of Economics* 108: 577–598.

Kim, S. 1995. Expansion of markets and the geographic distribution of economic activities: The trends in U.S. regional manufacturing structure, 1860–1987. *Quarterly Journal of Economics* 110: 881–908.

Marshall, A. 1890. *Principles of economics*. London: Macmillan.

Rosenthal, S., and W. Strange. 2001. The determinants of agglomeration. *Journal of Urban Economics* 50: 191–229.

Rosenthal, S., and W. Strange. 2003. Geography, industrial organization, and agglomeration. *Review of Economics and Statistics* 85: 377–393.

Rosenthal, S., and W. Strange. 2004. Evidence on the nature and sources of agglomeration economies. In *Handbook of urban and regional economics*, vol. 4, ed. J. Henderson and J.-F. Thisse, 2119–2172. Amsterdam: Elsevier.

Urban Economics

John M. Quigley

Abstract

Urban economics emphasizes: the spatial arrangements of households, firms, and capital in metropolitan areas; the externalities which arise from the proximity of households and land uses; and the public policy issues which arise from the interplay of these economic forces.

Keywords

Cities; Congestion; Density gradient; Diversity (ethnic and racial); Endogenous growth; External economies; Housing markets; Human capital; Labour markets; Land markets; Land use; Monopolistic competition; Patents; Pollution; Price discrimination; Property taxation; Regional economics; Residential segregation; Road pricing; Schelling, T; Social externalities; Spatial competition; Suburbanization; Tipping point models; Transport costs; Urban agglomeration; Urban consumption externalities; Urban economics; Urban production externalities; von Thünen, J; Zoning

JEL Classifications

R0

Cities exist because production or consumption advantages arise from higher densities and spatially concentrated location. After all, spatial competition forces firms and consumers to pay higher land rents – rents that they would not be willing to pay if spatially concentrated economic activity did not yield cost savings or utility gains. Economists have long studied the forces leading to these proximities in location, focusing first and foremost upon the importance of transport costs.

Early theorists (for example, von Thünen, as early as 1826; see Hall 1966) considered land use and densities in an agrarian town where crops were shipped to a central market. Early models of location deduced that land closer to the market would be devoted to producing crops with higher transport costs and higher output per acre. Cities in the 19th century at this time were characterized by high transport costs for both goods and people, and manufactured goods were produced in close proximity to a central node – a port or a railway from which goods could be exported to world markets. The high costs of transporting people also meant that workers' residences were located close to employment sites.

Transport improvements in the late 19th century meant that urban workers could commute cheaply by streetcar, thereby facilitating the suburbanization of population into areas surrounding the central worksite. More radical technical change in the first decades of the 20th century greatly reduced the cost of transporting materials and finished goods. The substitution of the truck for the horse and wagon finally freed production from locations adjacent to the export node. The introduction of the private auto a decade later further spurred the decentralization of US metropolitan areas.

Spatial Forces

The seminal literature in urban economics provides positive models of the competitive forces and transport conditions which give rise to the spatial structure of modern cities. These models emphasize the trade-off between the transport costs of workers, the housing prices they face, and the housing expenditures they choose to make. Relatively simple models can explain the basic features of city structure – for example, the gradient in land prices with distance to the urban core; the house price gradient; the relationship between land and housing price gradients; the intensity of land use; and the spatial distribution of households by income (see Brueckner 1987, for a review).

Empirical investigations of these phenomena reveal clearly that these gradients have been decreasing over time. Indeed, the flattening of price and density gradients over time has been

observed in the United States since as long ago as the 1880s. (Early work is reported in Mills 1972.) In interpreting these trends, it is important to sort out the underlying causes. The stylized model described above emphasizes the roles of transport cost declines (in part, as a result of technical change and the role of the private auto), increases in household income, and population growth in explaining suburbanization. These models also rely upon the stylized fact that the income elasticity of housing demand exceeds the income elasticity of marginal transport costs. The alternative, largely ad hoc, explanations stress specific causes, for example the importance of tax policies which subsidize low-density owner- occupied housing, the importance of neighbourhood externalities which vary between cities and suburbs, or the role of variations in the provision of local public goods. There is a variety of empirical analyses of the determinants of the variations in density gradients over time and space. A general finding is that levels and intertemporal variations in real incomes and transport costs are sufficient to explain a great deal of the observed patterns of suburbanization.

Durable Capital

But, of course, variations in many of these other factors are highly correlated with secularly rising incomes and declining commuting costs, so any parcelling out of root causes is problematic. The elegant and parsimonious models of urban form have proven easy to generalize in some dimensions – for example, to incorporate stylized external effects and variations in income distributions across urban areas. It has proven to be substantially harder to recognize the durability of capital in tractable equilibrium models. The original models assumed that residential capital is infinitely malleable, and that variations in income or transport costs would be manifest in the capital intensity of housing over space in a direct and immediate way. The decline in land rents with distance from the urban centre means that developers' choices of inputs vary with capital-to-land ratios – declining with distance to the core. Dwellings are small near the urban core and large at the suburban fringe. Tall buildings are constructed near the urban centre, and more compact buildings are constructed in peripheral areas. But, of course, these structures and housing units are extremely durable, with useful lives exceeding 40 years or more. Thus, insights derived from the perspective in which the capital stock adjusts instantly to its long-run equilibrium position in response to changed economic conditions are limited.

Incorporating durable housing into models of residential location and urban form implies some recognition of the fact that 'history matters' in the structure and form of urban areas. Cities with the same distribution of income and demographics and with identical transport technologies may be quite different in their spatial structures, depending upon their historical patterns of development. Extensions of these simple models analyse the form of urban areas when developers have myopic or perfect foresight and when development is irreversible. With myopic developers, land is developed at each distance from the centre to the same density as it would have been built with malleable capital, but, once built, its capital intensity is frozen. Thus, with increasing opportunity costs of land over time, population and structural densities may increase with distance from the urban core.

With perfect foresight, the developer maximizes the present value of urban rents per acre, which vary with the timing of urban development. The present value of a parcel today is its opportunity cost in 'agriculture' until development plus its market value after conversion (minus construction costs). With perfect foresight, developers choose the timing of the conversion of land to urban use as a function of distance to the urban core, and development proceeds in an orderly fashion over time. Locations are developed according to their distance from the centre.

Of course, durable residential capital also implies that structures may depreciate or become obsolete. In particular, a historical pattern of development along concentric rings from the urban core, together with rising incomes, means that the most depreciated and obsolete dwellings

are the more centrally located. But embedded in each of these parcels of real estate is the option to redevelop it in some other configuration. Obsolete and depreciated dwellings commanding low prices are those for which the option to exercise redevelopment is less costly.

Models of development with perfect foresight in which residential capital depreciates imply that the timing of initial redevelopment of residential parcels depends only on their distance from the urban core (since that indexes their vintage of development). These models imply that the capital intensity of land use does not exhibit the smooth and continuous decline with distance from the core. Capital intensity does decline with distance, on average, but the relationship is not monotonic.

With uncertainty, developers take into account their imperfect knowledge of future prices in making land use decisions today. But this means that developers may make mistakes by developing land too soon. As a consequence, land development may often proceed in a leapfrog pattern. Landowners may withhold some interior land from development in anticipation of higher rents and profitable development later on (see Capozza and Helsley 1990, for a unified treatment).

The key point in these modern models of urban form which incorporate durable residential capital is that the timing as well as the location of development affect the choices made by housing suppliers. History 'matters' in these models, just as it does in the decisions of housing suppliers in urban areas.

Externalities

Theory

Recent work has greatly extended these urban models to address explicitly the production and consumption externalities which give rise to cities. The basic models combine Marshallian notions of 'economics of localized industry' and Jacobs's (1969) notions of 'urbanization economies' with the perspective on monopolistic competition and product diversity introduced by Dixit and Stiglitz (1977).

On the consumption side, the general form of these models assumes that household utility depends on consumption of traded goods, housing, and the variety of local goods. The markets for traded goods and housing are competitive, while the differentiated local goods are sold in a monopolistically competitive market. If there is less differentiation among local goods, then variety loses its impact on utility; greater differentiation means that variety has a greater effect on utility. Under reasonable assumptions, the utility of a household in the city will be positively related to the aggregate quantity of local goods it consumes and the *number* of types of these goods which are available in the economy (see Quigley 2001, for examples).

On the production side of the economy, the importance of a variety of locally produced inputs can be represented in a parallel fashion. For example, suppose that the aggregate production function includes labour, space and a set of specialized inputs. Again, the markets for labour and space can be taken as competitive, while the differentiated local inputs are purchased in a monopolistically competitive market. If there is less differentiation among inputs, then variety loses its impact on output; greater differentiation means that variety has a greater effect on output. For example, a general counsel may operate alone. However, she may be more productive if assisted by a general practice law firm, and even better served by firms specializing in contracts, regulation and mergers. Again, under reasonable conditions, output in the city will be related to quantities of labour, space, and specialized inputs utilized and also to the number of different producer inputs available in that city.

The theoretical models built along these lines yield a remarkable conclusion: diversity and variety in consumer goods or in producer inputs can yield external scale economies, even though all individual competitors and firms earn normal profits. In these models, the size of the city and its labour force will determine the number of specialized local consumer goods and the number of specialized producer inputs, given the degree of substitutability among the specialized local goods in consumption and among specialized inputs in production. A larger city will have a greater

variety of consumer products and producer inputs. Since the greater variety adds to utility and to output, in these models larger cities are more productive, and the well-being of those living in cities increases with their size. This will hold true even though the competitive and monopolistically competitive firms in these models each earn a normal rate of profit (see Fujita and Thisse 2001, for a comprehensive treatment).

Applications: Pollution and Transport

As emphasized above, however, the advantages of urban production and consumption are limited. Explicit recognition of the land and housing markets and the necessity of commuting suggests that, at some point, the increased costs of larger cities – higher rents arising from the competition for space, and higher commuting costs to more distant residences – will offset the production and consumption advantages of diversity. Other costs like air and noise pollution no doubt increase with size as well. Nevertheless, even when these costs are considered in a more general model, the optimal city size will be larger when the effects of diversity in production and consumption are properly reckoned. Urban output will be larger and productivity will also be greater (see Quigley 1998).

The empirical evidence assembled to support and test these theoretical insights about the regional economy is potentially very valuable. Hitherto, much of the discussion about the sources of economic growth was framed at that national level, and most of the aggregate empirical evidence – time series data across a sample of countries – was inherently difficult to interpret. By framing these theoretical propositions at the level of the region, it is possible to investigate empirically the sources of endogenous economic growth by using much richer bodies of data within a common set of national institutions. Geographical considerations of labour market matching and efficiency (Helsley and Strange 1990), of the concentration of human capital (Rauch 1993), and of patent activity (Jaffe et al. 1993) have all been studied at the metropolitan and regional levels, and considerable effort is under way to use regional economic data to identify and measure more fully the sources of American economic growth. These are major research activities exploring urban externalities in urban economies throughout the developed world. This research programme is still in its infancy.

Of course, specialization, diversity and agglomeration are not the only externalities arising in cities. High densities and close contact over space reinforce the importance of many externalities in modern cities. Among the most salient are the external effects of urban transport – congestion and pollution. Most work trips in urban areas are undertaken by private auto. (Indeed, in 2000, less than four per cent of commuting was by public transit; see Small and Gomez-Ibanez 1999.) In most US cities, automobiles are the dominant technology for commuting from dispersed origins to concentrated worksites. This technology is even more efficient for commuting from dispersed residences to dispersed worksites in metropolitan areas. Since commuting is concentrated in morning and evening hours, roads may be congested during peak periods, and idle during off-peak periods. Road users pay the average costs of travel when they commute during peak periods. They take into account the out-of-pocket and time costs of work trips, and in this sense commuters consider the average level of congestion in their trip-making behaviour. But commuters cannot be expected to account for the incremental congestion costs their travel imposes on other commuters. This divergence between the marginal costs of commuting and the average costs of commuting may be large during peak periods on arterial roadways.

The imposition of congestion tolls, increasing the average costs paid by commuters to approximate the marginal costs they impose on others, would clearly improve resource allocation. In the absence of efficient road pricing, the rent gradients in metropolitan areas are flatter, and the patterns of residential location are more centralized than they would otherwise be. Land markets are distorted and the market price of land close to the urban core is less than its social value.

The obstacles to improved efficiency are technological as well as political. Until recently, mechanisms for charging road prices were expensive and cumbersome. But modern technology (for example, transponders to record tolls

electronically) makes road pricing easy on bridges, tunnels and other bottlenecks to the central business district. Regular commuters affix a device to their autos, a device which can automatically debit the traveller's account. It would be a simple matter to vary these charges by time of day or intensity of road use and to make the schedule of these changes easily available to commuters. So far, at least in the United States, about the only form of price discrimination on bridges, tunnels and bottlenecks has been by vehicle occupancy, not by time of day and intensity of road use. It is surely possible to profit from the experience of other countries (such as Singapore) in pricing scarce roadways.

Political resistance is a major factor inhibiting the diffusion of road pricing. Typically, tolls are imposed in new facilities and the proceeds are pledged to retire debt incurred in construction. Paradoxically, tolls are thus imposed on new uncongested roads. Later on, when the roads become congested, the initial debt has been retired, and there is political support for removing the toll. (After all, 'the investment in the bridge has been repaid.') This is surely an instance where economics can better inform public policy.

Applications: Social, Spatial and Neighbourhood

Urban areas have always been characterized by social externalities as well. The close contact of diverse racial and ethnic groups in cities gives rise to much of the variety in products and services which enrich consumption. But the city is also characterized by the concentration of poverty and by the high levels of segregation by race and class.

The spatial concentration of households by income is, of course, predicted by the models of residential housing choice described above. A central question is the extent to which poverty concentrations give rise to externalities which disadvantage low-income households *relative* to their deprived circumstances in the absence of concentration. A great deal of qualitative research by other social scientists suggests that this is the case. Quite recent econometric research, however, suggests that this proposition is quite hard to demonstrate quantitatively by reliance on non-experimental data (see Manski 1995.) Nevertheless, the view that concentrations of disadvantaged households lead to more serious social consequences simply because of concentration is widely shared. For example, in low-wage labour markets most jobs are found through informal local contacts. If unemployed workers are spatially concentrated, it follows that informal contacts will produce fewer leads to employment.

Economic models of residential location also suggest that households will be segregated by race – to the extent that race and income are correlated. Yet research clearly indicates that the segregation of black households in urban areas is far greater than can be explained by differences in incomes and demographic characteristics.

Until quite recently, these patterns of segregation could be explained by explicitly discriminatory policies in the housing market. During the period of black migration to northern cities, rents were substantially higher in black neighbourhoods than in white neighbourhoods. As levels of migration tapered off in the 1970s, price differentials declined. The patterns of residence by race may be explicable by the tipping point models of Thomas Schelling (1971). In these models, there is a distribution of tolerance among the population, reflecting the maximum fraction of neighbours of a different race tolerated by any household. In this formulation, the race of each household provides an externality to all neighbouring households. It is easy to show that the racial integration of neighbourhoods may be impossible to achieve under many circumstances.

Despite this, there is widespread evidence of conscious discrimination in the markets for rental and owner-occupied housing (Ross and Yinger 2002), four decades after passage of the first Fair Housing legislation.

Racial segregation in housing markets may have particularly important welfare implications as jobs continue to suburbanize at a rapid rate. Racial barriers to opening up the suburbs for residence may lead to higher unemployment rates among minority workers (see Glaeser et al. 2004.)

The barriers to the integration of the suburbs by race and income are also related to the fiscal

externalities which are conferred by one category of residents upon another category. Most local tax structures emphasize *ad valorem* property taxes, and in most urban areas towns are free to vary property tax rates to finance locally chosen levels of public expenditure. If local tax revenues are proportional to house value, and if local public expenditures are proportional to the number of households served, local governments have strong incentives to increase the property value per household in their jurisdictions. To achieve this outcome, local governments may simply use zoning regulations to prohibit construction of housing appropriate to the budgets of lower-income households. The prohibition of high-density housing and multi-family construction, the imposition of minimum lot-size restrictions and the imposition of development fees can all be used as devices to increase property tax revenue per household. Importantly, these rules also increase the price of low-income housing. Many of these regulations can also be cloaked in terms of ecological balance and environmental protection. The inability of higher levels of government to achieve balance and equity in new residential development in US urban areas is quite costly.

Summary

The field of urban economics emphasizes the spatial arrangements of households, firms, and capital in metropolitan areas, the externalities which arise from the proximity of households and land uses, and the policy issues which arise from the interplay of these economic forces.

See Also

- ▶ Hearn, William Edward (1826–1888)
- ▶ Urban Production Externalities
- ▶ Urban Transportation Economics

Bibliography

Brueckner, J. 1987. The structure of urban equilibria: A unified treatment of the Muth–Mills Model. In *Handbook of regional and urban economics*, vol. 2, ed. E. Mills. Amsterdam: North-Holland.

Capozza, D., and R. Helsley. 1990. The stochastic city. *Journal of Urban Economics* 28: 187–203.

Dixit, A., and J. Stiglitz. 1977. Monopolistic competition and optimum product diversity. *American Economic Review* 67: 297–308.

Fujita, M., and J.-F. Thisse. 2001. *Economics of agglomeration*. Cambridge: Cambridge University Press.

Glaeser, E., E. Hanushek, and J. Quigley. 2004. Opportunities, race, and urban location: The legacy of John Kain. *Journal of Urban Economics* 56: 70–79.

Hall, P. (ed.). 1966. *Von Thünen's isolated state.* Trans. C. Wartenberg. Oxford: Pergamon Press.

Helsley, R., and W. Strange. 1990. Matching and agglomeration economies in a system of cities. *Regional Science and Urban Economics* 20: 189–212.

Jacobs, Jane. 1969. *The economy of cities*. New York: Random House.

Jaffe, A., M. Trajtenberg, and R. Henderson. 1993. Geographic localization of knowledge spillovers as evidenced by patent citations. *Quarterly Journal of Economics* 108: 577–598.

Manski, C. 1995. *Identification problems in the social sciences*. Cambridge, MA: Harvard University Press.

Mills, E. 1972. *Studies in the structure of urban economy.* Baltimore: John Hopkins Press.

Quigley, J. 1998. Urban diversity and economic growth. *Journal of Economic Perspectives* 12(2): 127–138.

Quigley, J. 2001. The renaissance in regional research. *Annals of Regional Science* 35: 167–178.

Rauch, J. 1993. Productivity gains from geographic concentration of human capital. *Journal of Urban Economics* 34: 380–400.

Ross, S., and J. Yinger. 2002. *The color of credit*. Cambridge, MA: MIT Press.

Schelling, T. 1971. Dynamic models of segregation. *Journal of Mathematical Sociology* 1: 143–186.

Small, K., and J. Gomez-Ibanez. 1999. Urban transportation. In *Handbook of regional and urban economics*, vol. 3, ed. Paul C. Cheshire and Edwin S. Mills. Amsterdam: North-Holland.

Urban Environment and Quality of Life

Matthew E. Kahn

Abstract

Urban quality of life is a key determinant of where the educated choose to live and work. Recognizing the importance of attracting the

high-skilled, many cities are investing to transform themselves into 'consumer cities'. This essay examines the supply and demand for non-market urban quality of life. It provides an overview of hedonic methods often used to estimate how much households pay for non-market urban attributes such as temperate climate and clean air.

Keywords

Compensating differentials; Congestion; Crime; Hedonic functions; Internal migration; Pollution; Population density; Quality of life; Urban environment and quality of life; Urban growth; Urbanization

JEL Classifications

R13

Soon a majority of the world's population will live in cities. In 1950, 30 per cent of the world's population lived in cities. By 2000, this fraction had grown to 47 per cent. It is predicted to rise to 60 per cent by 2030 (United Nations Population Division 2004). While the popular media focus on the growth of 'mega-cities', much urbanization occurs through the development of new cities and the growth of smaller metro areas (Henderson and Wang 2004).

People have migrated to cities in pursuit of better economic opportunities than are available in the countryside (Harris and Todaro 1970). In the past, urbanization has been viewed as representing a trade-off. Urban workers earned higher wages than rural residents but suffered from a lower quality of life. In the 1880s, the average urbanite in the United States had a life expectancy ten years lower than that of the average rural resident (Haines 2001). Frederick Engels bemoaned the density and the squalor in Britain's manufacturing cities in the mid-19th century. Urban historians have provided indelible descriptions of US cities in the past. In the 19th century, dead horses littered the streets of New York City, and thousands of tenement-dwellers were exposed to stinking water, smoky skies, and ear-shattering din (Melosi 1982, 2001). During the 19th and early 20th centuries, the skies above such major cities as Chicago and Pittsburgh were dark with smoke from steel smelters, heavy industrial plants, and burning coal.

Since the early 20th century, many major cities in the developed world have experienced sharp improvements in quality of life. By 1940, the urban mortality premium had vanished (Haines 2001). Starting in the early 1970s, air pollution, water pollution and noise pollution have sharply fallen in many major US cities. While there are several causes of this progress, ranging from effective regulation to industrial transition from manufacturing to services and technological advance, the net result of this trend is that past 'production cities' are transforming themselves into 'consumer cities'.

Cities that have high quality of life will have greater success at attracting the footloose highly educated to live there. Empirical studies have documented that a location's stock of educated people plays an important role in generating urban growth (Glaeser et al. 1995).

First, I sketch how a city 'produces' quality of life. I then discuss the demand for urban quality of life using a household production function framework. While urban quality of life is a valued 'commodity', there are no explicit markets where it can be purchased. Utility maximizing households face a trade-off in choosing where to locate. In cities with higher quality of life, home prices are higher. Measuring this price premium for quality continues to be a major focus of much environmental and urban empirical research.

The Supply of Urban Quality of Life

Each city can be thought of as a differentiated product. Its attributes include some exogenous factors such as climate and risk of natural disasters, and endogenous factors such as average commuting times, pollution and crime. Some of these endogenous attributes are by-products of economic activity. A city of 10,000 bike-riding lawyers would have much cleaner air than another city with 500,000 old-car-driving steel workers. None of the steel workers driving old cars intends

to pollute local air. Pollution represents an unintended consequence of their daily commuting mode and of local industrial production. This example highlights the importance of scale, composition and technique effects in determining local environmental quality. In the above example, scale refers to whether the city has 10,000 people or 500,000 people. Composition effects focus on consumption patterns (such as bike versus car) and industrial patterns (such as law firms versus steel plants). If one controls for a city's scale and composition, urban environmental quality can be high due to technique effects brought about by government regulation or the free market designing new capital with low emissions (for example, hybrid cars such as the Toyota Prius).

Early research in urban economics emphasized scale effects such that the biggest cities suffered more quality-of-life degradation as they expanded (Tolley 1974). Anyone can migrate to a big city without paying an 'entry fee'. When an extra person moves to a big city from a smaller city, this migration causes net social damage (due to extra congestion and pollution). Migrants will ignore the fact that their choice degrades local public goods in the destination city, but a benevolent planner would not. In the absence of a big city entry fee, the big city grows beyond its efficient size.

Cross-city empirical research has documented that such urban challenges as crime, pollution and congestion are all greater in big cities than in smaller cities (Glaeser 1998; Henderson 2002). But this 'cost' of city bigness is declining over time. In the 1990s, crime fell fastest in the largest US cities (Levitt 2004). Ambient air pollution is improving in many major cities despite a continued increase in population (Glaeser and Kahn 2004). The suburbanization of employment in all major US metropolitan areas has meant that that population 'sprawl' has not increased commute times.

City size is not a sufficient statistic for determining a city's quality of life. Other relevant factors are the city's geography, industrial and demographic composition, and government policy. A city's geography determines its climate and its capacity for handling local pollution. Put simply, some cities have it and some cities don't. As Billy Graham once said, 'The San Francisco Bay Area is so beautiful, I hesitate to preach about heaven while I'm here.'

Cities differ in their ability to absorb growth without suffering urban quality-of-life degradation. World Bank researchers have recently documented the importance of climate and topological features of the city in determining how much air pollution is caused by economic growth (see Dasgupta et al. 2004). Windier cities and cities that receive more rainfall suffer less ambient pollution from a given amount of emissions.

The composition of city economic activity also plays a key role in determining the supply of quality of life. All else equal, a city that specializes in manufacturing relative to services will have a lower quality of life. Such a city will have greater levels of ambient particulate and sulphur dioxide pollution. Water pollution will be greater, and more hazardous waste sites will be created. The rise and decline of manufacturing in the US rust belt over the 20th century provides dramatic evidence documenting these effects (Kahn 1999). A similar 'natural-experiment' has played out as communism died. In major cities in the Czech Republic, Hungry and Poland air pollution improved in the 1990s because the phase out of energy subsidies contributed to the shutdown of communist era industrial plants (Kahn 2003). As major cities such as New York and London and Chicago have experienced an industrial transition from manufacturing to finance and services, more people work in the service and tourist industries, and these workers have a financial stake in keeping the city's quality of life high.

A city's demographics also play a role in determining its quality of life. A city filled with senior citizens will offer a different set of restaurants and cultural opportunities from a city filled with immigrants and young parents. If a city can attract the highly educated, then a virtuous circle can be set off. Since more highly educated people earn more income, this will attract better restaurants and other commercial amenities.

Government policy plays a role in determining a city's quality of life. Boston's Big Dig project has cost over US$14 billion and is intended to beautify Boston by submerging its ugly highways

connecting the city centre to the waterfront and increasing the supply of green parks. Successful Clean Air Act regulation has sharply reduced vehicle emissions in Los Angeles. Rudy Giuliani, Mayor of New York City, achieved wide acclaim for improved policing that some have argued contributed to the sharp decline in the city's crime rate in the 1990s.

The supply of urban quality of life varies across cities and within cities. Some variation such as proximity to a major park or body of water is exogenously determined, but public policy can also have differential effects on quality of life across a city's neighbourhoods. The Clean Air Act has reduced Los Angeles' smog by much more in inland Hispanic communities than along the Pacific Ocean (Kahn 2001). Economists are just starting to investigate the general equilibrium impacts of regulations that differentially improve urban quality of life in some parts of a city relative to other parts of the same city (Sieg et al. 2004). If the improvements in quality of life were unexpected, then homeowners in such areas will receive a windfall. Long-standing renters in communities that have experienced regulation-induced improvements in local public goods will pay higher rents and may no longer be able to afford to live in their old community.

Demand for Urban Quality of Life

The household production function approach offers a framework for modelling the demand for non-market local public goods such as climate, street safety and local environmental quality. A person gains utility from being healthy, safe and comfortable. To achieve these goals, one purchases market goods such as doctor visits, home alarm systems and home entertainment systems. In addition, this person might choose a city and a residential community within this city featuring a temperate climate, low smog levels and safe streets.

Each household must choose a city and a community within that city to live in.

Households that value quality of life face a trade-off in that each city represents a bundle of non-market attributes and economic opportunities. Some cities such as San Francisco are beautiful but home prices are very high. Other cities such as Houston offer warm winter weather and cheap housing but its residents face severe summer humidity. Market products can offset such city's disamenities. Before the advent and diffusion of cheap air conditioning, humid cities would feature much lower home prices to compensate households for summer humidity. The diffusion of the air conditioner has allowed households to enjoy the benefits of living in warmer cities such as Houston during winter without suffering from humidity in summer (Rappaport 2003). This market product has increased the demand for living in humid cities.

Households may reveal different willingness to trade off non-market goods depending on the household's age, income and demographic circumstances. A household with children may place greater weight on communities with good schools. Households may differ in their demand for urban attributes. Asthmatics will avoid highly polluted cities and skiers will not mind the cold New England winters. Household demand may also hinge on idiosyncratic factors; for example, an individual who grew up in a specific city may want to remain living near his childhood friends.

The Hedonic Equilibrium Approach for Valuing Urban Quality of Life

The theory of compensating differentials says that it will be more costly to live in 'nicer' cities (Rosen 2002). This theory is really a 'no arbitrage' result. If migration costs are low across urban areas and if potential buyers are fully informed about the differences in non-market urban attributes bundles, then real estate prices will adjust such that homes in cities with higher quality of life will sell for a premium.

An enormous empirical literature has estimated cross-city and within-city hedonic price functions to estimate the implicit compensating differentials for non-market goods. In these studies, the dependent variable is the price of home i in city j in community m in year t. Define X_{it} as home

i's physical attributes in year *t*. A_{jt} represents city *j*'s attributes in year *t* and A_{mjt} represents the attributes of community m located in city *j* in the year *t*. Given this notation, a standard real estate hedonic regression will take the form:

$$\text{Price}_{ijmt} = \beta_0 + \beta_1^* X_{it} + \beta_2^* A_{jt} + \beta_3^* A_{mjt} + \varepsilon_{ijmt} \quad (1)$$

Multivariate regression estimates of this regression yield estimates of the compensating differentials for city level local public goods (based on β_2) and community-level local public goods (based on β_3). These coefficients represent the marginal implicit prices for small increases in the consumption of local public goods. Studies that control for a vector of local public goods are able to pinpoint the relative importance of different features of cities and communities ranging from climate to air pollution to urban crime. Equation (1) highlights the fact that households face a rich set of choices both across cities and across communities within the same city.

Environmental studies have used this hedonic framework to estimate compensating differentials for a myriad of different environmental local characteristics. For example, Costa and Kahn (2003) examine the compensating differential for living in nice climate in 1970 and in the year 1990. In 1970, a person would have to pay $1,288 (1990 dollars) in higher home prices per year to purchase San Francisco's climate over Chicago's climate. In 1990, this yearly price differential increased by $6,259 (1990 dollars) to $7,547. Chay and Greenstone (2005) use 1980 and 1990 data for all US counties find that a ten per cent reduction in ambient total suspended particulates increased home prices by three per cent. While much of the urban quality of life literature has focused on US city data, a promising research trend is examining international evidence.

Conclusion

Urban economic development policymakers have pursued very different growth strategies. Some cities subsidize sports stadiums while others build airports or downtown cultural centres. Such targeted investment is unlikely to yield the key urban anchor. This essay has argued that cities than can provide and enhance urban quality of life will attract the high-skilled. An end result of attracting this group is a more vibrant, diversified local economy. As per-capita incomes continue to rise, the demand for living and working in high quality-of-life cities will increase. The empirical literature continues to inquire into what the key components of quality of life are.

See Also

- City and Economic Development
- Environmental Kuznets Curve
- Urban Agglomeration
- Urban Growth
- Urbanization

Bibliography

Chay, K., and M. Greenstone. 2005. Does air quality matter? Evidence from the housing market. *Journal of Political Economy* 113: 376–424.

Costa, D., and M. Kahn. 2003. The rising price of non-market goods. *American Economic Review* 93: 227–232.

Dasgupta, S., Hamilton, K., Pandey, K. and Wheeler, D. 2004. Air pollution during growth: Accounting for governance and vulnerability. World Bank Policy Research Working Paper 3383. Washington, DC: World Bank.

Glaeser, E. 1998. Are cities dying? *Journal of Economic Perspectives* 12(2): 139–160.

Glaeser, E., and M. Kahn. 2004. Sprawl and urban growth. In *Handbook of urban economics*, ed. V. Henderson and J. Thisse, Vol. 4. Amsterdam: North Holland.

Glaeser, E., J. Scheinkman, and A. Shleifer. 1995. Economic growth in a cross-section of cities. *Journal of Monetary Economics* 36: 117–143.

Haines, M. 2001. The urban mortality transition in the United States 1800–1940. Historical Paper 134. Cambridge, MA: NBER.

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

Henderson, V. 2002. Urban primacy, external costs, and quality of life. *Resource and Energy Economics* 24(1/2): 95–106.

Henderson, V. and Hyoung Gun Wang 2004. Urbanization and city growth. Online. Available at http://www.econ.

brown.edu/faculty/henderson/papers.html. Accessed 29 June 2005.
Kahn, M. 1999. The silver lining of rust belt manufacturing decline. *Journal of Urban Economics* 46: 360–376.
Kahn, M. 2001. The beneficiaries of Clean Air Act legislation. *Regulation* 24: 34–39.
Kahn, M. 2003. New evidence on Eastern Europe's pollution progress. *Topics in Economic Analysis & Policy* 3(1), article 4. Online. Available at http://www.bepress.com/bejeap/topics/vol3/iss1/art4. Accessed 28 June 2005.
Levitt, S. 2004. Understanding why crime fell in the 1990s: Four factors that explain the decline and six that do not. *Journal of Economic Perspectives* 18(1): 167–190.
Melosi, M. 1982. *Garbage in the cities: Refuse, reform and the environment: 1880–1980*. College Station: Texas A&M Press.
Melosi, M. 2001. *Effluent america: Cities, industry, energy and the environment*. Pittsburgh: University of Pittsburgh Press.
Rappaport, J. 2003. Moving to nice weather. Research Working Paper 03–07. Kansas: Federal Reserve Bank of Kansas City.
Rosen, S. 2002. Markets and diversity. *American Economic Review* 92: 1–15.
Sieg, H., V. Smith, H. Banzhaf, and R. Walsh. 2004. Estimating the general equilibrium benefits of large changes in spatially delineated public goods. *International Economic Review* 45: 1047–1077.
Tolley, G. 1974. The welfare economics of city bigness. *Journal of Urban Economics* 1: 324–345.
United Nations Population Division. 2004. *World Population Prospects: The 2004 Revision Population Database*. Online. Available at Accessed 28 June 2005.

Urban Growth

Yannis M. Ioannides and Esteban Rossi-Hansberg

Abstract

'Urban growth' refers to the process of growth and decline of economic agglomerations. The pattern of concentration of economic activity and its evolution have been found to be an important determinant, and in some cases the result, of urbanization, the structure of cities, the organization of economic activity, and national economic growth. The size distribution of cities is the result of the patterns of urbanization, which result in city growth and city creation. The evolution of the size distribution of cities is in turn closely linked to national economic growth.

Keywords

Agricultural development; Cities; Congestion; Demography; Economic growth; Neoclassical model; Industrial revolution; Internal migration; Knowledge spillovers; New economic geography; Population density; Production externalities; Quality ladder model of economic growth; Returns to scale; Rural growth; Spatial distribution; Symmetry breaking; Systems of cities; Urban agglomeration; Urban economic growth; Vs national economic growth; Urban economics; Urban growth; Urbanization; Zipf"s Law

JEL Classifications

R11

Urban growth - the growth and decline of urban areas - as an economic phenomenon is inextricably linked with the process of urbanization. Urbanization itself has punctuated economic development. The spatial distribution of economic activity, measured in terms of population, output and income, is concentrated. The patterns of such concentrations and their relationship to measured economic and demographic variables constitute some of the most intriguing phenomena in urban economics. They have important implications for the economic role and size distribution of cities, the efficiency of production in an economy, and overall economic growth. As Paul Bairoch's magisterial work (1988) has established, increasingly concentrated population densities have been closely linked since the dawn of history with the development of agriculture and transportation. Yet, as economies move from those of traditional societies to their modern stage, the role of the urban sector changes from merely providing services to leading in innovation and serving as engines of growth.

Measurement of urban growth rests on the definition of 'urban area', which is not standard throughout the world and differs even within the

same country depending upon the nature of local jurisdictions and how they might have changed over time (this is true even for the United States). Legal boundaries might not indicate the areas covered by urban service-providers. Economic variables commonly used include population, area, employment, density or output measures, and occasionally several of them at once, not all of which are consistently available for all countries. Commuting patterns and density measures may be used to define metropolitan statistical areas in the USA as economic entities, but major urban agglomerations may involve a multitude of definitions.

The study of urban growth has proceeded in a number of different directions. One direction has emphasized historical aspects of urbanization. Massive population movements from rural to urban areas have fuelled urban growth throughout the world. Yet it is fair to say that economics has yet to achieve a thorough understanding of the intricate relationships between demographic transition, agricultural development and the forces underlying the Industrial Revolution. Innovations were clearly facilitated by urban concentrations and associated technological improvements. A related direction focuses on the physical structure of cities and how it may change as cities grow. It also focuses on how changes in commuting costs, as well as the industrial composition of national output and other technological changes, have affected the growth of cities. Another direction has focused on understanding the evolution of systems of cities - that is, how cities of different sizes interact, accommodate and share different functions as the economy develops and what the properties of the size distribution of urban areas are for economies at different stages of development. Do the properties of the system of cities and of city size distribution persist while national population is growing? Finally, there is a literature that studies the link between urban growth and economic growth. What restrictions does urban growth impose on economic growth? What economic functions are allocated to cities of different sizes in a growing economy? Of course, all of these lines of inquiry are closely related and none of them may be fully understood, theoretically and empirically, on its own. In what follows we address each in turn.

Urbanization and the Size Distribution of Cities

The concentration of population and economic activity in urban areas may increase either because agents migrate from rural to urban areas (urbanization) or because economies grow in term of both population and output, which results in urban as well as rural growth. Urban centres may not be sustained unless agricultural productivity has increased sufficiently to allow people to move away from the land and devote themselves to non-food producing activities. Such 'symmetry breaking' in the uniform distribution of economic activity is an important factor in understanding urban development (Papageorgiou and Smith 1983). Research on the process of urbanization spans the early modern era (the case of Europe having been most thoroughly studied; De Vries 1984) to recent studies that have applied modern tools to study urbanization in East Asia (Fujita et al. 2004). The 'New Economic Geography' literature has emphasized how an economy can become 'differentiated' into an industrialized core (urban sector) and an agricultural 'periphery' (Krugman 1991). That is, urban concentration is beneficial because the population benefits from the greater variety of goods produced (forward linkages) and may be sustained because a larger population in turn generates greater demand for those goods (backward linkages). This process exploits the increasing returns to scale that characterize goods production but does not always lead to concentration of economic activity. The range of different possibilities is explored extensively in Fujita et al. (1999). These ideas have generated new lines of research; see several related papers in Henderson and Thisse (2004).

The process of urban growth is closely related to the size distribution of cities. As the urban population grows, will it be accommodated in a large number of small cities, or in a small number

of large cities, or in a variety of city sizes? While cities have performed different functions in the course of economic development, a puzzling fact persists for a wide cross-section of countries and different time periods. The size distribution of cities is Pareto-distributed, is 'scale-free'. Gabaix (1999) established this relationship formally. He showed that, if city growth is scale independent (the mean and variance of city growth rates do not depend on city size: Gibrat' s Law) and the growth process has a reflective barrier at some level arbitrarily close to zero, the invariant distribution of city sizes is a Pareto distribution with coefficient arbitrarily close to 1: Zipf's Law. (Empirical evidence on the urban growth process as well as Zipf's Law is surveyed by Gabaix and Ioannides 2004.)

These results imply that the size distribution of cities and the process of urban growth are closely related. Eeckhout (2004) extends the empirical investigation by examining in depth all *urban places* in the United States and finds that the inclusion of the lower end of the sample leads to a log-normal size distribution. Duranton (2004) refines the theory by means of a quality-ladder model of economic growth that allows him to model the growth and decline of cities as cities win or lose industries following technological innovations. Ultimately, the movements of cities up and down the hierarchy balance out so as to produce a stable, skewed size distribution. This theory is sufficiently rich to accommodate subtle differences across countries (in particular the United States and France) that constitute systematic differences from Zipf's Law. Rossi-Hansberg and Wright (2004) use a neoclassical growth model that is also consistent with observed systematic deviations from Zipf's Law: in particular, the actual size distribution of cities shows fewer smaller and larger cities than the Pareto distribution, and the coefficient of the Pareto distribution has been found to be different from 1 although centred on it. They identify the standard deviation of the industry productivity shocks as the key factor behind these deviations from Zipf's Law. The evident similarity of the conclusions of those two papers clearly suggests that the literature is closer than ever before to resolving the Zipf's Law 'puzzle.'

Urban Growth and City Structure

Understanding urbanization and economic growth requires understanding the variety of factors that can affect city size and therefore its short-term dynamics. All of them lead to the basic forces that generate the real and pecuniary externalities that are exploited by urban agglomeration, on one hand, and congestion, which follows from agglomeration, on the other. Three basic types of agglomeration forces have been used, in different varieties, to explain the existence of urban agglomerations (all of them were initially proposed in Marshall 1920): (*a*) knowledge spillovers, that is, the more biomedical research there is in an urban area, the more productive a new research laboratory will be; (*b*) thick markets for specialized inputs: the more firms that hire specialized programmers, the larger the pool from which an additional firm can hire when the others may be laying off workers; and (*c*) backward and forward linkages. Local amenities and public goods can themselves be relevant agglomeration forces.

The size of urban agglomerations is the result of a trade-off between the relevant agglomeration and congestion forces. Urban growth can therefore be the result of any city- specific or economy-wide change that augments the strength or scope of agglomeration forces or reduces the importance of congestion forces. One example that has been widely used in the literature is reductions in commuting costs that lead to larger cities in terms of area, population, and in most models also output (Chatterjee and Carlino 1999). Another example is the adoption of information and communication technologies that may increase the geographical scope of production externalities, therefore increasing the size of cities.

Changes of underlying economic factors cause cities to grow or decline as they adjust to their new equilibrium sizes. Another more subtle factor is changes in the patterns of specialization that are

associated with equilibrium city sizes. That is, the coexistence of dirty industry with high-tech industry generates too much congestion, and therefore cities specialize in one or the other industry. Adjustments in city sizes and patterns of specialization in turn may be slow, since urban infrastructure, as well as business structures and housing are durable, and new construction takes time (Glaeser and Gyourko 2005). However, this type of change lead only to transitional urban growth, as city growth or decline eventually dies out in the absence of other city-specific or economy-wide shocks. Even when any of the economy-wide variables, such as population, grows continuously, the growth rate of a specific city will dwindle because of new city creation (Ioannides 1994; Rossi-Hansberg and Wright 2004).

Much attention has also been devoted to the effect that this type of urban growth has on urban structure. Lower commuting costs may eliminate the link between housing location choices and workplace location. This results in more concentration of business areas, increased productivity because of, say, knowledge spillovers, and lower housing costs in the periphery of the city. Urban growth can therefore lead to suburbanization as well as multiple business centres, as in Fujita and Ogawa (1982) or Lucas and Rossi-Hansberg (2002). Those phenomena become increasingly important because of the decline in transport and commuting costs brought about by the automobile along with public infrastructure investments. In other words, urban growth is associated with sprawl (Anas et al. 1998).

Urban and National Economic Growth

Most economic activity occurs in cities. This fact links national and urban growth. An economy can grow only if cities, or the number of cities, grow. In fact, Jacobs (1969) and Lucas (1988) underscore knowledge spillovers at the city level as a main engine of growth. The growth literature has also argued that, in order for an economy to exhibit permanent growth, the aggregate technology has to exhibit asymptotically constant returns to scale (Jones 1999). If not, the growth rate in an economy will either explode or converge to zero. How is this consistent with the presence of scale effects at the city level? Eaton and Eckstein (1997), motivated by empirical evidence on the French and Japanese urban systems, study the possibility of parallel city growth, which is assumed to depend critically on intercity knowledge flows together with the accumulation of partly city-specific human capital across a given number of cities. Rossi-Hansberg and Wright (2004) propose a theory where scale effects and congestion forces at the city level balance out in equilibrium to determine the size of cities. Thus, the economy exhibits constant returns to scale through the number of cities increasing along with the scale of the economy. Hence, economic growth is the result of growth in the size and the number of cities. If balanced growth is the result of the interplay between urban scale effects and congestion costs, these theories have important implications for the size distribution of cities and the urban growth process. These implications turn out to be consistent with the empirical size distribution of cities, that is, Zipf's Law, and with observed systematic deviations from Zipf's Law.

To summarize: urban growth affects the efficiency of production and economic growth, and the way agents interact and live in cities. Understanding its implications and causes has captured the interest of economists in the past and deserves to continue doing so in the future.

See Also

- ▶ City and Economic Development
- ▶ Endogenous Growth Theory
- ▶ Location Theory
- ▶ New Economic Geography
- ▶ Power Laws
- ▶ Spatial Economics
- ▶ Symmetry Breaking
- ▶ Systems of Cities
- ▶ Urban Agglomeration
- ▶ Urban Economics
- ▶ Urban Environment and Quality of Life
- ▶ Urban Production Externalities
- ▶ Urbanization

Bibliography

Anas, A., R. Arnott, and K. Small. 1998. Urban spatial structure. *Journal of Economic Literature* 36: 1426–1464.

Bairoch, P. 1988. *Cities and economic development*. Chicago: University of Chicago Press.

Chatterjee, S., and G. Carlino. 1999. Aggregate metropolitan employment growth and the deconcentration of metropolitan employment. Working paper. Philadelphia: Federal Reserve Bank of Philadelphia.

De Vries, J. 1984. *European urbanization: 1500–1800*. Cambridge, MA: Harvard University Press.

Duranton, G. 2004. Urban evolutions: The still, the fast, and the slow. Working paper. London: Department of Geography and Environment, London School of Economics.

Eaton, J., and Z. Eckstein. 1997. Cities and growth: Theory and evidence from France and Japan. *Regional Science and Urban Economics* 27: 443–474.

Eeckhout, J. 2004. Gibrat's law for (all) cities. *American Economic Review* 94: 1429–1451.

Fujita, M., and H. Ogawa. 1982. Multiple equilibria and structural transition of nonmonocentric urban configurations. *Regional Science and Urban Economics* 12: 161–196.

Fujita, M., P. Krugman, and A. Venables. 1999. *The spatial economy: Cities, regions, and international trade*. Cambridge, MA: MIT Press.

Fujita, M., T. Mori, J. Henderson, and Y. Kanemoto. 2004. Spatial distribution of economic activities in Japan and China. In *Handbook of regional and urban economics*, ed. J. Henderson and J.F. Thisse, Vol. 4. Amsterdam: North-Holland.

Gabaix, X. 1999. Zipf's law for cities: An explanation. *Quarterly Journal of Economics* 114: 739–767.

Gabaix, X., and Y. Ioannides. 2004. The evolution of city size distributions. In *Handbook of regional and urban economics*, ed. J. Henderson and J.F. Thisse, Vol. 4. Amsterdam: North-Holland.

Glaeser, E., and J. Gyourko. 2005. Urban decline and durable housing. *Journal of Political Economy* 113: 345–375.

Henderson, J., and J.-F. Thisse, ed. 2004. *Handbook of regional and rban economics*. Vol. 4. Amsterdam: North-Holland.

Ioannides, Y. 1994. Product differentiation and economic growth in a system of cities. *Regional Science and Urban Economics* 24: 461–484.

Jacobs, J. 1969. *The economy of cities*. New York: Random House.

Jones, C. 1999. Growth: With and without scale effects. *American Economic Review P&P* 89(2): 139–144.

Krugman, P. 1991. Increasing returns and economic geography. *Journal of Political Economy* 99: 483–499.

Lucas, R. Jr. 1988. On the mechanics of economic development. *Journal of Monetary Economics* 22(1): 3–42.

Lucas, R., and E. Rossi-Hansberg. 2002. On the internal structure of cities. *Econometrica* 70: 1445–1476.

Marshall, A. 1920. *Principles of economics*. 8th ed. London: Macmillan.

Papageorgiou, Y., and T. Smith. 1983. Agglomeration as local instability of spatially uniform steady-states. *Econometrica* 51: 1109–1119.

Rossi-Hansberg, E., and M. Wright. 2004. Urban structure and growth. Working paper. Stanford: Department of Economics, Stanford University.

Urban Housing

D. Harvey

The specifics of history and geography indicate a vast array of urban housing types and patterns within equally divergent sets of urban circumstance. The problem is to bring all this particularity and variety into a frame of reference that will help us understand the social, economic, cultural and political significance of urban housing.

Housing means more than just shelter from the elements. It defines a space of social reproduction that necessarily reflects gender, familial, and other types of social relations. It can also function as a place of manufacture and commerce, of leisure, education and religious observance, of ordered social intercourse. Whether or how it performs such functions depends on the nature of the social order – its dominant mode of production, consumption and reproduction, its hegemonic class, gender and ethnic relations, its cultural requirements and form of urbanization.

The separation of working and living in the capitalist city, for example, arises out of a mode of production founded on wage labour. In the medieval city working and living were often kept under the same roof. In many pre-capitalist muslim cities, on the other hand, concern over privacy and the seclusion of women pushed the activities of artisans and traders out of the house and into the streets and markets. Thus did gender relations in one place forge a pattern of uses made necessary by class relations in another.

The ability of the house to accommodate diverse functions has also varied considerably.

Apart from technological and organizational limitations on housing construction and design (interior plumbing being perhaps the most important), competition for urban space and housing costs sometimes make it impossible to meet basic needs within the house. Working class immigrants to 19th-century Paris, for example, found such limited accommodation that most were forced to eat on the streets or in the cafés and cabarets. Recent muslim immigrants into Ibadan likewise find it hard to procure housing suited to the perpetuation of kinship relations. In both cases, problems of urban housing provision forced sometimes disruptive social adaptations. Eating out in the modern American city has, however, a quite different social signification (particularly for the upper classes), reflecting in part the attractions of urban alternatives and the limitations of time (rather than space) in a society where the role of women has undergone a significant transformation.

Considerations of law, social structure and culture therefore intersect with those of economy and technology to dictate the status of housing provision and use. Courtyards and compounds – the former being the oldest of all known urban housing types – suit a kinship system and can be easily adapted, as they have in the muslim world, to societies where polygamy and the seclusion of women are important. Within compounds, housing units are built or allowed to collapse as households form or dissolve within the kinship frame. Increasing density and the shift from collective to private property ownership put pressures on such flexible use of space, though the populations in the muslim quarters of even the highest-density African and Asiatic cities go to extraordinary lengths to preserve such possibilities by elaborate use of rooftops and interior adjustments of housing design. While kinship does not necessarily die out with modern high-rise apartment block living, it does become harder to sustain. This accounts for the social preference in many third-world cities for squatter settlements that subvert private property rights and permit the replication of more traditional social structures within more flexible spaces than those provided by public or commercial housing systems (often based on nuclear family concepts).

Housing construction entails more than a simple technological capacity to defy the laws of gravity and vault a roof over covered space. Though building technologies have changed over the ages housing styles have remained remarkably persistent. Geographical variation is much more emphatic because housing has to respond to different environmental conditions (climate, drainage, disease, etc.) using local construction materials and labour skills. The house also embodies local cultural preferences for light, ventilation and privacy, while responding to the needs of security, mobility, economy, and social structure. The range of primitive housing types extends from the eskimo igloo through the Berber tent to the cave dwelling, the wattle hut, the log cabin, and the adobe house. Out of these primitive house types a wide variety of vernacular pre-industrial architectures were evolved (see Rapoport 1969), many of which could be adapted to urban circumstances.

But urbanization always imposes social, economic and political as well as physical constraints. Modifications to rural vernacular led urban housing to evolve in quite distinctive ways. Furthermore, the pursuit of permanence and physical symbols of authority led to a much greater emphasis upon monumental forms of building in urban settings. These required new techniques and styles, elements of which could be incorporated into house building activity, in the first instance for the ruling class but later on for less privileged strata. Urbanization therefore entailed a complex interaction between these monumental styles and urban vernacular (Kostof 1985). Traditional vernacular forms do not necessarily disappear however. They can be reimported by rural migrants (in the contemporary shanty towns), or resurrected by architects as better physical or commercial adaptations (as with the use of Japanese or Islamic open style housing in California).

Urban housing cannot be understood independently of the kind of urbanization in which it is embedded. Urbanization depends upon the production, appropriation and geographical concentration of economic surpluses under the aegis of some ruling class. Different modes of production and appropriation associate with different class

structures and produce different types of urbanization. The theocratic cities of the ancient world, meso-America and pre-colonial Africa had much in common with each other in spite of wide separations in time and space, than they did with the bureaucratic cities of the Orient and Asia, the classical urbanization of Greece and Rome, the feudal cities of medieval Europe, the industrial and so-called 'post industrial' cities of advanced capitalism, and the contemporary cities of the neo-colonial or socialist bloc worlds. The quantities, qualities, and functions of urban housing reflect and help create such differences.

The organization of space and of dwelling units in a theocratic city that depended on the direct appropriation of agricultural surpluses expressed the symbolic and hierarchical requirements of the theocratic order. Where people lived and the nature and style of their building symbolized social position directly. As theocratic cities evolved into bureaucratic centres of organized political and military power, leavened by some degree of market exchange and artisanal production, so the organization of space and housing provision had to be opened up to more divergent functions and influences. It was, however, still subject to state regulation because location, style and function continued to signify positions of relative power and prestige within the social order. Political and social conflict over housing has often been, therefore, a signal of deeper stresses in society. Transitions from one mode of production and class domination to another have always coincided with rapid shifts in the forms and functions of urban housing. This general theme can perhaps best be illustrated by the evolution of urban housing in the passage from Western feudalism to capitalism and beyond.

Urbanization under feudalism was supported primarily out of the direct extraction of economic surpluses from the land (tithes, money rents, taxes) or their accumulation out of often tightly controlled trade. Towns were religious, political and military centres with ancillary market activities and artisanal production. Buildings, even those alienable under laws of private property, were not generally produced as commodities but as use values. Even though regulated guild labour (the predominant form of organization of construction) was frequently pushed to adapt vernacular or imported building styles for purposes of conspicuous display, the tight ordering of the social structure made the unrestrained flouting of individual wealth through building dangerous. Even as late as the 17th century, Louis XIV of France would brook no rivals in house building and the aristocracy applied the same pressure to a nascent bourgeoisie. Only in the Italian city states did a more powerful and autonomous merchant class permit a broader basis to conspicuous housing consumption (with effects visible to this day).

Merchants, shopkeepers, artisans and administrative officials built soberly under the jealous eye of church and state, but did so on a sufficient scale to house themselves, their families, servants, apprentices and employees. Only wage labourers (of which there were always fluctuating numbers) and the urban poor were forced to find independent means of shelter. Shanty towns, though frequent, were ephemeral and frowned on by authority, forcing the conversion of older structures into overcrowded lodging houses for rent. This commodification of housing had important consequences. Not only did it force the urban poor to become wage labourers in order to pay their rooming costs, but it also stimulated the production of housing as a speculative commodity wherever the wage labourers increased in numbers. Sold to landlords anxious to turn a profit out of the labourers' paltry wages, the new housing was cheaply built (often by casual rather than craft labour), thus forming the instant slums of many medieval towns. Ironically, the only other strata served by such speculative building were the aristocrats and high state functionaries needing to rent temporary accommodations close to the centres of political power.

The commodification of housing within the medieval frame had all kinds of social and physical consequences. Not only did it open up location and house building to the power of money and markets, but it also hinted at radical changes in social relations that came to fruition only with the rise of industrial capitalism. Entrepreneurial activity in building and the widespread use of wage labour in construction was an important

and often underestimated step in the transition from feudalism to capitalism.

Speculative house building was of two types. 'It is', noted Marx, 'the ground rent, and not the house, which forms the actual object of building speculation in rapidly growing cities.' The traditional power of landowners remained unchallenged while the builders remained undercapitalized and weak, usually beholden to the superior financial power of the landed capitalists. This style of speculation produced the large estates of upper class housing in older urban centres as well as in the new spas and resorts. The Georgian terraces in England provide excellent examples (Chalkin 1974). This system would later be adapted in England to the production of densely packed back-to-back housing for workers in industrial cities.

Speculative builders were of a different breed. Viewing land as just one of many inputs to their production process, they sought to produce housing for profit. Originating as craft workers who organized other craft workers and casual labour into an entrepreneurial production system, they were initially short on capital and credit, remained small-scale and locked into vernacular designs using traditional construction methods. Though entrepreneurial, they found it difficult to liberate themselves from the overwhelming power of landowners. And even when they could, the tyranny of lot size and land scarcity, coupled with lack of effective demand, constrained what they could build (awkward lot size played a key role, for example, in shaping the extraordinary design of New York's tenement housing). Only when finance capital (armed with new institutions and powers) came to control both land development and building jointly, did the industrialization of urban housing production become possible.

The rise of industrial capitalism changed the face of urbanization. The industrial city's business was the production of surpluses through the exploitation of wage labour in production organized within the urban frame. The radical reorganization of labour processes under industrial capitalism undermined craft distinctions and reorganized an ever-growing working class into a hierarchy of skills and authority positions to which there corresponded a hierarchy of wage rates. The housing needs of this vast and growing mass of heterogeneous wage labourers had to be met by commodity production. But how? Industrial capitalism had to find an answer to that question or simply collapse under the weight of its own inability to assure the reproduction of labour power as its most valuable commodity.

The conversion of pre-existing structures to lodging houses, tenements and rooming units, and the opening up of damp cellars and attics ill-protected from the elements formed one set of solutions. To this were added the efforts of speculative builders creating blocks of housing here (the back-to-backs of Britain, the *cités ouvrières* of France, the terrace housing of the United States) or jerry-built structures on odd land lots there. Only occasionally did employers seek to provide worker housing on their own account (Bourneville and Port Sunlight in England; Lowell and Pullman in the United States). Short of capital and low on income, most workers were forced to rent, and although the quality of housing varied, the net effect was to produce the dramatic overcrowding and all the signs of physical and social breakdown that made the industrial city such an appalling place. The more privileged strata of wage workers and the middle and upper classes for their part increasingly used their money power to escape contact with what they saw as dens of disease, vice and misery (to say nothing of the threats of the urban rabble). As early as 1844, Engels showed how suburbanization (later aided and abetted by revolutions in urban mass transportation) and residential differentiation reflected class-bound stratifications in money power. The housing spaces of the capitalist city came to be seen as 'natural' outcomes of competition and the commodification of housing provision and use in spatially segregated markets.

Bourgeois concerns over health, social unrest (the 'housing question' played an important role in sparking such events as the Paris Commune as well as movements towards municipal socialism) and the qualities of labour power, coincided with strengthening working-class movements to produce substantial housing reform movement throughout the capitalist world in the latter half

of the 19th century. If commodity provision and delivery (through landlordism) could not meet the needs and aspirations of the mass of urban workers, then other solutions to the housing question had to be found.

The first solution was to treat housing as an inalienable use value. Social housing produced through state action became a conspicuous feature of urban housing in all advanced capitalist countries in the 20th century, with the exception of the United States. The quantity, quality and location of this housing depends upon state policies (a question of political power and class interests), fiscal restraints (the tax base and borrowing costs), and the ability to organize production (land procurement, industrialization of building technologies, economies of scale, etc.). Organized state planning of housing provision was problematic, at least in part because it smacked heavily of some transition from capitalism to socialism. The second solution, therefore, was to reorganize housing provision around the principle of home ownership ('the last defence against bolshevism', as the head of the British Building Society movement put it in the 1920s). This solution depended upon rising worker incomes and job security coupled with new financial instruments that organized flows of savings and mortgage credit to the more privileged segments of the working class. State support was essential to the creation of such conditions, both in legalizing limited trade union power and supporting homeownership mortgage markets. Once the mass market had been clearly established (particularly after World War II), speculative building could gear up to mass produce privatized housing (the suburban building estates, the Levittowns, etc.). Revolutions in building technology together with the declining power of landowners relative to finance capital and the state, transformed housing provision. And finally, with increasing interest in tools for government intervention after the depression of the 1930s, housing became a target of fiscal and monetary policy. In the United States, for example, government policies accelerated suburbanization and access to individualized homeownership after 1945 so as to contain business cycles and sustain the post-war boom. The consequent deterioration of inner-city housing was an unfortunate side effect that played an important role in massive urban unrest. Housing continues to be a central issue in many urban social movements as well as in local and national politics.

Disparities in money power, modified by legal conditions of tenure and state intervention, typically generate marked residential differentiation in housing types and qualities. While the focus on the nuclear family is very strong, housing is still used to signal status, prestige and class power as well as life-style preferences, cultural affinity, gender relations and religious or ethnic identity. The 'housing question' remains a serious social and political issue in advanced capitalist cities and periodically becomes the focus of intense political struggles that sometimes touch at the very root of the capitalist form of urbanization. That this is so testifies to the cogency of Engels' remark that the 'manner in which the need for shelter is satisfied furnishes a measure for the manner in which all other necessities are supplied'.

See Also

- ▶ Housing Markets
- ▶ Property Taxation

Bibliography

Abrams, C. 1964. *Man's struggle for shelter in an urbanizing world*. Cambridge, MA: MIT Press.

Ball, M. 1981. The development of capitalism in housing provision. *International Journal of Urban and Regional Research* 5: 145–177.

Burnett, J. 1978. *A social history of housing, 1815–1970*. Newton Abbott: David and Charles.

Chalkin, C. 1974. *The provincial towns of Georgian England: A study of the building process*. London: Edward Arnold.

Dwyer, D. 1975. *People and housing in third world cities*. London: Longmans.

Engels, F. 1844. *The condition of the working class in England*. London: Parker, 1974.

Harvey, D. 1985. *The urbanization of capital*. Oxford: Blackwell.

Houdeville, L. 1969. *Pour une civilisation de l'habitat*. Paris: Editions Ouvrières.

Kostof, S. 1985. *A history of architecture: Settings and rituals*. Oxford: Oxford University Press.

Rapoport, A. 1969. *House form and culture*. Englewood Cliffs: Prentice-Hall.
Schwerdtfeger, F.W. 1982. *Traditional housing in African cities*. New York: Wiley.
Vance, J. 1966–7. Housing the worker. *Economic Geography* 42: 294–325; 43, 94–127.

Urban Housing Demand

Todd Sinai

Abstract

Urban housing demand is a reflection of households' desire to live in cities. In this article, I discuss possible reasons why US households have exhibited an increasing taste for urban living, including employment, urban amenities, and consumption opportunities. Next, I explain how growing urban housing demand led to rising house prices and a sorting of households across cities by income. That dynamic generated a divergence across housing markets in the value of the tax subsidy to owner-occupied housing as well as housing market risk. Those factors, in turn, had a feedback effect on urban housing demand.

Keywords

Housing markets; Housing supply; Housing tax subsidies; Internet, economics of; Monocentric city models; Population growth; Preference externalities; Productivity growth; Superstar cities; Urban housing demand; Willingness-to-pay

JEL Classification

R21

At its core, the demand for urban housing is just the manifestation of the demand for living in urban areas. On net, residence patterns suggest that most people want to live in or near cities, and that desire is increasing over time. In fact, by 1999, 75% of US households lived in cities (Rosenthal and Strange 2003). Today, urban America is where housing demand is most likely to exceed housing supply and generate rising house prices, where the tax system provides the greatest subsidy to owner-occupied housing, and where the housing market is the most volatile. In this article I discuss some of the causes and consequences of urban housing demand, and the supporting evidence.

Location

The classic explanation for the concentration of households in cities is that people want to live close to their jobs. That notion, developed in the monocentric city model of Alonso (1964), Mills (1972) and Muth (1969), leads to a prediction that is rarely as evident in reality as it is in theory: housing costs should rise as the distance to the employment centre falls since households would be willing to pay more in order to save time getting to work. Instead, households often settle for a longer commute in exchange for other positive qualities of a non-urban community, such as the density of development, the calibre of the school system, local taxes and amenities, and the similarity of the other residents to themselves.

Since the 1960s, the patterns of where people live have begun to shift back to cities, even though people are now less likely to work in the downtown areas. According to Glaeser et al. (2001), between 1960 and 1990 the rate of growth of commutes where the household lives in the city increased while the growth rate of commutes originating in the suburbs fell. Within cities, the high-income population has been moving closer to the central downtown area. Glaeser et al. argue that nowadays thriving cities are 'consumer cities', ones that attract highly educated households through appealing cultural amenities, such as museums, restaurants and opera. In fact, between 1977 and 1995, a temperate and dry climate, a coastal location, and more live performance venues and restaurants per capita predicted future population growth. By contrast, having more bowling alleys was correlated with population decline.

Indeed, the very congestion that urban economists typically point to as a reason that cities become unattractive may lead to an availability and quality of goods and services that are appealing. In a city, the large number of residents living in close proximity makes it feasible for even niche markets to be served since a critical mass of potential customers exists. Joel Waldfogel (2003), who in a series of papers termed this phenomenon 'preference externalities', found empirical support in the markets for broadcast radio, newspapers, and restaurants. For example, when there is a larger local consumer base for a certain format of radio station, calibre of newspaper, or style of restaurant, the more of them exist in a city. By revealed preference, that greater variety increases city dwellers' welfare, because the more options there are for residents that share a particular set of tastes the more they consume.

Even the advent of the Internet has not dampened the consumption appeal of living in cities. Since the Internet makes information and goods universally available to anyone, no matter where they live, it substitutes for living in a city. However, Sinai and Waldfogel (2004) find that the number and variety of websites focused on a city increases with the city's population. By enhancing the welfare benefits of living in cities – perhaps by mitigating the effects of congestion or facilitating communication and connection among city residents – these sites have an offsetting positive effect on urban housing demand.

Urban Housing Demand and House Prices

Two measures of the intensity of urban housing demand are house prices and the rate of house price growth. In some cities, housing is in inelastic supply because there is little or no open land and local regulations either restrict development or make it prohibitively expensive or slow. In that case, demand for a location leads to bidding up of the price of land in order to equilibrate housing demand with the available supply. Indeed, when one compares house prices across cities and town, areas that presumably have higher demand because they offer better amenities and fiscal conditions exhibit higher house prices (Roback 1982). Another indication of high demand for a city is population growth, which occurs when housing development is easy. I focus on high house prices because they can change the character of a city, which then has a further effect on urban housing demand.

Since the 1950s, a handful of metropolitan areas experienced real house price growth that significantly exceeded the national average, leading to a widening gap across locations in average house prices. For example, in 1950 the average house price in San Francisco was 37% higher than the average across all metropolitan areas. By 2000 the gap had grown to 218%. In order for land prices to continually grow in one location relative to another, the demand for that location must be growing as well. One possible explanation is that productivity growth in a handful of cities has exceeded the national average, and residents pay more to live in productive cities because their wage rises with their productivity. Another potential rationalization is that some cities are becoming more appealing over time and residents are paying more for increasingly higher quality.

Another possibility is that the rapid growth in the number and earnings of high- income households in the United States has led to an increased willingness-to-pay for scarce locations. Since some cities are in such limited supply, households have to outbid each other to live there, leading to land prices that grow with the aggregate spending power of the clientele that prefers that particular city. In 'Superstar Cities', Gyourko et al. (2006) show that inelastically supplied, high-demand cities have income distributions that are shifted to the right: low-income families can live there only if they have a very strong preference for the city, while high-income families can live there even if they only modestly prefer it. As the national high-income population grows, the greater number of high-income families outbid relatively low-income families (as well as some high-income families) who are unwilling or unable to pay a higher premium to live in their preferred location. Gyourko, Mayer and Sinai find that such superstar locations

experience supra-normal house price growth and a shift of their income distributions to the right as they experience inflows of high-income households and outflows of their lowest-income residents. This pattern has been intensified as cities have begun to 'fill up' due to the growing national population. For example, in 1960 only Los Angeles and San Francisco were demonstrably inelastically supplied. By 2000 more than 20 cities were. Gyourko, Mayer and Sinai show that cities that 'fill up' experience a right- shift in their income distributions and higher price growth after their transition into superstar city status.

These findings imply that there must be something unique and attractive about superstar cities, otherwise potential residents would turn to cheaper locations and superstar cities would not be able to sustain excess price growth. A niche-market appeal may be due to particular amenities, or the kind of preference externalities described by Waldfogel. As preference agglomerations form, the highest willingness- to-pay households are those that share the same preferences. If such sorting is along income lines, rising house prices can lead to high-income homogeneity, which itself makes an area more desirable to high-income residents. That dynamic implies that certain urban markets will evolve into luxury areas and grow increasingly unaffordable for the average household.

The inelasticity of housing supply also leads to price changes, and a correlated change in the demand for urban housing, in cities that are experiencing *declining* demand. Glaeser and Gyourko (2005) point out that, since housing does not quickly depreciate once built, if the demand for a city declines then house prices must fall since quantity cannot easily adjust downwards. That decline in prices can spur demand by low-income households that cannot afford to live anywhere else, leading to sorting into low-income enclaves rather than high-income ones.

The Tax Subsidy to Owner-Occupied Housing

Differences in house prices among cities also affect the benefits homeowners obtain from their houses, which in turn affect the demand for urban housing. One such benefit in the United States that often is especially valuable in cities is the favourable Federal income tax treatment for owner-occupied housing, worth a total of $420 billion in 1999. The nature of these tax benefits is well-documented in this edition of the *New Palgrave* (housing policy in the United States). Gyourko and Sinai (2004) note that two conditions are necessary in order to receive a high value of this tax subsidy: a high-priced home, so that the subsidy operates on a larger base, and a high tax rate, which in the progressive US tax system follows from having a high income.

Because of this, the very same superstar city dynamic discussed earlier leads to an unequal distribution of the housing subsidy across the country. Superstar cities experience both house price growth and relatively high-income residents, and thus should also have the highest tax subsidies, further increasing the demand for urban housing in hot markets. Indeed, the tax subsidy is highly concentrated in a handful of cities, with just five metropolitan areas receiving more than 85% of the total tax benefits in 1990. Between 1980 and 2000, the rise in house prices in superstar cities more than offset declining marginal tax rates, leading to a greater concentration of tax benefits in a handful of metropolitan areas.

This tax subsidy has been shown to lead to higher house prices, either because the subsidy induces households to consume a larger quantity of housing or simply because house prices capitalize the present value of the future tax savings. Recent estimates of the after-tax price elasticity of housing demand cluster around -0.5, and the income elasticity around 0.25. Urban areas tend to exhibit relatively high demand elasticities, as demand is more readily capitalized into land prices rather than the limited new supply. By contrast, rural areas have much lower measured elasticities of housing demand.

Risk and the Demand for Urban Housing

Since urban housing markets tend to have inelastic supply, they are more volatile as shocks to

housing demand are transmitted more completely into rents and prices. That higher risk may deter households from living in urban areas since they would face more uncertainty over housing costs, whether they rented or owned. Also, house price volatility generates an additional cost because it distorts other investment decisions (Flavin and Yamashita 2002). A mitigating factor, demonstrated in Sinai and Souleles (2005), is that long length-of-stay households can reduce their effective risk by owning their houses, in essence prepaying their housing costs. Other research suggests that uncertainty over house price growth simply may lead households to purchase housing in a city sooner than they otherwise would have in order to prevent housing costs from outpacing their income growth.

See Also

- ▶ Housing Policy in the United States
- ▶ Housing Supply
- ▶ Residential Real Estate and Finance
- ▶ Residential Segregation
- ▶ Tiebout Hypothesis
- ▶ Urban Agglomeration
- ▶ Urban Economics
- ▶ Urban Environment and Quality of Life
- ▶ Urban Political Economy
- ▶ Urbanization

Bibliography

Alonso, W. 1964. *Location and land use*. Cambridge, MA: Harvard University Press.

Flavin, M., and T. Yamashita. 2002. Owner-occupied housing and the composition of the household portfolio over the life cycle. *American Economic Review* 92: 345–362.

Glaeser, E., and J. Gyourko. 2005. Urban decline and durable housing. *Journal of Political Economy* 113: 345–375.

Glaeser, E., J. Kolko, and A. Sarz. 2001. Consumer city. *Journal of Economic Geography* 1: 27–50.

Gyourko, J., and T. Sinai. 2004. The (un)changing geographical distribution of housing tax benefits: 1980 to 2000. In *Tax policy and the economy*, vol. 18, ed. J. Poterba. Cambridge, MA: MIT Press.

Gyourko, J., C. Mayer, and T. Sinai. 2006. *Superstar cities*, Working paper, vol. 12355. Cambridge, MA: NBER.

Mills, E. 1972. *Studies in the structure of the urban economy*. Baltimore: Johns Hopkins University Press.

Muth, R. 1969. *Cities and housing*. Chicago: University of Chicago Press.

Roback, J. 1982. Wages, rents, and the quality of life. *Journal of Political Economy* 90: 1257–1278.

Rosenthal, S.S., and W.C. Strange. 2003. Evidence on the nature and sources of agglomeration economies. In *Handbook of urban and regional economics*, ed. J.-V. Henderson and J.-F. Thisse. Amsterdam: North-Holland.

Sinai, T., and N. Souleles. 2005. Owner-occupied housing as a hedge against rent risk. *Quarterly Journal of Economics* 120: 763–789.

Sinai, T., and J. Waldfogel. 2004. Geography and the Internet: Is the internet a substitute or complement for cities? *Journal of Urban Economics* 56: 1–24.

Waldfogel, J. 2003. Preference externalities: An empirical study of who benefits whom in differentiated product markets. *RAND Journal of Economics* 34: 557–568.

Urban Political Economy

Robert W. Helsley

Abstract

Models of local public finance generally emphasize the roles of household mobility and community heterogeneity in the provision of local public services. In contrast, the emerging field of urban political economy examines how economic and political institutions influence the formation of local public policies. Key issues include the strength of the local executive, whether local politicians are elected 'at large' or to serve the interests of particular wards, the norms that govern behaviour and decisionmaking within city councils, and institutional innovation, especially the growth of so-called 'private governments'.

Keywords

Common pool problem; Local public finance; Minimum winning coalitions; Parliamentary systems; Presidential systems; Principal and agent; Private government; Tiebout hypothesis; Urban political economy

JEL classification
R51

Economic models of local government applied to the United States generally suppress the roles of politics and political institutions. Indeed, the dominant model of local government, the Tiebout (1956) model of the provision of local public services to mobile residents, can be seen as an explicit attempt to eliminate the need for politics in cities. If individuals are highly mobile and communities offer a diverse menu of local taxes and expenditures, then there is no need for political expression – households can satisfy their demands for local public services by choosing to live in the community that provides their optimal bundle. Households 'vote with their feet'. When local politics are explicitly considered, the political process is usually treated as an idealized form of majority rule in which residents vote directly on tax and spending programmes, and the political outcome corresponds to the most preferred policy of the median voter. This 'institutionless' view of local public finance has in fact been quite successful in, for example, characterizing the demand for local public goods (Rubinfeld 1987).

Local Political Institutions

However, most local policy choices are not made directly by residents. According to the International City/Council Management Association (ICMA), 43.7% of US municipalities with populations over 2500 were governed by the combination of a mayor and a city council in 2000, while 48.3% were governed by the combination of a city council and a city manager. Thus, 90% of US municipalities were governed at least in part by a representative body. Council members may be elected 'at-large', that is, from the entire city, or by wards or districts within the city. Some cities adopt a mixed system, in which the council contains both at-large and ward representatives.

Mayors (or their offices) are traditionally classified as being either 'strong' or 'weak'. Strong mayors have broad powers, including a veto over some city council decisions. Strong mayors also prepare the city's budget, and have hiring and firing authority over the heads of city departments. In weak mayoral systems, most executive and legislative authority rests with the city council; the mayor performs largely ceremonial and organizational functions. Strong mayors are generally elected independently from members of the city council, and are more common in mayor–council systems. Baqir (2002), based on a sample of roughly 2000 US municipalities in 1990, reports that 98% of mayors in mayor–council systems were independently elected, compared to 65% of mayors in council–manager systems.

Strong mayors are generally associated with fiscal discipline, and there is some support for this view in other branches in the political economics literature. For example, the literature on comparative politics suggests that presidential systems have greater accountability to voters and less collusion within and between the branches of government than parliamentary systems (Persson et al. 1997, 1998, 2000). Persson et al. (2000) show that presidential systems have lower levels of government spending as a share of national product. Inman and Fitts (1990) show that between 1795 and 1988 'strong' presidents (those with 'independent political strength', identified from a survey of historians) were associated with lower levels of federal spending in the United States. Baqir (2002) suggests that a strong mayor may have a similar disciplinary effect on local government spending.

Many studies of local political institutions in North America examine the impacts of the reform movement of the early twentieth century. The reform movement brought a number of changes in local government structure that were allegedly designed to limit the exercise of private interest and patronage in city politics and promote the pursuit of public interests and professional management. Some of the specific institutional changes that followed included the introduction of at-large and non-partisan elections for city council (a change that has since been partially reversed), the council–manager form of local government, civil service exams as a basis for appointment and promotion in the bureaucracy, and, in some areas, the replacement of the

mayor–council form with a group of city 'commissioners', each of whom had executive and legislative responsibility for a different city department.

Early studies of reform governments expressed the hope that managerial expertise and autonomy in personnel matters could lead to lower costs for the delivery of local public services, and in particular, lower labour costs for municipal governments. However, subsequent empirical studies provide little support for this view: public expenditure levels and patterns in US cities seem to have been largely unaffected by the adoption of city managers, at-large representation and non-partisan elections.

The most compelling study of the reform movement in the recent economics literature is Rauch (1995). Rauch's hypothesis is that by creating a population of career bureaucrats in city government, the reform movement put in place incentives that encouraged investment in infrastructure and other 'long-gestation-period' projects. Rauch views the relationship between the city council and the bureaucracy as a principal–agent problem. Before reform, the agent, that is, the bureaucracy, is assumed to act as a political appointee who shares the council's immediate focus on retaining office. After reform, the bureaucracy is professionalized and the agent is assumed to have some job security and therefore a longer time horizon. The agent may then use his 'powers of information collection and expenditure oversight', in combination with costly or imperfect monitoring by the principal to direct resources towards longer-term projects that may further the agent's career. The implication is that this type of reform should increase the share of expenditures devoted to investment, as opposed to current public consumption. Using a panel of 144 cities over 23 years, Rauch regresses the infrastructure share of municipal expenditure on dummy variables for the use of civil service exams, the presence of a city manager, and the adoption of a commission form of local government. After accounting for the inertia generated by the durability of infrastructure investment, use of the civil service is found to have a positive impact on the share of expenditure devoted to infrastructure. Interestingly, in the cases where they are statistically significant, the presence of a city manager and the adoption of a commission form of government are both associated with lower levels of infrastructure spending.

The Common Pool Problem in City Councils

City councils are, in effect, local legislatures. One way to model the operation of a city council is by analogy with models of other legislative institutions. In that spirit, imagine a city council in which each councillor represents a well-defined local constituency. If councillors are elected by ward or district, then the constituencies will be geographic, as in most national, state and provincial legislatures. Councillors elected at-large may have non-geographic constituencies that are defined by a common ideology or policy initiative. Suppose that each councillor is motivated by holding office and that this gives him or her an incentive to pursue programmes and policies that provide net benefits to his or her constituents. It is generally assumed that the policies and programmes that are chosen by legislatures are 'distributive' in the sense that their costs are more widely distributed than their benefits. For example, benefits may be restricted to a particular district or group, while the supporting tax payments are made by residents of the entire city. Spending and tax choices are made by a majority vote of council members.

The literature on legislative decision-making discusses a number of issues that relate to the efficiency of the policy choices that will emerge in this context. First, there is an incentive for 'minimum winning coalitions' within the legislature to form for the purpose of approving distributive policies (Riker 1962). A minimum winning coalition is the smallest set of legislators that can guarantee passage of a proposal under majority voting. If proposals or projects have spillover costs and benefits, as distributive policies generally do, then the exclusion of the interests of delegates outside of a winning coalition will lead to inefficient choices. Second, minimum winning coalitions should be highly unstable, since

excluded delegates have strong incentives to alter the coalition structure. Each member of the legislature faces some probability that he or she will be excluded from the minimum winning coalition for any particular policy proposal. Third, Weingast et al. (1981), Shepsle and Weingast (1984), and others suggest that the resulting uncertainty helps explain the practice of 'universalism', in which the size of coalitions and the set of approved projects exceed the minimum winning size. In its extreme form, universalism involves a 'norm of reciprocity' in which each delegate supports the project of every other, and so a project for every delegate or constituency is approved. Finally, Weingast et al. (1981) argue that politicians in such a setting have an incentive to count the resource costs of geographically earmarked programmes as benefits. They refer to this as the 'Robert Moses' effect: 'pecuniary gains in the form of increased jobs, profits, and local tax revenues go to named individuals, firms, and localities from whom the legislator may claim credit and exact tribute' (1981, p. 648). Such 'political cost-accounting' will obviously encourage individual representatives to support higher than efficient levels of public spending.

More formally, following Persson and Tabellini (2000, section 7.1), imagine that there are M seats on the city council and that the fixed population of each constituency is N. Thus, the aggregate population of the city is MN. If council members are elected by district or ward, so the constituencies are geographic, then the assumption of fixed constituencies implies that the population is immobile. Suppose that all residents are identical and have quasi-linear preferences of the form $U(g) + x$, where g is per capita consumption of a publicly provided good, x is the numeraire and the sub-utility function $U(\cdot)$ is increasing and strictly concave. All residents have the same exogenous income y. Public services are financed through lump-sum taxes that balance the city's budget. Each councillor is assumed to be a perfect representative of his or her constituent group.

If one takes utilitarianism as a normative benchmark, the efficient provision of public services in this symmetric setting maximizes aggregate utility $M(U(g) + x)$ subject to the resource constraint $M N(y - x - g) = 0$. The first-order condition for this problem implies $U'(g) = 1$: the marginal benefit of the public service should equal its marginal cost in every constituency. Represent this efficient level of provision by g^*.

In contrast, under extreme universalism, or with decentralized provision and centralized finance, each delegate chooses a level of the public service to maximize the utility of a representative constituent, taking the levels chosen by other delegates as fixed. If one lets g^0 represent the conjectured level chosen by others, the balanced budget requirement implies that the lump-sum tax τ for any group satisfies $\tau M = g + (M - 1)g^0$. Thus, an individual delegate chooses g to maximize

$$y - \frac{g + (/M - 1)g^0}{M} + U(g). \tag{1}$$

The first-order condition for this problem implies $U'(g) = 1/M$ Each member of the legislature perceives that they pay only a fraction $1/M$ of the costs of the public services that they acquire. This is known as the common pool problem. If one lets g^U represent the level of provision under this extreme form of universalism, the concavity of $U(\cdot)$ implies $g^U > g^*$. The common pool problem thus leads to overprovision. Persson and Tabellini (2000, p. 163) summarize the nature of the distortion as follows: 'The problem here lies in the collective choice procedure, in which the tax rate is residually determined once all spending decisions have been made in a decentralized fashion. Concentration of benefits and dispersion of costs lead to excessive spending when such spending is residually financed out of a common pool of tax revenue.

The first-order condition for g^U implies

$$\frac{dg^U}{dM} = -\frac{1}{MU''(g)} > 0 \tag{2}$$

by concavity. Thus, the level of overprovision increases as the constituencies become smaller, *ceteris paribus*. Finally, if one allows $G^U = M g^U$ to represent aggregate spending, we have

$$\frac{dG^U}{dM} = g^U + M\frac{dg^U}{dM} > 0. \quad (3)$$

This is an instance of Weingast et al. (1981) 'law of 1/n': aggregate spending, and therefore the inefficiency of excessive spending increases with the number of constituencies or the size of the legislature.

This implication of the common pool problem seems to be supported by the evidence. Landbein et al. (1996), based on a sample of 192 cities in 1980, all of which have a council–manager form of government and a weak mayor, find that local government expenditure per capita is positively related to the number of elected members of the city council. Baqir (2002) finds that the size of US local governments (measured by expenditures or employment per capita or expenditures as a share of total income) increases with the size of the city council. Baqir also finds that expenditures (per capita or as a share of total income) are not significantly different in councils where a majority of members are elected at-large, but that local government employment per capita is lower when at-large councillors are in the majority. However, evaluated at the sample means, employment per capita is actually higher where a majority of councillors are elected at large. This is consistent with the hypothesis that at-large councillors serve their (non-geographic) constituencies in much the same manner that ward councillors serve the interests of their wards. The positive relationship between the size of the government and the size of the council is unaffected by the presence of at-large elections. Baqir also examines the impact of a strong city executive, and finds that expenditures do not increase with council size when the city has a strong mayor with the power to veto city council decisions. As noted above, this is consistent with recent models and results from the literature on comparative politics.

Private Government

Private governments are voluntary, exclusive organizations that supplement services provided by the public sector. There are two broad classes of private governments in the United States. Residential private governments, sometimes called residential community associations (RCAs), common interest developments (CIDs), or homeowner associations (HOAs), exist to further the interests of residential property owners. Commercial private governments, sometimes called business improvement districts (BIDs) or business investment areas (BIAs), exist to further the interests of their member firms. Private governments are highly controversial. Garreau (1991) labels them 'shadow governments', and argues that they are undemocratic, discriminatory, and operate outside of the constitutional restrictions that public governments face.

Residential private governments are an increasingly important component of housing markets and local government systems in North America. Garreau (1991, p. 189) estimates that there may have been as many as 130,000 RCAs in the United States in 1988. McKenzie (1996) reports that the number of CIDs in the United States grew from a few hundred in 1960 to 150,000 in 1993 and that they then housed 32 million people. The Community Associations Institute (an industry trade association) maintains that there were 231,000 RCAs in the United States in 2002, housing 57 million people. The 2001 American Housing Survey from the US Bureau of the Census reports that 28% of all new-housing residents paid community association fees in 2001. Residential private governments provide security and sanitation services, and manage and maintain common facilities, including recreational facilities and infrastructure. They also regulate property use and individual conduct through covenants, codes, and restrictions in property deeds.

There are fewer commercial private governments, but their impacts are also substantial. Pack (1992) estimates that there were 400 BIDs in the United States in 1992, while Mitchell's (2001) survey found 404 independently managed BIDs in the United States in 1999. BIDs typically provide security, marketing and sanitation services. Mitchell reports that 94% of BIDs engage in marketing, 85% provide maintenance and

sanitation services, and 68% provide security. Mitchell's survey also found that 88% of BIDs engaged in some form of policy advocacy, like lobbying governments on behalf of business interests. BIDs have become a key component of downtown revitalization strategies in many, if not most, major North American cities.

Private governments raise a number of interesting economic issues. First, they may have significant impacts on the traditional public sector. To the degree that private and public spending are substitutes, private governments provide a mechanism for the public sector to withdraw from certain activities. Helsley and Strange (1998, 2000a) show that such 'strategic downloading' is a characteristic of equilibrium in a game where public and private governments simultaneously choose levels of provision to maximize the welfare of their citizens and members, respectively. Cheung (2004) finds evidence of strategic downloading in a sample of California communities. Second, membership in private governments may be inefficient. Helsley and Strange (1999) argue that one of the essential features of gated communities is that they divert crime to other areas. This increases the incentive for other communities to engage in similar private policing activities (the activities are strategic complements), and may lead to excessive gated community development. Third, private governments have complex welfare effects. Citizens with high demands for public services are generally made better off by this form of privatization. By joining the private government, they can supplement what is for them an inadequate level of public provision. If strategic downloading occurs, citizens with low demands, who choose not to join a private government, are also better off, since the level of public provision falls towards their optimal level.

However, citizens in the middle of the distribution – whose demands were relatively well served by the traditional public sector – are made worse off.

The field of urban political economy is in its infancy. The roles that economic and political institutions play in the formation of local public policies are clearly deserving of further study.

See Also

- ▶ Local Public Finance
- ▶ Systems of Cities
- ▶ Tiebout Hypothesis
- ▶ Urban Economics
- ▶ Urban Environment and Quality of Life

Bibliography

Baqir, R. 2002. Districting and government overspending. *Journal of Political Economy* 110: 1318–1354.

Caro, R.A. 1974. *The power broker: Robert Moses and the Fall of New York*. New York: Knopf.

Cheung, R. 2004. *The interaction between public and private governments: An empirical analysis*. Mimeo: Department of Economics, Florida State University.

Garreau, J. 1991. *Edge cities: Life on the new frontier*. New York: Doubleday.

Helsley, R.W. 2004. Urban political economics. In *Handbook of regional and urban economics*, vol. 4, ed. J.V. Henderson and J.F. Thisse. Amsterdam: North-Holland.

Helsley, R.W., and W.C. Strange. 1998. Private government. *Journal of Public Economics* 69: 281–304.

Helsley, R.W., and W.C. Strange. 1999. Gated communities and the economic geography of crime. *Journal of Urban Economics* 46: 80–105.

Helsley, R.W., and W.C. Strange. 2000a. Potential competition and public sector performance. *Regional Science and Urban Economics* 30: 405–428.

Helsley, R.W., and W.C. Strange. 2000b. Social interactions and the institutions of local government. *American Economic Review* 90: 1477–1490.

Inman, R.P., and M.A. Fitts. 1990. Political institutions and public policy: Evidence from the U.S. historical record. *Journal of Law, Economics, and Organization* 6: 79–132.

Landbein, L.I., P. Crewson, and C.N. Brasher. 1996. Rethinking ward and at-large elections in cities. *Public Choice* 88: 275–293.

McKenzie, E. 1996. Homeowner association private governments in the American political system. *Papers in Political Economy*, 75, University of Western Ontario.

Mitchell, J. 2001. Business improvement districts and the new revitalization of downtowns. *Economic Development Quarterly* 15: 115–123.

Pack, J.R. 1992. BIDs, DIDs, SIDs, SADs: Private government in urban America. *The Brookings Review* 10: 18–21.

Persson, T., and G. Tabellini. 2000. *Political economics*. Cambridge, MA: MIT Press.

Persson, T., G. Roland, and G. Tabellini. 1997. Separation of powers and political accountability. *Quarterly Journal of Economics* 112: 1163–1202.

Persson, T., G. Roland, and G. Tabellini. 1998. Towards micropolitical foundations of public finance. *European Economic Review* 42: 685–694.
Persson, T., G. Roland, and G. Tabellini. 2000. Comparative politics and public finance. *Journal of Political Economy* 108: 1121–1161.
Rauch, J.E. 1995. Bureaucracy, infrastructure, and economic growth: Evidence from U.S. cities during the progressive era. *American Economic Review* 85: 968–979.
Riker, W.H. 1962. *The theory of political coalitions*. New Haven: Yale University Press.
Rubinfeld, D.L. 1987. The economics of the local public sector. In *Handbook of public economics*, vol. 2, ed. A.J. Auerbach and M. Feldstein. Amsterdam: North-Holland.
Shepsle, K.A., and B.R. Weingast. 1984. Political solutions to market problems. *American Political Science Review* 78: 417–434.
Tiebout, C. 1956. A pure theory of local expenditure. *Journal of Political Economy* 64: 416–424.
Weingast, B.R., K.A. Shepsle, and C. Johnsen. 1981. The political economy of benefits and costs: A neoclassical approach to distributive politics. *Journal of Political Economy* 89: 642–664.

Urban Production Externalities

Antonio Ciccone

Abstract

Urban production externalities (agglomeration effects) are external effects among producers in areas with a high density of economic activity. Such external effects are the main explanation for why productivity is usually highest in those areas of a country where economic activity is densest. There is some disagreement about the strength of urban production externalities. What is clear, however, is that even weak urban production externalities can explain large spatial differences in productivity.

Keywords

Congestion; Increasing returns; Instrumental variables; Intermediate inputs; Knowledge spillovers; Labour productivity; Local technological externalities; Localization economies; Mincerian wage regression; Non-transportable input sharing; Non-transportable intermediate inputs; Outsourcing; Schooling externalities; Spatial externalities; Skill-biased technical change; Urban agglomeration; Urban production externalities; Urbanization economies; Spatial wage differentials

JEL Classifications

R0

Urban production externalities (agglomeration effects) are external effects among producers in areas with a high density of economic activity. Such external effects are the main explanation for why productivity is usually highest in those areas of a country where economic activity is densest. The best understood urban production externalities are technological externalities due to knowledge spillovers and non-transportable input sharing, both of which are already discussed by Marshall (1920).

That local technological externalities translate into increasing returns at the city level is demonstrated formally by Henderson (1974). Building on the analysis of Chipman (1970), he also shows that such increasing returns are compatible with perfect competition. Abdel-Rahman (1988), Fujita (1988, 1989), and Rivera-Batiz (1988) present a rigorous analysis of decentralized market equilibria with increasing returns at the city level due to intermediate input sharing. These contributions build on the formalization of monopolistic competition of Spence (1976) and Dixit and Stiglitz (1977) to show how increasing returns to city size emerge when some intermediate inputs are non-transportable and produced subject to increasing returns at the plant level.

There is some disagreement about the strength of increasing returns to the density (or scale) of local economic activity and therefore about the strength of urban production externalities. This is partly because the best approach to estimation is still unclear. What is clear, however, is that even weak urban production externalities can explain much of the large spatial differences in productivity observed in many countries. This is because spatial differences in the density of economic

activity are very large, so that even a small degree of increasing returns to density can explain sizable spatial productivity differences. Moreover, mobile physical capital and tradable intermediate inputs imply that the strength of increasing returns to density exceeds the strength of urban production externalities (by approximately a factor of two).

The remainder of this article first illustrates the link between the strength of urban production externalities and the strength of increasing returns to the density of economic activity (or increasing returns at the city level). It then turns to the advantages and drawbacks of different empirical approaches to urban production externalities.

A Model of Urban Production Externalities and Increasing Returns

The link between urban production externalities and increasing returns to the density of economic activity is easily illustrated using the technology-spillover model of Ciccone and Hall (1996) extended to include costlessly tradable intermediate inputs. This extension is important for understanding why the strength of increasing returns to density is approximately twice the strength of urban production externalities. Including urban production externalities due to non-transportable intermediate-input sharing in the model would be straightforward (see Ciccone and Hall 1996) but not change any of the relevant conclusions.

Model Set-Up

Let $f(n_f, k_f, m_f; Q_c, A_c)$ be the production function that describes the amount of output produced by firm f on an acre of space when employing n workers, m units of costlessly tradable intermediate inputs, and k units of capital (lower-case letters denote per-acre quantities). The acre is embedded in county c with total output Q and total acreage A (upper-case letters denote total county-level quantities). The simplest production function to deal with is one where the externality depends multiplicatively on the density of economic activity Q/A, and the elasticity of $f(n, k, m; Q, A)$ with respect to all its arguments is constant. In this case,

$$qf = \left(a_f^{\alpha} k_f^{\beta} m_f^{1-\alpha-\beta}\right)^{1-\rho} \left(\frac{Q_c}{A_c}\right)^{\lambda} \qquad (1)$$

$\lambda \geq 0$ captures the strength of urban production externalities (agglomeration effects); for example, $\lambda = 2\%$ means that a doubling of the density of economic activity is associated with a two per cent increase in the output of the firm (for a given amount of inputs used by the firm). $0 \leq \alpha \leq 1$ and $0 \leq \beta \leq 1$ determine the relative importance of labour, capital and intermediate inputs in production. And $0 \leq \rho \leq 1$ captures possible decreasing returns to labour, capital and intermediate inputs when holding the amount of land used in production constant (congestion effects).

Input Demand and Value Added

Firms maximize profits taking factor prices and aggregate output in each county as given. Profit maximization implies that firms employ capital up to the point where its marginal product is equal to the national rental price of capital R (measured in units of output), which gives rise to a demand for capital equal to $k_f = \beta(1-\rho)q_f/R$. The demand for intermediate inputs can be determined analogously as $m_f = (1-\alpha-\beta)(1-\rho)q_f$, where we have assumed that one unit of intermediate inputs can be produced with one unit of output. Substituting these factor demand functions in (1) and solving for output at the firm level yields that q_f is proportional to $(n_f)^{\alpha(1-\rho)/(1-(1-\rho)(1-\alpha))} (Q_c/A_c)^{\lambda/(1-(1-\rho)(1-\alpha))}$. Moreover, the demand for intermediate inputs implies that value added at the firm level (y) and county level (Y) are a fraction $1-(1-\alpha-\beta)(1-\rho)$ of the total value of production at the firm and county level respectively, that is, $y_f = q_f - m_f = (1-(1-\alpha-\beta)(1-\rho))q_f$ and $Y_c = Q_c - M_c = (1-(1-\alpha-\beta)(1-\rho))Q_c$. Hence, value added at the firm level is linked to firm-level employment and county-level value added by

$$y_f = (\gamma n_f)^{\frac{\alpha(1-\rho)}{1-(1-\rho)(1-\alpha)}} \left(\frac{Y_c}{A_c}\right)^{\frac{\lambda}{1-(1-\rho)(1-\alpha)}}, \qquad (2)$$

where γ is an unimportant constant.

Increasing Returns to Density

The amount of labour N employed in a county is taken to be distributed uniformly in space; $n_f = N_c/A_c$ for all firms f in the county. Substituted in (2), this yields that output per acre in a county is linked to employment per acre by

$$\frac{Y_c}{A_c} = \left(\gamma \frac{N_c}{A_c}\right)^{1+\theta}, \tag{3}$$

where the strength of increasing returns to the density of economic activity θ is given by

$$\theta = \frac{\lambda}{\alpha\,(1-\rho) - (1-\rho)}. \tag{4}$$

As expected, increasing returns to density are stronger when urban production externalities λ are strong and congestion effects ρ are weak. A necessary condition for productivity to be greater in areas with dense economic activity is that urban production externalities (agglomeration effects) more than offset congestion effects, $\theta > 0$.

From Increasing Returns to Density to Urban Production Externalities

The relationship between increasing returns to the density of economic activity θ and the strength of net agglomeration effects $\lambda - \rho$ in (4) depends on $\alpha(1 - \rho)$, the elasticity of output with respect to labour. In equilibrium, this elasticity equals the share of labour in the total value of production. In the United States, the share of labour in value added is around two thirds (for example, Gollin 2002) and the share of intermediate inputs in value added around one half (for example, Basu 1995), which implies a share of labour in the total value of production of around one third. To see that this implies that urban production externalities are magnified, note that for small values of $\lambda - \rho$ (4) implies

$$\theta \cong \frac{\lambda - \rho}{\alpha\,(1-\rho)} = (\lambda - \rho), \tag{5}$$

where we have made use of $\alpha(1 - \rho) = 1/3$. A one-point increase in the strength of urban production externalities therefore implies a three-point increase in the strength of increasing returns to the density of economic activity. Much of this magnification is due to the presence of intermediate inputs. In a model without intermediate inputs where physical capital earns one third of value added, the factor of magnification would have been (only) 3/2.

Empirical Approaches and Results

Increasing Returns to City or Industry Size

Early empirical studies of urban scale effects by Sveikauskas (1975), Segal (1976), Moomaw (1981, 1985), and Tabuchi (1986) focused on the link between city size and productivity at the city and the city-industry level. The empirical results indicate that doubling city size increases productivity by between three and eight per cent. Nakamura (1985) and Henderson (1986, 2003) extend the analysis by distinguishing between urbanization economies and localization economies. Localization economies are increasing returns related to the size of city industries, while urbanization economies refer to increasing returns to overall city size. Henderson concludes that scale effects are mostly at the industry level, but Nakamura finds evidence for both urbanization and localization economies.

Most studies of the strength of agglomeration economies measure output as the value of production or value added from the U.S. Census Bureau's Census of Manufacturers. This data-set does not contain information on the value of services that plants purchase in the market or obtain from headquarters. Census of Manufacturers data will therefore overstate the value added of city industries. This bias is likely to be greater in larger cities, for two reasons. First, there is more service outsourcing in larger cities, due to the larger variety of services available (Holmes 1999; Ono 2007). Second, headquarter services are more likely to be counted twice in larger cities, as such cities are more likely to contain both a plant

and its headquarters. The total value of production from the Census of Manufacturers has the additional disadvantage of counting twice all intermediate inputs traded within and across industries located in the same city.

Increasing Returns to Density and the Productivity of US States

In the United States, the finest level of geographical detail with reliable data on value added is the state level. Ciccone and Hall (1996) therefore estimate increasing returns to the density of economic activity by combining state-level value added data with the model in (3). Aggregating county-level value added to the state level yields that labour productivity in state s, Y_s/N_s, is equal to

$$\frac{Y_s}{N_s} = D_c(\theta) \equiv \sum_{c=1}^{C_s} \left(\frac{\gamma N_c}{A_c}\right)^{1+\theta} \frac{N_c}{N_s}, \qquad (6)$$

where C_s is the number of counties in the state. Hence, the strength of increasing returns to the county-level density of economic activity can be estimated by combining cross-state variation in labour productivity and the state-level density index $D_c(\theta)$, which depends on county-level employment density and the distribution of employment across counties. Ciccone and Hall find θ to be just above five per cent, using a least-squares approach. Because of large differences in the density of economic activity, this limited degree of increasing returns to density can explain more than half of the sizable differences in output per worker across US states.

Ciccone and Hall's work is about the degree of increasing returns to the density of economic activity, not about the strength of urban production externalities. Going from one to the other is rather straightforward, however. Using (5) yields that θ equal to five per cent corresponds to a net agglomeration effect $\lambda - \rho$ of 1.7 per cent. According to the *Flow of Funds Accounts of the United States, 1982–1990* prepared by the Board of Governors of the Federal Reserve System (1997), the share of land in the total value of production ρ in the private sector outside of agriculture and mining is around 0.5 per cent. Hence, λ is between 2 and 2.5 per cent, which implies that a doubling of the density of economic activity in a county is associated with a 2–2.5 per cent increase in the output firms produce with a given amount of inputs (see (1)). Mobile physical capital and tradable intermediate inputs therefore imply that the strength of increasing returns to density exceeds the strength of urban production externalities by a factor of two. Hence, more than half of the differences in output per worker across US states can be explained by rather weak urban production externalities.

An important concern when estimating agglomeration economies is potential feedback from productivity to the density of economic activity. To address this possibility, Ciccone and Hall (1996) use an instrumental variables approach. The instruments for the state-level density index used are the population and population density of US states between 1850 and 1880, as well as the presence or absence of a railroad in each state in 1860 and the distance of states from the eastern seaboard. The identifying assumption is that the original sources of agglomeration in the United States have remaining influences only through the preferences of people about where to live. The instrumental variables estimates of θ are between 5.5 and 6.1 per cent, and therefore very similar to the least squares estimates.

Agglomeration Effects in Europe

For many European countries, value added data is available at a much finer level of geographic detail than for the United States. Employing such data for France, Germany, Italy, Spain and the UK, Ciccone (2002) finds an average degree of increasing returns to the local density of economic activity of between four and five per cent, only slightly below estimates for the United States. Rice et al. (2006) find a similar result using geographically detailed earnings data for the UK. They also take into account the scale of production in neighbouring locations weighted by travel times, and find that productivity benefits diminish quickly with travel distance.

Human Capital Externalities?

An open question is whether there are agglomeration economies associated with the geographic

concentration of human capital. Rauch (1993) examines this issue by augmenting a standard Mincerian wage regression (for example, Card 1999) with data on the characteristics of cities where people live. His empirical model relates wages of individuals i in cities c, w_{ic} to relevant individual characteristics (for example, education, experience), X_{ic}, and to the average level of schooling of the city, S_c, and other city characteristics, Z_c,

$$\log w_{ic} = aX_{ic} + bS_c + cZ_c + \varepsilon_{ic}, \qquad (7)$$

where ε_{ic} summarizes all other (unobserved) factors affecting individual wages across cities. Least-squares estimation of (7) using US data for 1980 yields a positive and significant coefficient on average schooling in the city (b), and Rauch therefore concludes that there are human capital externalities at the city level.

A drawback of Rauch's approach is that it cannot account for time-invariant unobserved city characteristics that increase both schooling and wages. Another drawback is that city-level schooling is taken to be exogenous. Acemoglu and Angrist (2001) address these drawbacks by taking US states, rather than cities, to be the relevant aggregate unit in (7). In this case, the data allow for an analysis of the effects of increases in average state-level schooling on individual wages. Moreover, Acemoglu and Angrist show that changes in average schooling at the state level can be instrumented by state-level changes in compulsory-schooling and child-labour laws. Their approach yields no evidence of significant schooling externalities between 1960 and 1980.

Ciccone and Peri (2006) show that a positive effect of average schooling in a Mincerian wage equation like (7) may not reflect human capital externalities but a downward sloping demand function for human capital. They therefore propose an alternative approach, which exploits the fact that the wage differential between more and less educated workers reflects differences in marginal social products of the two worker types when human capital externalities are absent. This approach does not yield evidence of significant human capital externalities at the level of US cities or states between 1960 and 1990.

Moretti (2004a) finds that US cities where the labour force share of college graduates increased most between 1980 and 1990 also saw the largest wage increase for college graduates. Using Census of Manufacturers plant-level data, Moretti (2004b) finds that the output of plants in high-tech city industries rises with levels of schooling in other high-tech industries in the same city. This evidence is consistent with human capital externalities. An alternative explanation could be that skill-biased technological progress translated into increases in the productivity and wages of college graduates in high-tech industries. Cities that specialized in industries experiencing rapid productivity growth would in this case see faster output growth and attract more college graduates. This alternative hypothesis is especially plausible for the 1980–90 period, which was marked by rising college wage premia due to skill-biased technological progress (for example, Katz and Murphy 1992).

See Also

- ▶ Externalities
- ▶ New Economic Geography
- ▶ Urban Agglomeration

Bibliography

Abdel-Rahman, H.M. 1988. Product differentiation, monopolistic competition and city size. *Regional Science and Urban Economics* 18: 69–86.

Acemoglu, D., and J. Angrist. 2001. How large are the social returns to education: Evidence from compulsory schooling laws. In *NBER macroeconomic annual 2000*, ed. B. Bernanke and K. Rogoff. Cambridge, MA: MIT Press.

Basu, S. 1995. Intermediate goods and business cycles: Implications for productivity and welfare. *American Economic Review* 85: 512–531.

Board of Governors of the Federal Reserve System. 1997. *Flow of funds accounts of the United States, 1982–1990*. Washington, DC: Federal Reserve.

Card, D. 1999. The causal effect of education on earnings. In *Handbook of labor economics*, ed. O. Ashenfelter and D. Card. Amsterdam: North-Holland.

Chipman, J.S. 1970. External economies of scale and competitive equilibrium. *Quarterly Journal of Economics* 84: 347–385.
Ciccone, A. 2002. Agglomeration effects in Europe. *European Economic Review* 46: 213–227.
Ciccone, A., and R.E. Hall. 1996. Productivity and the density of economic activity. *American Economic Review* 86: 54–70.
Ciccone, A., and G. Peri. 2006. Identifying human capital externalities: Theory with applications. *Review of Economic Studies* 73: 381–412.
Dixit, A.K., and J.E. Stiglitz. 1977. Monopolistic competition and optimum product diversity. *American Economic Review* 67: 297–308.
Fujita, M. 1988. A monopolistic competition model of spatial agglomeration: Differentiated product approach. *Regional Science and Urban Economics* 18: 87–124.
Fujita, M. 1989. *Urban economic theory: Land use and city size*. Cambridge: Cambridge University Press.
Gollin, D. 2002. Getting income shares right. *Journal of Political Economy* 110: 458–474.
Henderson, J.V. 1974. The sizes and types of cities. *American Economic Review* 64: 640–656.
Henderson, J.V. 1986. Efficiency of resource usage and city size. *Journal of Urban Economics* 19: 47–70.
Henderson, J.V. 2003. Marshall's scale economies. *Journal of Urban Economics* 53: 1–28.
Holmes, T. 1999. Localization of industry and vertical disintegration. *Review of Economics and Statistics* 81: 314–325.
Katz, L.F., and K.M. Murphy. 1992. Changes in relative wages, 1963–1987: Supply and demand factors. *Quarterly Journal of Economics* 107: 35–78.
Marshall, A. 1920. *Principles of economics*. 8th ed. London: Macmillan.
Moomaw, R.L. 1981. Productivity and city size: A critique of the evidence. *Quarterly Journal of Economics* 96: 675–688.
Moomaw, R.L. 1985. Firm location and city size: Reduced productivity advantages as a factor in the decline of manufacturing in urban areas. *Journal of Urban Economics* 17: 73–89.
Moretti, E. 2004a. Estimating the social return to higher education: Evidence from longitudinal and repeated cross-sectional data. *Journal of Econometrics* 121: 175–212.
Moretti, E. 2004b. Workers education, spillovers, and productivity: Evidence from plant-level production functions. *American Economic Review* 94: 656–690.
Nakamura, R. 1985. Agglomeration economies in urban manufacturing industries: A case of Japanese cities. *Journal of Urban Economics* 17: 108–124.
Ono, Y. 2007. Market thickness and outsourcing services. *Regional Science and Urban Economics* 37: 220–238.
Rauch, J.E. 1993. Productivity gains from geographic concentration of human capital: Evidence from the cities. *Journal of Urban Economics* 34: 380–400.
Rice, P., A.J. Venables, and E. Patacchini. 2006. Spatial determinants of productivity: Analysis for the regions of Great Britain. *Regional Science and Urban Economics* 36: 727–752.
Rivera-Batiz, F.L. 1988. Increasing returns, monopolistic competition, and agglomeration economies in consumption and production. *Regional Science and Urban Economics* 18: 125–153.
Segal, D. 1976. Are there returns to scale in city size? *Review of Economics and Statistics* 58: 339–350.
Spence, M. 1976. Product selection, fixed costs, and monopolistic competition. *Review of Economic Studies* 43: 217–235.
Sveikauskas, L. 1975. The productivity of cities. *Quarterly Journal of Economics* 89: 393–413.
Tabuchi, T. 1986. Urban agglomeration, capital augmenting technology, and labor market equilibrium. *Journal of Urban Economics* 20: 211–228.

Urban Transportation Economics

Alex Anas

Abstract

Advances by economists in understanding the demand, capacity and supply, pricing, finance and performance of urban transportation systems is reviewed. The economics of urban transportation has emphasized externalities such as traffic congestion. The effects of transport systems and travel behaviour on real estate prices, urban land use and density and urban expansion as well as the reciprocal effects of urban form on the nature and utilization of transport systems are studied by economists.

Keywords

Congestion; Congestion tolls; Derived demand; George, H.; Labour market discrimination; Land markets; Land taxes; Land use; Land values; Land-use zoning; Pigouvian taxes; Residence location choice; Samuelson, P.; Spatial mismatch; Static assignment models; Stationary state models; Transport externalities; Transport system performance; Urban economics; Urban sprawl; Urban

transportation economics; Value of time; Work location choice

JEL Classifications
R4; R14; H23; H41; H54

The study of transportation in urban areas relates to urban economics and to public economics and finance. The development of cities and their land use patterns cannot be understood without studying the transportation systems that shape them, nor can urban transportation systems be understood independently of the urban economy.

Unique aspects of urban transportation economics relate to demand, capacity and supply, the performance of urban transportation systems, and pricing and finance. We provide a discussion of the key conceptual issues and knowledge in each of these areas of the field and point out some challenges that remain. (For reviews of transportation economics focused less on its relationship to urban economics and more on technical issues internal to transportation, see Arnott and Kraus 2003; Small and Verhoef 2006.)

Demand

The demand for transport is 'derived demand'. Travel provides utility mostly because it is a means to an end, be it a consumer purchase, getting to work or to recreation. The travel itself usually has a disutility which varies according to the quality, reliability and safety of the transport system or the particular trip. Hence, virtually all transport choices involve a trade-off between the inconvenience and cost of a trip on the one hand and the frequency with which that trip is chosen relative to other trips on the other.

Beginning with the emergence of the telephone, the demands for travel and for communication have become increasingly interlinked in an urban setting. While travel and communication are substitutes because a phone call, fax or e-mail (or a messenger or letter in the pre-telephone days) may reduce the need for a trip, they are also complements because cheaper communication generates higher demand for goods, services and personal contacts. From this higher demand more travel is subsequently derived.

An important aspect of urban travel is the fact that the out-of-pocket cost of travel can be low relative to the value of time expended in that travel. As such, travel competes with leisure and with work as a key activity to which time must be allocated. The dominance of time–cost means that market prices are less important than full opportunity costs in the explanation and measurement of travel behaviour. Values of time vary greatly among consumers since wage rates vary but also because of other factors. Thus, consumers who undertake similar trips frequently incur vastly different opportunity costs.

The demand for urban travel by consumers is derived from a complex set of hierarchically linked choices. At the top of the hierarchy and slowest to change are decisions relating to where to work and where to reside. Lower in the hierarchy and more malleable are choices about the number and type of cars to own including the possibility of dispensing with cars and relying on walking or public transport for some trips. Yet lower on the hierarchy are choices about the destinations and frequency of discretionary trips, the frequency of commuting (to the extent that work arrangements do not require daily commuting), the destination of the commute being implicit in the residence–workplace choice, and the mode of transport (private automobile, taxi, public transit or walking) that will be used on each such round trip. There is the choice of the time of day during which particular trips are made and the trip chaining decision about whether several trips may be combined into a tour. In a tour, the trips are linked sequentially, beginning and ending at the point of origin (for example, the consumer's home). Finally, also important are the choices of particular routes (of the highway network, for example) on which trips occur. Firms make a similar set of choices. At the top of the hierarchy is the location of the plant vis-à-vis suppliers and markets, followed by the choice of a vehicle fleet and associated decisions of the modes (barge, rail, truck and so on) to use to get products to market or procure them from suppliers.

The study of travel demand using econometric techniques has not yet advanced to the point where a unified theory of travel can be tested that deals simultaneously with all or even most of the levels in the hierarchy. Led by McFadden (1973), transport economists have mostly focused on mode choice: the split of the demand for travel between competing modes such as auto, urban rail or bus and car pooling on a particular trip or round trip. This has resulted in the widespread use of a rich variety of discrete choice models (such as logit, nested logit or probit) that are designed to predict the probability that a randomly selected traveller will choose a particular mode to commute to work. Modelling the choice of residence or job locations has not received much attention from transportation economists. Virtually all studies in this area can be found in the urban economics literature, and some have emphasized the joint choice of residence location and mode of commuting (Anas and Duann 1985). Most travel demand studies focus exclusively on commuting, ignoring the fact that discretionary non-work travel is continually increasing with incomes, car ownership and suburbanization. And the trade off between commuting and discretionary trip-making under a time budget constraint has remained relatively unexamined.

Another area of demand that has received attention is the choice of travel route on a congested highway network. The work of Beckmann et al. (1956) counterfactually conceived traffic on a network as a stationary state process of steady flow, rather than as a system of queues and bottlenecks causing complex flow dynamics. Despite this, the simplicity of the model led to the prolific development of static assignment models by operations researchers. These models use system optimization principles to simulate how travellers choose the least costly route on a congested network. Stochastic cost perceptions have been introduced into this type of stationary state models (Daganzo and Sheffi 1977). More recently, a variety of dynamic simulation models that recognize the queue and bottleneck nature of traffic are under development based on the principle that travellers choose not only a route but also departure–arrival times (Arnott et al. 1990).

Capacity and Supply

One aspect of supply is that most transport is made possible in large part by consumer effort, time, and by inputs purchased by the consumer such as car, gasoline, vehicle maintenance and garage. Viewed this way supply becomes virtually inseparable from demand. As such it would make sense to model a part of the supply decision within a household production context.

Another aspect of supply is that the public sector is involved in the planning, provision and operation of most travel infrastructure. This includes highways as well as buses and urban rail. A key decision variable is capacity, measured as the throughput of passengers per hour that can be transported in a particular direction at a given time during the day on a particular facility. This throughput determines the user's travel time. Also important, however, are the safety, privacy, reliability and quality of the travel time and its components such as in-vehicle time, waiting at a station, searching for parking and time walking to and from stations and parking lots.

The key supply decision is the quantity of streets and highways and public transit rights-of-way. Road capacity relative to demand determines, in part, the level of traffic congestion in an area. Since land is the prime input in roads, more road building reduces the land available for other uses such as housing or production, raising the market price of land in such uses. In turn, the price of land chiefly determines how much land is allocated to create road capacity in an area. Thus, the most congested areas are also the ones where land is the most expensive. With extremely expensive land as in Tokyo, London, Paris or downtown New York, the substitution of capital for land results in tunnelling for transit systems (subways) and even for some roads.

An important supply question is whether economies of scale exist in congested highway traffic flow. Congestion occurs when the vehicles sharing the same road segment at the same time reach a critical value relative to road capacity. The addition of one more vehicle then begins to delay the other vehicles. The total cost experienced by the vehicle stream increases by a marginal cost that is

higher than the cost privately born by the marginal vehicle's passengers. The difference between this social marginal cost and the private average cost is the monetary value of the sum of delays the marginal vehicle imposes on all the vehicles travelling with it on the road segment. The evidence seems to suggest that this congestion process exhibits no economies of scale at least at a crude level. Scaling up (or down) road capacity and the volume of traffic in the same proportion, would not increase the average cost of travel. Capital costs of highways, on the other hand, were found to exhibit significant scale economies by the engineering estimates of Meyer et al. (1965), but since then Keeler and Small (1977) and others have found statistical evidence of weak or virtually no scale economies.

In contrast to highways, rail-based public transit systems are subject to scale economies and, more importantly, to economies of density. As more passengers use these systems (for example by reducing headways between successive trains), the per-passenger total average cost comes down because of the high fixed costs involved in system construction and maintenance. This is the chief reason why such rail infrastructure is uneconomical in US cities below some critical size such as one million or more people, or in suburban areas of low land use densities where the passengers' time-costs of accessing transit stations can be high (Kim 1979).

System Performance

The reconciliation of supply and demand results in system performance. Unlike other markets in which price is the only salient outcome of market performance, in transport the outcomes include travel time, the level of congestion or travel delay, air pollution from car traffic, accidents, system reliability (that is, the variability of travel time from day to day or hour to hour), and pecuniary and non-pecuniary externalities caused by the transport system. While travel time, congestion and reliability costs are primarily born by the travellers, air pollution and accidents have costs that are born by travellers as well as by non-travellers. Thus, the economic performance of a transport system cannot be measured completely without evaluating the social costs and benefits created by these external effects.

The purely pecuniary externalities of transport are pervasive. For example, as noted, the creation of transport capacity has a direct effect on the supply and price of land available for other uses and can thus cause land scarcity. But this is only the direct effect of capacity provision. The indirect effect on land values and land use is quite different and works at both the extensive and intensive margins. At the extensive margin, cities endowed with more road and transit capacity can expand to land areas that were previously inaccessible. At the intensive margin, transport systems work by changing the *relative* accessibility of land parcels. Areas that are made relatively more accessible than before gain value, while areas made relatively less accessible lose value. As a result of these shifts, windfall gains and losses in land markets should be among the chief measures of transport system performance evaluation. The aggregate change in land values can be positive or negative. Since most land is owned by consumers (such as homeowners) transport system changes play an important role in redistributing private wealth and public revenues from *ad valorem* property taxes by changing an existing pattern of accessibility.

Transportation, land use and land prices are the central foci in urban economics and a variety of models have been developed. Virtually all of these assume that all jobs are located at a predetermined centre, an anachronism given that current downtowns in US cities contain no more than ten per cent of the jobs. Versions of this basic model based on linear programming have been developed to model road capacity provision and transit investment in congested cities (Mills 1972; Kim 1979).

Other pecuniary externalities centre on the improved discretionary mobility enabled by transport systems. Such mobility improvements have received praise as well as criticism. Improved mobility enables easier, cheaper and more frequent contacts among firms and among firms and consumers. This should result in positive social benefits enhancing productivity and boosting economic

growth. It has been noted in the large literature on *spatial mismatch* that central-city minorities in the United States who are less-mobility enabled, are at a disadvantage competing for suburban employment. While discrimination and suburban land use exclusion cause minorities to be clustered and socially cloistered in central cities, lower car ownership may also hinder their ability to compete for distant suburban jobs.

Improved mobility induces economic agents to locate in a more spread out pattern, substituting cheaper outlying land for more expensive, centrally located land. The resulting urban land use pattern, common in the United States, has been dubbed 'urban sprawl'. Sprawl has been blamed for a variety of ills stemming from the increased dependence on cars and reduced pedestrian mobility that sprawled land use promotes. Among such perceived ills, for example, is the alleged demise of social and neighbourhood cohesion and the rising obesity of American children and adults.

Pricing and Finance

In practice, urban roads and transit systems are subsidized. In the United States a large part of the cost of highways and roads comes from general income taxes. The rest of the cost comes from taxes on gasoline and taxes on real property. Urban rail systems are also heavily subsidized with fares covering only about half of the operating and maintenance costs. Hence, for all forms of urban transport with the possible exception of unregulated taxis and jitneys, market-based user fees and marginal cost prices do not play the role they do in other markets.

What does economic theory tell us about how urban transport systems should be priced and financed? The answer will be different for highway and rail systems, primarily because the latter are subject to economies of scale.

The congestion externality is key in highway pricing and investment (Vickrey 1969). Economic efficiency requires that each traveller pay his full marginal social cost on each road segment that he uses. As we saw earlier, the full marginal social cost includes the monetary value of the delay each traveller imposes on his cotravellers. This is higher where congestion is high, falling to zero where congestion is not present. It has been shown that if congestion tolls can be properly calculated and levied on travellers, then with no economies of scale in roads, the tolls collected from the vehicles using a particular road segment would in the long run cover the amortized costs of road construction and maintenance. The only requirement is for road planners to build more (less) road capacity where toll revenue exceeds (falls short) of these amortized costs.

The congestion toll has three coincident theoretical interpretations. First, it is a Pigouvian tax (Pigou 1947) because it levies, on the source of a negative externality, a tax that closes the gap between the social marginal cost and the private average cost. In this role, the toll causes travellers to economize on travel by internalizing the negative externality they create. Second, because a road can be viewed as a (congested) public good, the congestion toll in the long run serves to equate the marginal benefit of road capacity with the marginal cost of supplying it, the Samuelson rule for the optimal finance of a public good (Samuelson 1954). The toll itself is a marginal benefit since it measures the reduction in total travel cost that would be realized if one more unit of road capacity were to be added, while the marginal cost of the capacity is the cost of purchasing the additional capacity. Third, since the aggregate toll revenue from the road segment is equal to the land rent the road would fetch in an alternative use, the aggregate toll is equivalent to a confiscatory tax on the owners of the land, the Henry George rule (George 1879). On the view that the land used for roads is privately owned and operated by competitive or contestable firms, the Pigouvian pricing described above would be the outcome of profit maximization, and the aggregate toll revenue would confiscate the profits of these private road owners. On the alternative view that the land used for roads is owned by society, the congestion tolls are the fees travellers pay society for the right to use the road, and in the long run these fees add up to the rent on land, provided land markets are competitive.

Keeler and Small (1977) empirically estimated what congestion tolls should be in the San Francisco Bay Area on the assumption of fixed land use. The effects of tolls on urban form have been studied within the naïve theoretical urban model that assumes all jobs are at a central point (Arnott and MacKinnon 1978) or a central point and a suburban ring (Sullivan 1983). Simple simulations based on such models show small efficiency gains of up to one per cent of income for reasonably congested cities. Recent studies, based on modern assumptions of completely dispersed employment, show similar efficiency gains (Anas and Xu 1999). All of these studies show that congestion tolls could significantly reduce travel times. But the welfare benefits of tolls would come mostly from changes in travel mode and the timing of travel during the day, rather than from land use adjustments.

Congestion tolls have become more popular in recent years and have seen such prominent implementation as in central London. But the correct calculation of first-best tolls is a quagmire. Chief among the difficulties is the fact that one must know how the value of travel time is distributed among travellers using the same road segment. If I share the road with higher (lower) income drivers, the toll on me should be higher (lower). Without knowledge of the distribution, accurate first-best tolls cannot be computed because values of time vary so widely among people. A second difficulty is that road use varies enormously throughout the day, requiring that first-best tolls should similarly vary. The problem is simplified somewhat by dividing the day into peak and off-peak periods. A third difficulty is that the technology used to detect congestion and calculate tolls should not be so obtrusive on travel as to create more congestion than the tolls would alleviate. Automatic vehicle identification by several means is feasible and not expensive. This may contribute to a wider use of tolls in the future.

Although the calculation of first-best tolls is highly daunting, a number of second-bests are available. Tolls levied on major roads but not on local roads may be effective second-bests. A tax on the market price of parking in heavily congested destinations such as the downtowns of major cities would achieve some of the efficiencies of first-best tolls. Taxes on gasoline are not nearly as effective, because gasoline usage is not closely related to the congestion created on a trip. Such taxes heavily penalize driving on congestion-free roads, for example.

Unlike highways, rail transit should be priced as a regulated natural monopoly. Since marginal cost is below average cost at any scale, marginal cost pricing ensures efficiency but requires a subsidy to the transit operator to cover fixed costs. Thus for transit systems, theory tells us that fares should be set to cover variable operating costs, while other taxes should be used to purchase the fixed inputs, including land (right-of-way). The debate then, should be about what these other taxes should be. Considerable evidence exists showing that land around transit stations appreciates in value after a transit investment is announced or constructed. Anas and Duann (1985) used an empirically estimated general equilibrium model to predict prior to construction that residential property values around the proposed stations of the Chicago Midway line would increase, with the increase sharply tapering off with distance from the stations. They estimated that the aggregate increase could pay for about 40 per cent of the construction cost. McMillen and McDonald (2004) used *ex post* data on housing sales and confirmed that these predictions were accurate. Taxing such windfall gains is one source of revenue for fixed facilities, although there are practical complications about how to accurately measure and document the land value appreciation in a legal-administrative context.

Transportation as a Tool to Shape Land Use

It has been observed that the underpricing of road travel, especially as it relates to the unpriced congestion externality makes travel cheaper than its marginal cost. This not only causes excessive urban expansion but also induces planners to use faulty cost-benefit measures and thus invest in too much road building as argued by

Kraus et al. (1976). Excessive road capacity in turn reinforces the excessive urban expansion.

In view of the many pecuniary externalities of transportation, and since perfect pricing is not possible, a combination of judicious capacity provision and land-use zoning to ensure better accessibility to main roads and rail lines could have significant benefits. Such economies of transport–land use interdependence may be possible to exploit in urban planning and urban design at the level of smaller areas and neighbourhoods. Boarnet and Crane (2000) have examined whether land use policy and urban form can significantly affect travel behaviour in such settings. Similar concerns exist at the macro urban level (Gordon et al. 1989). In the future, planners could use such knowledge when major decisions are made on how much capacity to supply, where to supply it and how much to restrict development around it. More often than not, however, when urban planners intervene with land use controls they may fail to find the golden rule, causing distortions in land markets that could outweigh the efficiencies that can be gained by influencing travel.

See Also

- ▶ Coase Theorem
- ▶ Congestion
- ▶ External Economies
- ▶ George, Henry (1839–1897)
- ▶ Land Markets
- ▶ Pigouvian Taxes
- ▶ Time Use
- ▶ Urban Economics
- ▶ Value of Time

Bibliography

Anas, A., and L. Duann. 1985. Dynamic forecasting of travel demand, residential location and land development: Policy simulations with the Chicago area transportation/land use analysis system. *Papers of the Regional Science Association* 56: 38–58.

Anas, A., and R. Xu. 1999. Congestion, land use and job dispersion: A general equilibrium model. *Journal of Urban Economics* 45: 451–473.

Arnott, R., and M. Kraus. 2003. Principles of transport economics. In *Handbook of transportation science*, 2nd ed, ed. R. Hall. Boston: Kluwer.

Arnott, R., and J. MacKinnon. 1978. Market and shadow land rents with congestion. *American Economic Review* 68: 588–600.

Arnott, R., A. de Palma, and R. Lindsey. 1990. Departure time and route choice for the morning commute. *Transportation Research B* 24: 209–228.

Beckmann, M., C. McGuire, and C. Winsten. 1956. *Studies in the economics of transportation*. New Haven: Yale University Press.

Boarnet, M., and R. Crane. 2000. *Travel by design: The influence of urban form on travel*. New York: Oxford University Press.

Daganzo, C., and Y. Sheffi. 1977. On stochastic models of traffic assignment. *Transportation Science* 11: 253–274.

George, H. 1879. *Progress and poverty*, , 1955. New York: Robert Shalckenbach Foundation.

Gordon, P., A. Kumar, and H. Richardson. 1989. The influence of metropolitan spatial structure on commuting time. *Journal of Urban Economics* 26: 138–151.

Keeler, T., and K. Small. 1977. Optimal peak-load pricing, investment, and service levels on urban expressways. *Journal of Political Economy* 85: 1–25.

Kim, T. 1979. Alternative transportation modes in an urban land use model: A general equilibrium approach. *Journal of Urban Economics* 6: 197–215.

Kraus, M., H. Mohring, and T. Pinfold. 1976. The welfare costs of non-optimum pricing and investment policies for freeway transportation. *American Economic Review* 66: 532–547.

McFadden, D. 1973. Conditional logit analysis of qualitative choice behavior. In *Frontiers in econometrics*, ed. P. Zarembka. New York: Academic Press.

McMillen, D., and J. McDonald. 2004. Reaction of house prices to a new rapid transit line: Chicago's Midway line, 1983–1999. *Real Estate Economics* 32: 462–486.

Meyer, J., J. Kain, and M. Wohl. 1965. *The urban transportation problem*. Cambridge, MA: Harvard University Press.

Mills, E. 1972. Markets and efficient resource allocation in urban areas. *Swedish Journal of Economics* 74: 100–113.

Pigou, A. 1947. *A study in public finance*, 3rd ed. London: Macmillan.

Samuelson, P. 1954. The pure theory of public expenditure. *Review of Economics and Statistics* 36: 387–389.

Small, K., and E. Verhoef. 2006. *Urban transportation economics*, 2nd ed. London: Routledge.

Sullivan, A. 1983. The general equilibrium effects of congestion externalities. *Journal of Urban Economics* 14: 80–104.

Vickrey, W. 1969. Congestion theory and transport investment. *American Economic Review, Papers and Proceedings* 59: 251–260.

Urbanization

Sukkoo Kim

Abstract

Cities first arose in the Fertile Crescent a few thousand years after the discovery of agriculture. Yet the history of urbanization is not one of steady progress. Pre-industrial urbanization rose with technological advances in agriculture and transportation which fostered population growth and trade, but fell with famine and disease. Just as important, cities rose and fell with the military fortunes of city states, territorial empires and nation states. With the Industrial Revolution, urbanization rose dramatically. As population shifted out of agriculture into manufacturing and services, cities became the dominant landscape of human civilization.

Keywords

Amenities; Civilization; Division of labour; Early industrialization; Face-to-face interaction; Feudalism; Globalization; Industrial Revolution; Labour markets; Labour productivity; Marshallian externalities; Middle East; North, D; Transaction costs; Transportation costs; Transportation revolution; Urbanization

JEL Classifications

R11

The Rise of Cities in the Ancient Middle East

The first city in human history is believed to have emerged around 3200 BC in Sumer, Mesopotamia, between the Tigris and Euphrates rivers, as a consequence of the Neolithic Revolution which saw a shift in food production from hunting and gathering to agriculture based on domesticated plants and animals (Childe 1950). The emergence of cities in Sumer marked the beginning of an 'urban revolution', but the revolution was an exceedingly slow one. Since agriculture began in the Fertile Crescent around 8500 BC, the first city emerged several thousand years after the discovery of agriculture. Moreover, the emergence of cities was not unique to Mesopotamia. Interestingly, cities emerged independently in at least two other places, China and the New World, places where major domestication of plants and animals arose independently.

While it is extremely difficult to determine the causes of the emergence of cities in ancient times, scholars such as Childe (1950) and Bairoch (1988) believe that agriculture caused cities to form because it increased population growth and provided surplus food for a non-agricultural population. Since demand for food is believed to be income inelastic, an increase in income from a rise in agricultural productivity will increase demand for secondary and tertiary products (Wrigley 1987). The urban concentration of secondary and tertiary employment, such as crafts and commerce, enabled the exploitation of the division of labour, fostered technological innovations in many areas of the economy from irrigation, transportation, metallurgy to writing, and lowered the costs of coordinating long-distance trade.

Even more importantly, cities were centres of states before the rise of territorial empires and nation states. A city state was composed of a governing city and its food-producing hinterlands. It was a distinct geographical, political and economic unit of organization. By establishing a body of formal and informal rules of property rights, city states provided their citizens with the incentives to improve productivity or, more fundamentally, to acquire more knowledge (North 1981). Cities probably became centres of government because close face-to-face interactions between the ruling elite and its administration increased the efficiency of governing by lowering the costs of generating, collecting and processing information. In addition, dense, walled cities provided effective defence against raiders.

The cities that arose in Sumer, Mesopotamia, were independent city states. Archaeological

evidence indicates that there may have been as many as 15 city states by 3000 BC. A typical city state may have contained population of 25,000 with rural population of about 500,000. Hammond (1972) suggests that the impetus for city states in Sumer may have been the need to coordinate the provision of public goods such as large-scale irrigation, drainage, communal storage, and defence against other city states. Over the following millennium, the number and size of cities grew in this region, some reaching populations of 100,000 or more.

When the Sumerian cities were conquered by Babylon by 1800 BC, the era of early city states gave way to the era of territorial empires such as those of Babylon, Egypt and Canaan. Scholars generally believe that urbanization suffered under these empires except for cities, like Babylon, that served as political and military centres. Babylon at its height may have reached populations of 200,000–300,000. While the exact causes of the rise of these territorial empires are not clear, among them may have been the growing benefits of trade over longer distances and changes in military technology which allowed for the control of larger areas.

Of the major empires, it was in Egypt that cities declined most significantly. Indeed, many scholars would argue that Egypt was an empire without cities. To the extent that cities existed in Egypt, they were centres of religion and administration. The lack of cities in Egypt is often attributed to the Pharaoh's centralized control of irrigation, trade and all other facets of the economy (Hammond 1972). Cities also declined in the other territorial empires, but probably less, as many remained relatively independent. Despite the general decline of urbanization during this period, a new type of coastal city emerged in Phoenicia. These cities, which grew in numbers around 1200 BC, arose principally to trade goods throughout the Aegean and the Mediterranean. Cities first emerged in the Middle East, but reached their greatest heights in the Mediterranean in the ancient era. Most likely, cities arose later in the Mediterranean because agriculture arrived in this region a century and a half after its discovery in the Fertile Crescent. It arose first in the Fertile Crescent because of favourable geography and climate, which provided an abundance of indigenous species suitable for domestication (Diamond 1997). From there, agriculture spread about 0.7 miles per year and reached the Mediterranean region sometime between 7000 and 6000 BC. By 2000–1450 BC, Aegean civilization seems to have constructed small city states with relatively limited agricultural hinterland and trade. However, urbanization flourished under the Greek and Roman civilizations, as the Mediterranean Sea became the highway of transport and communications.

The Greeks formed small, independent city states composed of coastal cities and their adjoining farmlands between 800 and 146 BC. Their largest city, Athens, may have reached 100,000 in population, but most other cities rarely exceeded 40,000. Bairoch (1988) estimates that perhaps as much as 15–25 per cent of the Greek population lived in cities with more than 5000 inhabitants. Because the Greek soil was relatively poor, a large portion of the population was engaged in local and long-distance trade in the Mediterranean. The Greek invention of coined money facilitated market exchanges. The Greek cities are also known for their political innovations. While most city states and territorial empires, with the exception of the Phoenician city states, were ruled by monarchy or aristocracy, democracy arose in many Greek cities.

The formation of the Roman Empire between 146 BC and AD 300 represented the largest political and economic integration of territory in the ancient world. Rome, as the military and administrative centre, grew to an astonishing size, surpassing 800,000 inhabitants and perhaps reaching a million by AD 2. Unlike in the earlier territorial empires such as Egypt, cities prospered under the Roman Empire. Politically, cities became military and administrative centres and collected taxes from the surrounding countryside for Rome; economically, cities acted much like independent city states. Under conditions of peace and uniform law a great numbers of cities emerged as commercial activity increased. Bairoch (1988) estimates that about 8–15 per cent of the population resided in cities. While most cities were small, 20 or more may have reached 20,000 or more inhabitants.

When the Roman Empire disintegrated around AD 476, it signalled the decline of the Mediterranean world. In the resulting so-called Dark Ages between AD 500 and 800, urbanization fell as frequent wars and invasions contributed to economic insecurity. But the two centuries following this period were a period of urban renaissance. By this time, Europe was divided into two regions. The southern Mediterranean part was conquered by Arab Muslims while the northern part was composed of Christian Europe. In Muslim Spain, the urban population rose rapidly. In Christian Europe, urbanization grew in some places for defensive reasons, but in Italy commercial city states rose to great heights (Bairoch 1988). Ruled by merchant elites, Italian city states engaged in extensive long-distance trade using newly developed technology such as the compass, but cities like Venice also grew because their naval powers provided protection for their ships in the Adriatic Sea and beyond (Lane 1973).

The Growth of Cities in Western Europe

The period called the Middle Ages, between AD 1000 and 1500, marked the beginning of the rise of western Europe. Because this region was further away from, and possessed a very different climate from the Fertile Crescent, agriculture arrived between 6000 and 2500 BC, much later than in the Mediterranean (Diamond 1997). In the Middle Ages, technological advances such as the heavier plough, horse collar, and the three-field system increased the agricultural productivity of western Europe significantly. Between 1000 and 1340, there was a strong growth in urbanization based on feudal city states. Under feudalism, a lord provided protection using a castle and knights; in exchange for protection, slaves, serfs and free labourers offered free labour.

The rise of western Europe was interrupted between 1300 and 1490 as famine, disease and wars led to major losses of population and de-urbanization. Famines between 1315 and 1317, the plague and the Black Death between 1330 and 1340 and between 1380 and 1400, and series of wars, civil wars, feudal rebellion and banditry afflicted much of Europe and contributed to the decline of the medieval economy (Palmer and Colton 1965). Urbanization fell as feudal city states crumbled. North (1981) argues that a series of major technological advances in military warfare, such as the introduction of the pike, longbow, cannon, and eventually the musket, as well as a growing market for mercenaries, contributed to the decline of feudalism since a feudal lord could no longer provide adequate protection for his manor against these new military developments.

The modern period in history begins with the year 1500. Between 1500 and 1800, the opening of Atlantic commerce and the formation of nation states fundamentally transformed Europe and the world. Advances in ocean shipping, which led to the discovery of maritime routes to the Americas and Asia, ushered in a new era of international trade. Although cities in the Mediterranean remained important, the focus of urbanization shifted toward the Atlantic as urbanization grew rapidly in nations with easy access to the open ocean such as Portugal, Spain, the United Kingdom and the Netherlands. In addition, nation states based on monarchies arose throughout Europe and established colonies in Africa and the New World. These new nation states were supported by an unprecedented growth in military and civil administration financed by taxes and debt (Brewer 1990).

The rise of western Europe and globalization were accompanied by a significant growth in urbanization. In Europe, while the upward trend was not uniform over time, de Vries (1984) finds that the number of cities with populations of at least rose from 154 to 364 between 1500 and 1800. The growth in urbanization was concentrated in the very largest cities, whose main functions were to serve either as a government capital or as a port city (de Vries 1984; Bairoch 1988). The concentration of merchants in cities lowered the costs of financing and coordinating trade around the globe. Likewise, governments became concentrated in cities since the efficiency of military and government operations involved the collection and processing of an enormous amount of information (Brewer 1990).

The pre-industrial cities in the Middle East and Europe possessed a variety of political regimes

across space and over time. Most often, city states were ruled by kings with absolute power, but in some instances they were ruled by merchant elites. While the rulers of city states provided order and stability, they also imposed heavy tax burdens on their subjects. Between 1000 and 1800, city states that were ruled by absolutist governments grew much less significantly than those governed by merchants or assemblies (De Long and Shleifer 1993).

Industrialization and Urbanization

The Industrial Revolution, which began in Britain around 1700, transformed the modern world in a short period of time. In pre-industrial times, which spanned thousands of years from the ancient and medieval periods to the first two centuries of the modern era, cities became important centres of government, trade and artisan manufacturing, but most of the population lived in rural farms and villages. Yet, within a little over a century after the onset of the Industrial Revolution, the majority of people in Britain lived in cities. As industrialization spread to other European nations and the United States in the 19th century, and eventually to an ever growing list of nations in the 20th century, world urbanization rose rapidly.

The Industrial Revolution's transformation of the urban order began in the countryside. The early factories arose in rural villages and towns rather than in established major urban centres, and less urbanized places industrialized more rapidly in Europe and in the United States (Bairoch 1988). However, as industrialization matured it was significantly correlated with urbanization. The rise of the manufacturing sector was not only responsible for the emergence of new industrial cities, but also contributed to the growth of traditional urban centres. Between 1800 and 1900, the share of the urban population in industrialized countries almost tripled, from 11 to 30 per cent (Bairoch 1988).

It remains unclear why early industrialization was rural or why late industrialization was urban. Scholars generally believe that early factories chose rural locations because of the availability of water power, coal or unskilled labour (Bairoch 1988). Because early factories used women and child labour intensely, Goldin and Sokoloff (1984) believe that, in the United States, early industrialization arose in rural New England rather than in the rural South because the relative labour productivity of women and children was less than that of men in the former region. For Rosenberg and Trajtenberg (2004), industrialization caused urbanization as firms adopted the Corliss steam engine as their primary power source.

Kim (2005a, b) suggests that explanations of why industrialization led to urbanization are likely to rest on the rise of division of labour and the labour market. Prior to industrialization, goods were produced by self-employed artisans who made the entire product. With industrialization, factory owners hired workers in a labour market and employed them in specialized tasks. Because early industrialization was concentrated in a limited number of industries and was limited in scale, firms located in rural places since the costs of recruiting workers were relatively modest. However, as industrialization rose in scale and spread to numerous industries, the agglomeration of workers and firms in cities deepened the extent of the division of labour and lowered labour market transaction costs.

One of the major developments associated with the Industrial Revolution was a transportation revolution. In the pre-industrial period, the bulk of long-distance trade occurred over bodies of water such as canals, rivers, lakes, seas and oceans. With the introduction of the railroad and later trucks and airplanes, overland transportation costs fell dramatically. In the United States, the integration of regional economies led to a significant increase and then decrease in regional specialization (Kim 1995). The rise of US regional trade led to the rise of many large inland cities like Chicago, which emerged to coordinate the increase in domestic trade.

In the second half of the 20th century, although the rate of urbanization slowed in industrialized nations, cities remain a vital component of the modern economy. Scholars believe that one or more factors, such as Marshallian externalities

(technological spillovers, non-traded industry specific inputs, and labour market pooling), market size and natural advantages, cause the formation and growth of cities (Henderson and Thisse 2004). Moreover, while economic factors are much more significant in modern than in pre-modern times, political factors, such as tariffs and dictatorships, remain important for urbanization (Ades and Glaeser 1995).

Patterns of World Urbanization

Urbanization was not confined to the Middle East and Europe. Cities arose indigenously in India (1000–400 BC), China (700–400 BC) and the New World (100 BC). Cities diffused to Korea (108 BC–AD 313) and Japan (AD 650–700) from China, and to south-west Asia (AD 700–800) from China and India. In the New World, cities arose independently in Mesoamerica in Mexico and the Andes in South America. With the arrival of maize agriculture from Mesoamerica, urbanization reached the south-western and eastern parts of North America. Whereas China and Japan contained some of the largest pre-industrial cities, and had urbanization rates comparable to, or perhaps even higher than, pre-industrial European societies, the cities in the New World were smaller, fewer in number, and less stable.

In Africa, there is considerable uncertainty as to whether agriculture and cities arose independently. While there is evidence of domesticated agriculture in Sahel (5000 BC) and tropical West Africa (3000 BC), scholars remain unsure whether founder crops arrived from elsewhere (Diamond 1997). There is evidence of cities in Africa as early as 1000 BC, but it is also not clear whether these cities resulted from outside influences (Bairoch 1988). In Australia, no cities arose indigenously as the aboriginal population remained hunters and gatherers.

The coming of the modern era in 1500 not only transformed western Europe but also decisively altered the path of development around the world. As European nations colonized the New World and parts of Africa and Asia, they transplanted their technologies, agriculture, germs and political institutions to their colonies. From the colonies, the Europeans extracted new plants and resources and traded them around the globe. The demography of the New World colonies was fundamentally altered as the native population suffered and, to a varying extent, was supplanted by European immigrants and African slaves. In general, colonization and globalization were accompanied by a rapid growth in urbanization in Europe and the colonies.

In the 19th and the 20th centuries, the uneven diffusion of the Industrial Revolution around the globe determined the patterns of world urbanization. While there is no general consensus on the causes of uneven economic development, explanations usually rest on one or more factors related to geography, technology, trade and institutions (see Aghion and Durlauf 2005). In nations that developed, industrialization led to a rapid growth in urbanization; in those that did not, urbanization remained relatively low. In addition, most of the urban population in poor nations became concentrated in a handful of very large cities.

Summary

Cities in history arose with the advent of agriculture as they became centres of governments, crafts, religion and universities. As markets and trade developed, cities became centres of finance and commerce. While the patterns of urbanization differed greatly over space and time, the causes of urbanization were the same around the world. Pre-industrial urbanization rose with technological advances in agriculture, artisan manufacturing, transportation and trade, but fell with environmental degradation, famine and disease. Just as importantly, cities rose and fell with military and administrative efficiency of city states, territorial empires and nation states.

With the Industrial Revolution, cities became centres of factories and offices. Although early industrialization began in rural areas, the shift in the employment composition from agriculture to manufacturing and services led to rapid urbanization. The formation and growth of industrial or modern cities are attributed to benefits from a

variety of market and non-market factors such as the division of labour, lower search costs of matching specialized workers and firms, information spillovers, market size, and non-traded intermediate inputs (Henderson and Thisse 2004). With the rise in disposable incomes, cities became centres of arts, entertainment and other amenities.

The history of civilization is the history of urbanization. Without cities, the pillars of civilization – literature, science and the arts – would not exist. But as the industrial era gives way to the information era, will cities disappear? Leamer and Storper (2001) believe that face-to-face interactions in cities are likely to remain important for some time to come. For these scholars, the coordination of complex, unfamiliar and innovative activities depends on the successful transfer of uncodifiable messages and requires long-term relationships, trust, closeness and agglomerations. Wherever the new innovative activity may arise, be it in commerce, finance, politics, arts or science, the future of civilization is likely to rest on the success of its citizens.

See Also

- ▶ Central Place Theory
- ▶ Location Theory
- ▶ Marketplaces
- ▶ New Economic Geography
- ▶ Regional Development, Geography of
- ▶ Spatial Economics
- ▶ Systems of Cities
- ▶ Urban Agglomeration
- ▶ Urban Economics
- ▶ Urban Growth
- ▶ Urban Political Economy
- ▶ Urban Production Externalities

Bibliography

Ades, A., and E. Glaeser. 1995. Trade and circuses: Explaining urban giants. *Quarterly Journal of Economics* 110: 195–227.

Aghion, P., and S. Durlauf (eds.). 2005. *Handbook of economic growth*. Amsterdam: Elsevier.

Bairoch, P. 1988. *Cities and economic development*. Chicago: University of Chicago Press.

Brewer, J. 1990. *The sinews of power: War, money and the English state, 1688–1783*. Cambridge, MA: Harvard University Press.

Childe, V. 1950. The urban revolution. *Town Planning Review* 21: 3–17.

De Long, J., and A. Shleifer. 1993. Princes and merchants: European city growth before the industrial revolution. *Journal of Law and Economics* 36: 671–702.

de Vries, J. 1984. *European urbanization 1500–1800*. Cambridge, MA: Harvard University Press.

Diamond, J. 1997. *Guns, germs, and steel*. New York: W.W. Norton.

Goldin, C., and K. Sokoloff. 1984. The relative productivity hypothesis of industrialization: The American case, 1820–1850. *Quarterly Journal of Economics* 99: 461–488.

Hammond, M. 1972. *The city in the ancient world*. Cambridge, MA: Harvard University Press.

Henderson, J., and J.-F. Thisse (eds.). 2004. *Handbook of regional and urban economics*, vol. 4. Amsterdam: Elsevier.

Kim, S. 1995. Expansion of markets and the geographic distribution of economic activities: The trends in U.-S. regional manufacturing structure, 1860–1987. *Quarterly Journal of Economics* 110: 881–908.

Kim, S. 2005a. Industrialization and urbanization: Did the steam engine contribute to the growth of cities? *Explorations in Economic History* 42: 586–598.

Kim, S. 2005b. *Division of labor and the rise of cities: Evidence from U.S. industrialization, 1850–1880*. Mimeo. St. Louis: Washington University.

Lane, F. 1973. *Venice: A maritime republic*. Baltimore: Johns Hopkins University Press.

Leamer, E., and M. Storper. 2001. The economic geography of the internet age. *Journal of International Business Studies* 32: 641–665.

North, D. 1981. *Structure and change in economic history*. New York: W.W. Norton.

Palmer, R., and J. Colton. 1965. *The history of the modern world*. New York: Alfred Knoft.

Rosenberg, N., and M. Trajtenberg. 2004. A general purpose technology at work: The Corliss steam engine in the late-nineteenth-century United States. *Journal of Economic History* 64: 61–99.

Wrigley, E. 1987. *People, cities and wealth*. Oxford: Basil Blackwell.

Ure, Andrew (1778–1857)

William Lazonick

Andrew Ure, MD, was professor of chemistry and natural science at Anderson's College, Glasgow

from 1804 to 1830. In 1830, he introduced the word 'thermostat' into the English language in conjunction with a patent that he secured (Standfort 1982, p. 659). At about the same time, he moved to London to serve as a consultant in analytical chemistry to the Board of Customs. From 1832 to 1834, his major research assignment was to ascertain the wastage rate of raw material in sugar refining in order to determine the rebates on raw sugar import duties that British refiners could legitimately claim. Ure (1843, p. iv) complained that his research saved the exchequer £300,000 but yielded him only £800 in remuneration and cost him his health.

To recuperate, he 'spent several months in wandering through the factory districts of Lancashire, Cheshire, Derbyshire, &c., with the happiest results to his health; having everywhere experienced the utmost kindness and liberality from the mill-proprietors' (Ure 1835, p. viii). Two important books were the result. *The Philosophy of Manufactures* (1835) and *The Cotton Manufacture of Great Britain* (1836) are detailed technical treatises on the industry at the heart of Britain's industrial revolution, interlaced with commentary on the salutary moral, intellectual and physical effects of factory life on the workers.

Ironically, it was Karl Marx who established Ure's place in the history of economics. Ure's blatantly pro-capitalist stance, combined with his obvious technical expertise, made him the perfect 'horse's mouth' in Marx's attempt to show how capitalists used technology to throw adult males out of work and turn women and children into mere appendages of the machine. Marx (1867 [1977], pp. 560, 563–4) invoked the authority of Ure, who, in the conflict-ridden 1830s, argued that the diffusion of more automated technology would 'put an end … to the folly of trades' unions', proving that 'when capital enlists science into her service, the refractory hand of labour will be taught docility' (Ure 1835, pp. 23, 368).

Writing some three decades later, however, Marx failed to distinguish pro-capitalist ideology from ongoing reality. Contrary to Ure and Marx, adult male workers had not been definitively humbled, even in the presence of mechanization. Rather, certain groups of workers had maintained substantial control of work organization and had built up considerable union power (Lazonick 1979). In effect, Marx's uncritical use of Ure provided the 'evidence' needed to confirm that, in their confrontation with capitalists armed with technology, workers had 'nothing to lose but their chains'. Theory and history were parting company in Marx's theory of capitalist development (Lazonick 1986).

See Also

► Taylorism

Selected Works

1835. *The philosophy of manufactures*. London: Knight.
1836. *The cotton manufacture of Great Britain*. London: Knight.
1843. *The revenue in jeopardy from spurious chemistry*. London: Ridgway.

Bibliography

Lazonick, W. 1979. Industrial relations and technical change: The case of the self-acting mule. *Cambridge Journal of Economics* 3: 231–262.
Lazonick, W. 1986. Theory and history in Marxian economics. In *The future of economic history*, ed. A.J. Field. The Hague: Kluwer-Nijhoff.
Marx, K. 1867. *Capital*, vol. I. New York: Vintage, 1977.
Standfort, J.T. 1982. Thermostat. In *Encyclopedia Americana*, vol. 26. New York: Grolier.

US Mortgage and Foreclosure Law

Zachary K. Kimball and Paul S. Willen

Abstract

A mortgage is an exchange of a collection of rights between a borrower and a lender. In this article, we describe those rights and explain both their economic logic and their implications

for economic analysis and policy. We briefly discuss the medieval origins of the American mortgage contract and its evolution into its present form. We then turn to topics relevant for contemporary economic research – including title and lien theory; recording and registration of documents; judicial versus power-of-sale foreclosure; deficiency judgments and recourse; assignments; the Mortgage Electronic Registration System; and methods for avoiding foreclosure, including deeds-in-lieu and short sales. Our discussion focuses on real property law and its economic implications; we do not discuss, for example, securities law related to mortgage contracts.

Keywords

Equity; Default; Foreclosure; Housing; Law and economics; Mortgage; Residential real estate and finance

JEL Classifications

G21; G22; G28

Introduction

Mortgages underlie a great deal of property ownership in the USA, both commercial and residential – more than 69% of US owner-occupied housing units are subject to a mortgage (US Census Bureau, 2010 Census). These critical tools aid in the smooth operation of the housing market, and residential mortgages allow borrowers to live in homes that they otherwise could not afford to own. But for such an instrumental component of the economy, mortgages are widely – and wildly – misunderstood. We explain the complex legal and economic structure of a modern mortgage, including its applications to foreclosures and public policy. Our goal is to provide a conceptual overview, not comprehensive coverage of all aspects of mortgage law; this article is not written to provide legal advice.

To understand mortgage law, it is useful to go back to its historic origins in medieval England. The original common-law mortgage was a repurchase agreement in which the borrower sold the property to the lender and promised to buy it back by repaying the loan, plus interest, on an agreed date known as law day. If the borrower failed to appear on law day, the repurchase agreement was void and the lender received *clean* title to the property – title unencumbered by the borrower's right to repurchase. English courts of equity viewed this contract as unfair because the value of the property could exceed the balance of the loan, in which case failing to appear on law day would lead to an excessive transfer of wealth from borrower to lender. To remedy this, courts in 16th century England gave the borrower the right to repurchase, or *redeem*, the property, even if he or she had failed to appear. The borrower could exercise this repurchase right by paying off the loan, including interest and any associated costs. The courts understood that there needed to be some limit on this *right of equitable redemption*, as it became known, because otherwise lenders could never obtain clean title and, under such circumstances, no property could ever serve as good collateral for a loan. To solve this issue, courts allowed lenders to petition to *foreclose* the borrower's right of equitable redemption. This basic legal concept is the principle behind foreclosure to this day.

As we explain in section "Legal Frameworks: Title Theory and Lien Theory" below, in 30 US jurisdictions a mortgage contract is still a repurchase agreement as it was in medieval England. But even where it is not, the repurchase metaphor goes a long way towards explaining some counterintuitive concepts about mortgage law. For example, in common usage a mortgage transaction involves the lender giving a mortgage to the borrower; however, in the eyes of the law, the borrower is actually granting the mortgage to the lender. The logic is that the mortgage transaction, as in medieval England, is the grant of the property from the borrower to the lender.

In a sense, a mortgage contract establishes a form of shared ownership of a residential property, in which each party can extinguish the other's ownership rights under certain conditions. For the borrower, the equitable right of

redemption provides the power to claim the property by paying some agreed amount of money. In the event that the borrower fails to repay the loan as promised, the lender gets the right to extinguish the borrower's ownership interest by following the appropriate foreclosure procedure.

In this article, we first discuss the mortgage contract, including the two principal legal regimes behind the transfer of ownership rights, as well as the standards for recording mortgage documents with the relevant authorities. We then discuss foreclosure proceedings, including how they are affected by the different legal regimes, why such formal procedures are necessary, and some repercussions and complications that can emerge.

The Mortgage Contract

A mortgage is an exchange of a collection of rights between a borrower and a lender. Although in common usage a mortgage refers to a loan secured by real estate, legally the loan is called a 'note' and is secured by a separate instrument called a 'mortgage'. The *note* is a debt contract which specifies the amount lent and the schedule for repayment – including the interest rate, amortisation schedule, prepayment penalty and any other relevant information such as the index used for adjustable-rate mortgages. With the *mortgage*, the borrower conveys to the lender an interest in the property in order to secure the debt evidenced by the note.

The contract works simply. If the borrower makes the scheduled payments according to the note, the lender can exercise essentially no property rights. The lender cannot sell the property, or even set foot on it, without the permission of the borrower. By repaying the loan, the borrower satisfies the note and redeems the mortgage, extinguishing the lender's security interest in the property. If the borrower violates the terms of the note, or *defaults*, the lender can exercise the right in the mortgage to foreclose the borrower's equity of redemption and extinguish the borrower's interest in the property. What constitutes default is specified in the note; it includes failing to make a scheduled payment, but may also include other violations of the contract, such as failing to insure the property or renting the property without permission. It is important to recognise that problems with the borrower's finances or with the collateral property do not directly constitute default on a mortgage, and that this is different from many other types of loan. For example, the borrower is not required to maintain a certain amount of equity in the property, as would be required with a margin loan against a stock holding. We discuss default and foreclosure proceedings in more detail in section "Foreclosure Proceedings".

Almost all mortgage contracts today are uniquely connected to both a particular property and a particular borrower. Historically, however, lenders offered both assumable and portable mortgages. An *assumable* mortgage is tied to the property and can be transferred to a new owner after a sale; today, a typical mortgage includes a due-on-sale clause which requires the borrower to pay off the loan when the property is sold. A *portable* mortgage follows a borrower from property to property, allowing the borrower to keep, for example, a low interest rate even if market rates have increased; such loans are rare.

In the remainder of this section, we discuss two topics relating to mortgage contracts. The first involves the two conceptually distinct US legal frameworks for mortgages, known as title theory and lien theory. The second is the role of the recording process for mortgage documents.

Legal Frameworks: Title Theory and Lien Theory

Although the USA inherited much of its property law from England, individual states have since developed distinct branches. The main difference across jurisdictions relates to the ownership framework. According to the first such framework, *title theory*, the mortgage actually conveys *legal* title to the property from the borrower to the lender with a mortgage *deed*. It is only upon satisfying the mortgage note – by paying off the debt – that the borrower becomes a legal homeowner. Meanwhile, the borrower retains *equitable* title to the property and, for all intents and purposes, remains the apparent and effective owner of the property. In Massachusetts, the

Supreme Judicial Court described the phenomenon thusly: 'to all the world except the [lender], a [borrower] is the owner of the mortgaged lands'. Dolliver v. St. Joseph Fire & Marine Ins. Co. 128 Mass. 315, 316 (1880).

Over time, because the lender could not exercise any property rights despite having legal title to the property, some jurists argued that 'the [lender] could no longer be said to have legal title. . . and the interest of the [lender] was only a security interest which was called a lien' (Osborne 1951, p. 311). Statutes were enacted in a minority of states, the earliest of which was South Carolina in 1791, officially effecting this change and establishing the alternate theory of conveyance, *lien theory*, which allows the borrower to maintain legal title to the property. In a lien-theory state, property rights are conveyed by a mortgage *lien*. In contrast to a deed, a lien does not transfer title to the property. Instead, a lien grants the right to recover debt through the sale of the property if the borrower defaults on the note, although this usually requires a lawsuit.

Thus title theory, which is used in 30 US jurisdictions, including Arizona, California and Nevada, is the most directly analogous to the medieval legal regime in which the mortgage was an explicit repurchase agreement. However, lien theory, used in the other 21 jurisdictions, including Florida, Illinois and New York, can also be construed in roughly the same terms.

So under either framework the borrower retains effective ownership of the property until one of two possible events occurs. The first is that the borrower satisfies the note by paying off the debt, at which point either, under title theory, legal title reverts to the borrower or, under lien theory, the lien is extinguished. The other possibility is that the borrower defaults, usually by failing to make a periodic payment. A borrower in default maintains the right to repay the debt in full – including late payments, fees and other expenses – and thereby satisfy the note. Under title theory, a defaulted borrower satisfying the note does not automatically regain legal title unless the lender reconveys it, but a court can compel this reconveyance; under lien theory, satisfying the note automatically extinguishes the lien, and insodoing the borrower immediately has free and clear ownership of the property (Osborne 1951, pp. 836–7).

As long as the borrower maintains the equity of redemption – the ability to satisfy the note and regain title to the property – the lender's ownership is in question. In order to remove this *cloud* from the title, a lender must foreclose on the borrower's equity of redemption. By doing so, the borrower loses his or her right to redeem the mortgage and regain clear title; this process of extinguishing the borrower's equity of redemption is the *foreclosure*. In a title-theory state, foreclosure is usually carried out through a foreclosure auction (see section "Foreclosure Types: Strict Foreclosure and Foreclosure by Sale"). In a lien-theory state, because the lender does not yet possess legal title to the property, the lender usually must go to court to effect the foreclosure. Thus, in most lien-theory states, foreclosure is carried out through a judicial process (see section "Foreclosure Methods: Judicial Foreclosure and Power-of-Sale Foreclosure").

Borrowers in all states can redeem the mortgage debt before foreclosure. About half the states also have a *statutory right of redemption* explicitly permitting borrowers or their successors a limited time – generally six months to two years – to redeem the mortgage after a foreclosure, usually for the price of the foreclosure sale (Nelson and Whitman 1985, p. 616). This action nullifies the foreclosure sale, but it is hardly ever used in practice (Nelson and Whitman 1985, p. 622).

Mortgage Records: Document Recording, Registration, and Priority

As a general rule, all mortgages in the USA are publicly recorded at a town or county registry, while records of the note are kept privately by the borrower and the lender.

Public records generally contain most transfers of interests in a property from one party to another. However, they typically do not contain the title to the property because there is usually no physical document showing title. To establish that a particular person has title to a property, one must show that anyone with a previous interest in the

property has relinquished it – in other words, all previous owners have deeded the property to someone else, forming a chain of ownership, and all previous mortgages have been discharged. This system of recording transfers and inspecting the historical record to establish a chain of title is the de facto standard throughout the country.

However, in some jurisdictions there is title registration for property more analogous to title registration for vehicles, which usually involves physical documents. Such a method 'registers and determines title to land so as to amount to practically a government patent to each purchaser', but these systems have not been fully utilised even where they have been formally adopted (Osborne 1951, p. 496). In Massachusetts, for example, one of the states which permits land registration, only about 20% of land is registered.

The role of the public records differs across states. In some states, only recorded documents have legal standing. In other states, a deed is valid even if it is never recorded. In the latter states, however, a recorded deed almost always takes precedence over an unrecorded deed, so failure to record is rare. To understand the issue, consider an example. Suppose Alex sells a house to Bailey and Bailey does not record the sale deed. In some jurisdictions, this transaction is perfectly valid and Bailey has title to the property. Now suppose Alex also sells the house to Casey and Casey's deed is recorded. Then both Bailey and Casey think they own the house, but in some states, the courts would hold that Casey, as the first to record the purchase, is the legal owner of the property. In other words, recording provides protection to the buyer, even if the contract is valid without being recorded. (Of course, Bailey could then sue Alex for damages, but Bailey has no power over Casey's valid ownership. However, if Casey had knowledge of Bailey's deed before buying the house from Alex, Casey's deed would be considered invalid in some, but not all, states. These details are governed by state recording statutes, which are known generally as one of 'race', 'notice' or 'race-notice'. See Hunt et al. (2011) for a brief discussion.)

Documents eligible for recording include purchase deeds between homeowners, mortgage deeds and liens between borrowers and lenders, and more controversially, *assignments* which transfer ownership of mortgages between lenders. Unlike deeds involving homeowners and borrowers, which are almost always recorded, assignments between lenders frequently go unrecorded. To make matters even more confusing, in the 1990s the mortgage industry set up the Mortgage Electronic Registration System (MERS). MERS was created for many reasons. One was because an increase in securitisation meant that assignments were more common and consequently their frequency was more burdensome. Another was that most registries still required paper copies. But a third important reason was that, during the savings and loan crisis of the 1980s and 1990s, bank failures resulted in serious title problems as mortgages changed hands, often several times and without clear documentation.

To use MERS, the lender assigns MERS as the mortgage owner of record in the registry. MERS, in turn, keeps track of any underlying assignments of ownership from one lender to another. If and when the controlling lender needs a recorded interest in the property, MERS goes to the registry and records an assignment of the mortgage to that lender. MERS has worked smoothly since its inception in 1995, but during the foreclosure crisis that started in 2007, some people raised questions about its legality, in particular with respect to foreclosures. In general, appellate courts have found in favor of MERS, but it remains controversial and litigation continues.

Most US jurisdictions respect the doctrine that 'the mortgage follows the note', meaning that any time a mortgage note is sold from one party to another, ownership of the mortgage goes along with it automatically, without requiring a separate assignment. The two notable exceptions are Massachusetts and Minnesota. Reliance on this doctrine may simplify foreclosure, because the foreclosing party need only demonstrate possession of the note in order to have the right to foreclose that is provided by the mortgage. However, this simplification has recently been called into question, particularly regarding its interplay with MERS assignments.

Foreclosure Proceedings

The fundamental principles of foreclosure date back centuries, but the actual procedures have evolved considerably over time. To understand the logic of the foreclosure proceeding, it is important to understand the motivations of the two key parties. For the lender, the primary goal is to ensure that there is no risk that the original borrower can recover the property, which allows a new buyer of the property to get clean title. For the courts, the overarching concern is to prevent a lender from injuring a borrower by taking more than the lender is owed. Under current foreclosure law, the court is not generally concerned with whether the original loan was suitable for the borrower or whether the lender made efforts to prevent foreclosure, because the court's narrow focus is on whether borrower and lender upheld their respective ends of the mortgage contract. In particular, neither the mortgage contract nor current legal principles oblige the lender to modify the loan or work with the borrower to prevent foreclosure.

Foreclosure Types: Strict Foreclosure and Foreclosure by Sale

Before the 19th century, foreclosures were what are now called 'strict', meaning that the lender took possession of the property and it was disposed of at the lender's discretion. However, US courts found that this was unsatisfactory for largely the same reason the English courts of equity distrusted the original medieval mortgage – the value of the property could exceed the amount owed and in that case 'there is injustice to the [borrower]' (Osborne 1951, p. 904). The solution to this issue became known in the USA as 'foreclosure by sale', whereby foreclosure is effected by a public auction of the property and the borrower receives any proceeds in excess of the amount owed (Osborne 1951, p. 908). It is important to emphasise that the auction does not take place after the foreclosure; the auction itself is the legal foreclosure. In other words, the equitable right of redemption vanishes the moment the auctioneer sells the property.

Most often, the auction is something of a formality where the lender is the high bidder and the property ends up in the 'real-estate owned', or REO, portfolio of the lender. At the auction, the lender usually starts the bidding, often with an offer much higher than the actual market value of the property. This at first appears puzzling, but recall that the goal of the lender in the foreclosure process is to ensure that title to the property is clean. Since a borrower could contest the foreclosure if the lender does not get the best possible price at auction, a low winning bid could potentially cloud title to the property. By setting the opening bid sufficiently high – often only slightly less than the borrower owes, which is generally more than the property is worth – the lender can forestall any challenge to the foreclosure.

Foreclosure Methods: Judicial Foreclosure and Power-of-Sale Foreclosure

Two types of foreclosure by sale emerged in US law. The first is foreclosure by judicial sale, in which the lender petitions the court and the court orders a foreclosure auction. Judicial sale is available in every jurisdiction. The alternative approach is that, when the mortgage is originated, the borrower gives the lender the right to carry out a foreclosure auction in the event of default, a right known as the 'power of sale' (Osborne 1951, p. 992). Although rare in the early 19th century, power-of-sale foreclosure became more common in the USA over time (Osborne 1951, p. 993).

Power-of-sale foreclosure is available in a majority of states. In general, states in the south and west of the country offer power of sale and states in the north and east are judicial; whether power-of-sale or judicial foreclosure is the preferred method aligns almost exactly with whether the state follows title or lien theory, respectively. Of the states with the most severe foreclosure problems in the current crisis, Arizona, California and Nevada all allow power-of-sale foreclosure, while Florida only allows judicial foreclosure. Other notable judicial states include Illinois, New York and New Jersey. For fuller discussion of judicial and power-of-sale foreclosure, see Gerardi et al. (2011) and National Consumer Law Center (2010).

Some have suggested that the judicial procedure, by giving the borrower an opportunity to

appear in court, is friendlier to the borrower. Meanwhile, power of sale is generally viewed as lender-friendly because lenders face no official supervision by the courts. But the truth is more nuanced. Under power of sale, the desire for clean title leads to implicit supervision. Specifically, most buyers of residential real estate need title insurance – lenders usually require it before funding a loan – and title insurers will not insure a foreclosed property if there is any chance that a previous owner could contest the title and that the courts could declare the foreclosure invalid. So, in a sense, title insurers act as third-party enforcers in place of the courts in power-of-sale states.

In judicial states the courts provide explicit supervision, but that supervision provides surprisingly little, if any, additional protection to borrowers. To get the court to order a foreclosure in a judicial state, a representative of the lender must attest that three key conditions are met: the borrower took out a mortgage, pledged the property as collateral, and defaulted on the mortgage. This attestation usually comes in the form of an affidavit certifying that the representative has reviewed the borrower's loan file. Since effectively all borrowers facing foreclosure meet these conditions, the borrower has little to contest in court and borrowers rarely succeed in blocking foreclosure; borrowers contesting judicial foreclosures usually yield only delays. During the recent foreclosure crisis, some lenders' representatives signed affidavits without complete knowledge of the loan files, a practice often referred to as 'robosigning'. Because of these affidavits, borrowers were able to raise questions about the validity of the attestations despite the fact that the foreclosure files met the three key conditions.

The data suggest that judicial foreclosure is borrower-friendly and lenderunfriendly only in the sense that it extends the foreclosure timeline. Gerardi et al. (2011) show that fewer than half of initiated foreclosures are completed within three years in 14 of the 18 judicial states, whereas the same is true in only seven of the 33 power-of-sale jurisdictions. Nor do borrowers benefit from judicial foreclosure in other ways – they are not more likely to cure a serious delinquency in judicial states than they are in power-of-sale states, and they are not more likely to receive a mortgage modification.

Legal scholars have long argued that the power-of-sale procedure can replicate the protections of the judicial process at much lower cost. Nelson and Whitman (1985, p. 536), for example, write that

> The underlying theory of power of sale foreclosure is simple. It is that by complying with the above type statutory requirements the [lender] accomplishes the same purposes achieved by judicial foreclosure without the substantial additional burdens that the latter type of foreclosure entails. Those purposes are to terminate all interests junior to the mortgage being foreclosed and to provide the sale purchaser with a title identical to that of the [borrower] as of the time the mortgage being foreclosed was executed.

Mortgage Deficiency: Deficiency Judgments, Lender Recourse, and Second Liens

The concern of the courts, as we have discussed, has historically been that a foreclosure would injure a borrower who owned a property worth more than the loan balance. That is, of course, ironic because a borrower with a property worth more than the loan balance is virtually never foreclosed upon; the borrower can simply sell the property and pay off the debt. Rather, the overwhelming majority of foreclosures involve the opposite scenario: borrowers with negative equity. In this situation, when the borrower owes more on the loan than the value of the property, the amount recovered from the auction will not cover all the money the borrower owes. In the language of the mortgage contract, the security for the mortgage will not cover the debt specified in the note. This gap is called a *deficiency*, but it is not automatically a debt owed by the borrower, since the note has been extinguished. (The balance of a deficiency may become a debt due the lender as part of a strict foreclosure proceeding, although these are relatively rare (Nelson and Whitman 1985, p. 595).) Historically, lenders in the USA could sue the borrower and get a *deficiency judgment*, which converts the deficiency into an unsecured debt. This process of pursuing a mortgage deficiency is called *recourse* and is common throughout the world. During the Great

Depression, however, US lenders abused deficiency judgments by underbidding at auction in order to inflate deficiencies; consequently, some states enacted anti-deficiency statutes. The anti-deficiency laws often have confusing subtleties: in California, for example, deficiency judgments can only be pursued for judicial foreclosures of refinance mortgages, and even then the deficiency is limited and the borrower is provided a right of redemption. According to Ghent and Kudlyak (2011), only 11 states have legal systems that are effectively non-recourse. Statutorily, most mortgages in the USA are recourse loans.

It is true, however, that lenders generally do not pursue deficiency judgments on first mortgages and, as mentioned earlier, set an opening bid that is close to the amount owed on the loan in order to ensure a small deficiency, even if the REO sale price (i.e. the price the lender receives when reselling the property to a third party) leads to a large loss. The main reason lenders do not pursue deficiency judgments is that borrowers who lose their homes typically have few or no other assets. US bankruptcy also law allows borrowers to discharge deficiency judgments. Moreover, a deficiency judgment may cause the court to question the fairness of the foreclosure, in particular the auction and bidding process, regardless of the circumstances. Thus, lenders often forgo pursuit of a deficiency judgment because of the inclination of mortgage law towards borrower protection; in this way, most first mortgages are effectively non-recourse in practice.

In contrast to lenders of first mortgages, lenders of second mortgages have only a limited amount to gain from foreclosure, because the collateral for the second mortgage is not the property itself but the borrower's equitable right of redemption. In other words, if the second lender forecloses, the buyer at auction acquires the property with the first mortgage still in force. Such foreclosures on second mortgages are rare; much more common is foreclosure of a first mortgage on a property which also has a second mortgage. In that case, the second lender is entitled to recover its debt from the proceeds of the foreclosure sale, but only after the first lender's debt has been satisfied. If the auction price is insufficient to cover the amount owed on the second mortgage, as is usually the case, the second lender can pursue a deficiency judgment against the borrower; deficiency judgments for second mortgages are much more common than deficiency judgments for first mortgages.

Avoiding Foreclosure: Deeds-in-Lieu and Short Sales

Given the time, expense and complexity of foreclosure, many in the current crisis have asked if there are alternatives. Alternatives such as mortgage modifications that allow the borrower to retain ownership do not generally involve real property law and are beyond the scope of this article. However, there are two procedures designed to generate the same outcome as a foreclosure at a lower cost.

A deed in lieu of foreclosure, or simply a *deed-in-lieu*, is when the borrower deeds the property to the lender in exchange for forgiveness of most or all of the mortgage debt. A *short sale* is when the borrower sells the property for less than the outstanding balance of the loan and the lender agrees to discharge the mortgage despite the deficiency. Both procedures benefit the lender, saving time and expense, and the borrower, who gets a cleaner exit from an unfortunate situation and a less damaged credit history than would result from foreclosure. Further, lenders agreeing to deeds-in-lieu or short sales generally choose to forgo the possibility of a deficiency judgment; choosing to forgo the deficiency also helps to preclude accusations of exploiting the borrower. The downside of a deed-in-lieu or short sale from the borrower's perspective is that he or she cannot live rent-free while waiting for a foreclosure auction – a wait that can often take months or years.

Deeds-in-lieu and short sales each face a serious obstacle: they avoid a true foreclosure. At first, it may be surprising that this is an obstacle at all, particularly in light of concerns about foreclosures and their impact during the recent financial crisis. The courts, however, have historically viewed the foreclosure process, and the foreclosure auction in particular, as a central protection for the borrower. Without an auction, there is no way to know for sure whether the borrower

surrendered a property worth more than the mortgage debt. The courts might then question whether the lender coerced or misled the borrower into giving up the right to a full foreclosure and thereby circumvented, or *clogged*, the borrower's equity of redemption. It is for this reason that 'the deed in lieu of foreclosure can create substantial problems for the [lender] and is often, from its perspective, a dangerous device' – the same holds true for a short sale (Nelson and Whitman 1985, p. 474). This is even more true when the borrower has a second mortgage, because deeds-in-lieu and short sales negotiated with a first lender do not extinguish subordinate mortgages; in those cases, 'the only prudent alternative for the [first lender] is to foreclose' (Nelson and Whitman 1985, p. 476).

See Also

- ▶ Foreclosure, Economics of
- ▶ Household Portfolios
- ▶ Subprime Mortgage Crisis

Bibliography

Gerardi, K., L. Lambie-Hanson, and P.S. Willen. 2011. Do borrower rights improve borrower outcomes? Evidence from the foreclosure process. *Public Policy Discussion Paper 11-9*. Boston: Federal Reserve Bank of Boston. Available at http://www.bostonfed.org/economic/ppdp/2011/ppdp1109.pdf

Ghent, A.C., and M. Kudlyak. 2011. Recourse and residential mortgage default: Evidence from US states. *Review of Financial Studies* 24(9): 3139–3186.

Hunt, J.P., R. Stanton, and N. Wallace. 2011. The end of mortgage securitization? Electronic registration as a threat to bankruptcy remoteness. *Manuscript*. Available at http://faculty.haas.berkeley.edu/stanton/papers/pdf/mers.pdf

National Consumer Law Center. 2010. *Foreclosures*, The consumer credit and sales legal practice series, 3rd ed. National Consumer Law Center.

Nelson, G.S., and D.A. Whitman. 1985. *Real estate finance law*, Hornbook series, student edition, 2nd ed. St. Paul: West Publishing.

Osborne, G.E. 1951. *Handbook on the law of mortgages*, Hornbook series. St. Paul: West Publishing.

U.S. Census Bureau, 2011. 2010. Census summary file 1 – United States.

Disclaimer The views expressed are those of the authors and do not necessarily reflect the position of the Federal Reserve Bank of Boston or the Federal Reserve System.

Use of Experiments in Health Care

Joseph P. Newhouse

Abstract

Two of the best known randomised trials in health economics are described in detail in this chapter: the RAND Health Insurance Experiment and the Oregon Health Insurance Experiment. The RAND Experiment randomised participants to health insurance plans that varied the cost of care from free care at one extreme to an approximation of a large deductible at the other. Those on the free care plan increased their use of services by about 30% relative to those on the large deductible plan. For the average participant there appeared to be little or no effect on health outcomes from this change in use. Among the poor with hypertension (high blood pressure), however, blood pressure was better controlled on the free care plan, which projected to about a 10% decrease in the risk of mortality. The Oregon Experiment randomised its participants to Medicaid or to remaining uninsured; those with Medicaid insurance used more services and suffered less from depression.

Keywords

Field experiments; Health insurance; Health insurance experiment; Randomised trials

JEL Classifications

I13; I18; C9

Medicine adopted the randomised controlled trial in the immediate post-Second World War period to determine the clinical effects of various drugs,

procedures and devices. Since that time results from tens of thousands of clinical trials have been reported in the medical literature, in part because such a trial is generally required to obtain approval to market a new drug in the USA, the European Union and many other countries. Trials in the economics of health care are, of course, much less common than trials in clinical medicine, but there have been well-known influential health economics trials, and one may expect their use to increase. Such trials are the subject of this chapter. As a semantic note, medicine tends to use the word *trial* and social scientists tend to use the word *experiment* to mean the same thing; I shall use the two words interchangeably.

One virtue of a randomised experiment is well known; assuming the randomisation is successful and there is no selection introduced from refusal to participate or attrition during the experiment, the treatment assignment is independent of both observed and unobserved variables. As a result, ignoring issues around refusal and attrition, a simple comparison of means between the treatment and control groups – or of means between alternative treatment groups – yields an unbiased estimate of the treatment effect. Rather than simply comparing means, however, analysts frequently estimate a regression equation with a treatment effect and covariates to lower residual variation. In principle this improves power and can adjust for any minor imbalances across treatment and control groups, but it can also lead to overfitting, meaning that mean square error can be larger if covariates are included than if they are not (Duan et al. 1983). Thus a reasonable precept in designing an experiment is that power should be sufficient without using covariates.

Randomisation thus directly addresses the problem that often arises when economists use observational data to estimate causal effects, namely that the allocation of persons to various treatments may not be independent of unobserved – or more accurately uncontrolled – variables that also influence the outcome, thereby creating bias in the estimate of treatment effects. For example, one may want to know how the generosity of the insurance contract affects demand for medical services, but standard theory implies that, *ceteris paribus*, individuals who have chosen more generous insurance expect to use more services, typically because they are, or expect to be, sicker (Rothschild and Stiglitz 1976). In other words, when estimating a demand curve from observational data the consumer's insurance contract is endogenous, at least in individual insurance markets where consumers have a choice of plan. Failure to address the endogeneity can result in overestimating the effect of insurance generosity on the use of services; some of the additional use by those with generous insurance contracts stems from their being sicker in ways that the analyst cannot control for. If, however, there is a negative correlation between risk aversion and sickness, the bias may result in an underestimate; in that case the healthy could buy a more generous insurance policy than the sick because they are more risk-averse.

The standard econometric treatment for endogeneity is, of course, the use of instrumental variable(s), but finding suitable instruments can be difficult in observational data. The randomised experiment is in fact a special case of instrumental variable analysis, in which the instrument is the initial randomised treatment assignment. Under standard assumptions, instrumental variable analysis can be used in the context of a randomised experiment to adjust for differential refusal to accept various treatment assignments or differential attrition among treatment assignments. The Oregon Health Insurance Experiment (OHIE) that I describe below used instrumental variable analysis in this fashion.

In addition to addressing the issue of endogeneity, there are other, less well appreciated virtues of designed experiments. Most importantly, the experimenter can choose the treatments whose effects are to be assessed, whereas the analyst using observational data must take the treatments as given. If the observational data do not include the region of interest, they are not very useful. Moreover, if there are multiple treatments of interest, the experimenter can take account of any differential interest in the various treatments when determining the proportion of the total sample to be assigned to each treatment. The example of the RAND Health Insurance Experiment

(RHIE) described below illustrates both these features.

The field of health economics began in the 1960s with the seminal theoretical paper of Kenneth Arrow (1963), and empirical work was not long behind. In fact, the RHIE, one of the best-known experiments in health economics, was carried out relatively early in the development of the field in the 1970s and early 1980s. I directed that experiment and describe it next; more details are available in Newhouse and the Insurance Experiment Group (Newhouse and The Insurance Experiment Group 1993). I then describe a much more recent experiment, the OHIE (Finkelstein et al. 2011; Baicker et al. 2013; Taubman et al. 2014). Like the RHIE, the OHIE is also well known; in addition to their prominence in the literature, I have chosen to describe those two experiments here since I participated in both and know them well.

Of course, other experiments have been conducted in health economics, both in the USA and elsewhere, for example the China Health Insurance Experiment (Sine 1994); a randomised experiment on cost sharing in Ghana (Ansah et al. 2009; Powell-Jackson et al. 2014); the MI FREEE trial of making effective drugs free for post-heart attack patients (Choudhry et al. 2011, 2014); the Accelerated Benefits Demonstration and Evaluation Project, a randomised trial to determine if earlier access to Medicare would improve health or promote an earlier return to work among new Disability Insurance recipients (Michalopoulos, et al. 2011); and the PNPM Generasi Program in Indonesia, a randomised trial of a community block program to improve health and education (Olken et al. 2011). I close with some data on the frequency of randomised experiments and a pointer toward additional experiments in health economics that are now in progress.

The RAND Health Insurance Experiment (RHIE)

This experiment sought to measure the price elasticity of demand for medical services. Because many persons, even many economists, do not want to assign normative meaning to observed demand curves, the RHIE also sought to determine the effects on health outcomes of induced changes in the use of services from different degrees of cost sharing. When the RHIE began in the early 1970s, the USA was debating various legislative proposals to establish a national health insurance plan, and a key point of contention was the role of cost sharing for services. One school of thought, represented most prominently by Senator Edward Kennedy (D-MA), held that medical care should be free at the point of service. Adherents of this view often argued that if persons had to pay for care they would put off receiving care and/or not comply with prescribed clinical regimens, with detrimental effects for their health and potentially higher subsequent cost of treatment. The opposing school of thought, led by the Nixon Administration, relied on standard economic theory, and argued in effect that free health care would encourage moral hazard, or in the vernacular, would lead to the overuse of care. Ultimately, as is well known, the USA did not adopt a national insurance plan at that time, but to generate evidence on the effect of cost sharing the Nixon Administration authorised the RHIE.

The RHIE was to some degree modelled after so-called income maintenance experiments, several of which began in the 1960s and were being carried out at the time the RHIE was conceived (Robins 1985). These experiments sought to determine the effect on labour supply of a minimum guaranteed income that was taxed away as earnings rose. The treatments in the income maintenance experiments varied the level of the guarantee and the size of the tax rate.

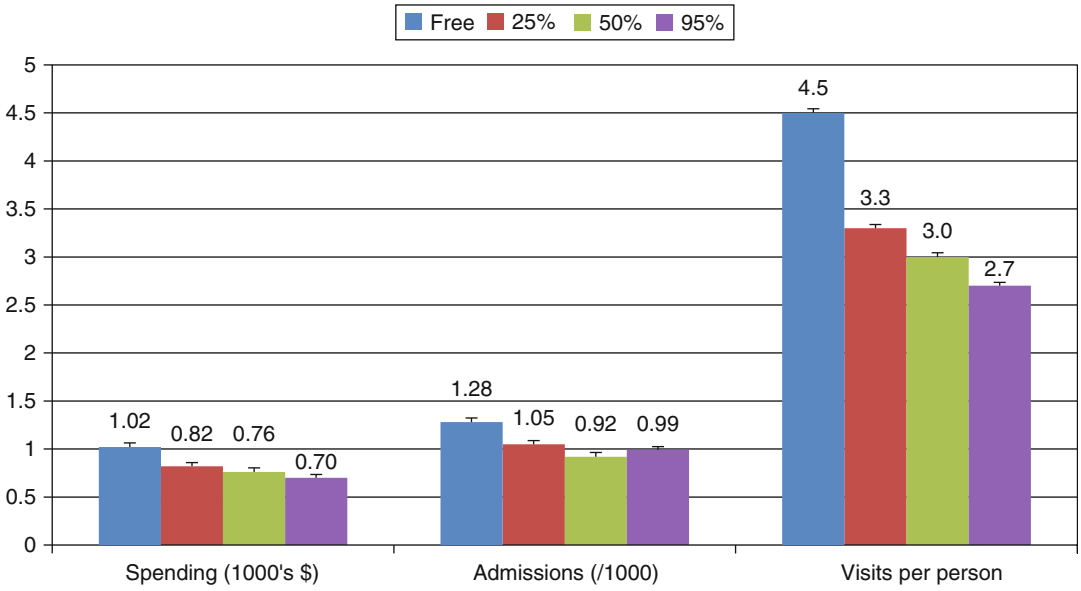
The treatments in the RHIE were patterned after the prevailing indemnity health insurance plans of the early 1970s. Those insurance plans passively reimbursed participants' expenses for medical services subject to a deductible and a coinsurance rate, which was often in the range of 20%. Some insurance plans, however, especially in unionised industries, had little or no cost sharing. There were rather loose upper limits on fees that could be charged. Insurers did not generally intervene in the delivery of care; managed care techniques that are common today in the USA

were rare. Thus, most of the 2,000 families in the RHIE were allocated to plans in which they either paid nothing for care, which I will term the free plan, or paid 25% coinsurance up to a stop loss, or paid 95% coinsurance up to a stop loss. The 95% coinsurance plan was intended to approximate a deductible, but reimbursed, 5% of the family's medical expense to give the family a modest incentive to file a claim. (At the end of the experiment an audit study showed that claim filing rates were around 7–9% less in dollar value in the 95% plan than in plans with lower coinsurance rates and correspondingly higher incentives to file a claim. As will be seen below, 7–9% is small relative to the differences in use among the plans.) Because there was some interest in the shape of the demand curve between 25% and 95%, the RHIE also allocated some families to a plan with 50% coinsurance, but it received a smaller share of families because of less interest in such a coinsurance rate. Almost all medical services and drugs were covered because information on use came from claims forms that the families filed with the RHIE, so there would be no administrative information on the use of a service if it were not covered.

In 1973 the United States Congress enacted the Health Maintenance Organization Act, which was intended to spur the growth of organisations that would agree to accept a capitated or fixed amount in return for supplying 'necessary' medical services. Support for this act was based on observational data from a few Health Maintenance Organizations (HMOs). Those data suggested the HMOs had less utilisation with no obvious detrimental health effects, although little effort had been made to determine effects on health outcomes. There also was concern that the lower observed utilisation might have arisen from favourable selection. One part of the RHIE therefore randomised families to an HMO. Ignoring one small and inconclusive randomised prior experiment in a new HMO, the RHIE was the first effort to carry out a randomised experiment with an HMO and remains the only such experiment I know of.

The RHIE sought to enrol a representative sample of the non-elderly American population. (The Nixon Administration did not want to experiment with cost sharing in the Medicare plan that covered those over 65 years of age.) Given the budget constraint for the experiment, the fixed costs of operating in a given site, and estimates of between-site and within-site variances, the optimal number of sites was six. In order to ensure a degree of representativeness, the sites were purposively chosen to represent all four census regions, to vary in population from two small rural sites to a large metropolitan area (Seattle, WA), and to vary in the wait for an appointment for a non-urgent problem. (The latter dimension was an observational study within the RHIE; because of the likelihood that a national plan with free care would lead to some form of non-price rationing, at least in the short run, the experiment sought to determine what observable consequences there were, if any, from variation across sites in waiting times to see a physician for a non-urgent problem. The only observable effect turned out to be that more persons sought care in the emergency department if waiting times for office visits were longer.) To determine if low-income families were differentially affected by cost sharing, those families were oversampled; in addition, the upper 3% of the national income distribution was excluded, as were individuals who were institutionalised at baseline. Thus, the sample consisted of persons under age 62 at the time of enrolment who were living in the community. They participated for either a three-or five-year period. (Those randomly assigned to the five-year enrolment period were under age 60 at the time of enrolment so that they would not become eligible for Medicare during the experiment.)

Most enrolled families, of course, had health insurance prior to the experiment, typically through the employer of an adult in the family. If those families were randomly assigned to a plan with cost sharing, they could well rationally refuse to enrol because their current insurance could be more generous than the experimental plan. Therefore the RHIE included an unconditional lump sum side payment that guaranteed the family they would not be financially worse off from participating and in fact with near certainty would be better off from participating and



Use of Experiments in Health Care, Fig. 1 Annual spending and utilisation, by coinsurance (Notes: The spending results are those predicted from a model that transformed raw spending into the logarithm of spending and then retransformed the logarithm back to dollars, but results from raw means are similar. Spending data are in 1991 dollars. Error bars show one standard deviation error. Further details are in Newhouse and the Insurance Experiment Group (Newhouse and The Insurance Experiment Group 1993))

completing the experiment irrespective of the plan to which they were assigned. Some random variation was built into this lump sum to estimate the income effect that the side payments caused; the results showed the income effect to be quantitatively unimportant. Interestingly, the side payments in the RHIE are analytically similar to an employer's deposits in a Health Savings Account.

Despite there being no financially rational reason to refuse or drop out, both refusals and attrition were somewhat lower on the free plan than on the plans with cost sharing; this difference in those refusing, however, was not related to any observable pre-experiment variable, including the use of services or self-rated health status. There have been two critiques of the RHIE that have emphasised the differential refusal and attrition rates as potentially leading to an overestimate of the demand response (Nyman 2007; Aron-Dine et al. 2013), but subsequent observational data have generally been consistent with the RHIE results (Newhouse et al. 2008; Chandra et al. 2010 2014; Baicker and Goldman 2011).

Figure 1 shows the main results on the annual utilisation of services and spending in plans with four different coinsurance rates, the free plan (zero coinsurance), and plans with 25%, 50% and 95% coinsurance. In the latter three plans annual out-of-pocket spending was capped at $1000 ($750 in the 95% plan); this cap was scaled down for low-income families. In general, utilisation decreased as coinsurance increased. The one exception was hospital admissions in the 25%, 50% and 95% plans; many people in these plans who were hospitalised, however, reached the annual out-of-pocket limit, in which case the cost of an admission was simply the amount of the out-of-pocket limit ($750 or $1000; less for lower income persons) less any other out-of-pocket health spending during the year. Thus a more meaningful comparison of hospital admission rates is between the free plan and all the cost sharing plans taken together. That difference is substantial.

Spending was about 30% less in the 95% coinsurance rate plan and about 20% less in the 25% coinsurance plan than in the free plan. Hospital

admissions were about 20% less in all the cost-sharing plans taken together, and there were between one and two more physician visits per year by those with free care than by those in plans with cost sharing.

In standard welfare economics, of course, the additional utilisation in the free plan would be treated as a welfare loss from moral hazard (Pauly 1968). Doing so overstates the loss, since there is a gain from reduced risk in the free plan, but the overstatement is modest because of the stop loss feature (Newhouse and the Insurance Experiment Group 1993). Many, however, reject the application of standard welfare economics in this context because of consumer ignorance and agency problems (e.g. Evans 1984; Rice 1998; Cookson and Claxton 2012). To address these concerns the RHIE collected many measures of physiologic health, self-rated general health and health habits such as smoking and exercise. In this context the question was whether the additional services in the free plan translated into better health on these dimensions.

In general the answer to that question was that they did not. The one important exception was that there was better control of hypertension (high blood pressure) in the free plan among persons in the lowest 20% of the income distribution and the lowest 25% of the health distribution, which implied a 10% reduction in that group's predicted risk of future mortality. Also there was marginally better corrected vision (eyeglasses were a covered service), and there were fewer unfilled cavities (dental services were covered) in the free plan. But for most of the population the additional visits and hospital admissions in the free plan did not result, on average, in better outcomes.

Why this was so must be speculative, but, as mentioned above, those who participated in the RHIE were under 62 at the time of enrolment and were living in the community, meaning they were not institutionalised. Most such persons were reasonably healthy. Although the marginal additional services in the free plan almost certainly benefited some individuals, they could well have harmed others, either through medical error or poor quality care, such as prescribing antibiotics for a viral condition (Institute of Medicine 1999, 2001). On average, the potential benefits and harms in this population seemed to offset except for poor hypertensives, where the prior odds of benefiting from additional services were higher. Interestingly, more of the difference in blood pressure control across the plans was from a higher likelihood of a diagnosis conditional during a visit in the free plan rather than the higher likelihood of a visit.

Although undertaken to support a decision about a national health insurance plan in the 1970s, the results of the RHIE were published in the 1980s when the American political environment had changed and there was no active discussion of such a plan. As a result, the RHIE had its immediate impact on commercial health insurance. Cost-sharing in commercial health insurance increased substantially after the results were published, especially cost-sharing for hospital services; because most of those with commercial insurance were not in the lowest 20% of the income distribution, this was probably on average a welfare gain that considerably outweighed the cost of the experiment (Manning et al. 1987).

Although the data from the RHIE are now old, its results are still generally accepted as the best available estimates of the effects of cost-sharing (Congressional Budget Office 2006, 2010, 2013). Indeed, the Congressional Budget Office continues to use the RAND results to estimate the cost of various legislative proposals, including the Patient Protection and Affordable Care Act of 2010. Perhaps the RHIE results have held up despite the changes in medical treatment because much of the effect of cost-sharing appeared to be on the patient's decision to seek care at all; once having sought care, medical problems appeared to be treated similarly on the various plans in the experiment.

The Oregon Health Insurance Experiment (OHIE)

The OHIE began with a decision by the state of Oregon to expand its Medicaid program to reduce the number of uninsured in the state.

The expansion was to take place among childless adults, exactly the population that the Medicaid expansion of the Affordable Care Act targeted. Because federal law required that eligibility for an expansion be granted on a non-discriminatory basis, the state randomised a pool of uninsured either to eligibility for Medicaid or to a control group. This pool of uninsured persons was formed from those who applied for a lottery in which 10,000 winners and members of their household would be eligible for Medicaid. After a five-week intensive marketing campaign, almost 90,000 persons applied. To apply, an adult in the household had to send in a postcard with basic information such as name, address, telephone number and language preference. Alternatively, a person could transmit that information to the state using the telephone or the web.

The state sent the 10,000 winners of the lottery a packet of forms, the purpose of which was to determine their eligibility for Medicaid, meaning that the household had to describe, among other things, its income and assets. Only about 60% of those who were sent packets returned the application, and of that 60% only around half were eligible for Medicaid. In the end, only about a quarter of those who 'won' the lottery were ultimately enrolled in Medicaid. Because the randomisation took place among those who returned the postcard with the basic information, there was no assurance that those who finally enrolled, i.e. the quarter of the group who won the lottery, resembled the control group. Therefore the OHIE investigators calculated both intent-to-treat results (that is, results including all those who were randomised to the treatment group irrespective of whether they returned the forms to determine eligibility or whether they were even eligible for Medicaid) as well as results using the randomised assignment as an instrumental variable for Medicaid eligibility. The latter estimate effectively assumed that all of the effect of the treatment that was estimated in the intent-to-treat analysis came from those who ultimately enrolled and that none of it came from those who did not. Because only a quarter of those randomised to the treatment group ultimately enrolled in Medicaid, the instrumental variable estimate of the treatment effect was four times that of the intent-to-treat estimate ($4 = 1/0.25$).

Relative to the RHIE, the strength of the OHIE was that one of its two arms was an uninsured group and the OHIE thus provided a direct estimate of the effect of expanding insurance among the uninsured. The RHIE did not and could not have such a group, since it would have been both unethical and impractical to randomise families with insurance to be uninsured. In addition, the OHIE was able to collect data on the principal economic purpose of health insurance: to protect the household against large random losses of wealth. In particular, it obtained information from credit rating agencies on bankruptcies and bills sent to collection, among other indicators of financial strain. The RHIE had not been able to do that because credit rating agencies were in their infancy when it was carried out.

The OHIE, however, had three weaknesses relative to the RHIE. Whereas the RHIE had all its participants file claims forms (except for the HMO group, where utilisation information was obtained from the HMO's administrative data), the OHIE could not do this for the uninsured group. The OHIE was, however, able to obtain administrative data on hospitalisations and emergency department use from data that hospitals filed with the state for all admissions; it successfully matched these data to its participants. Information on physician office visits and pharmaceutical use, however, was limited to self-reports by participants. This had two consequences. First, data on total spending in the control group was inferred from average payments by the uninsured in a national survey. Second, data on exactly what was done during the physician visit and in particular what procedures were carried out was lacking.

A second weakness of the OHIE relative to the RHIE was less control over refusal and attrition. Both were at rather modest levels in the RHIE, because it was never in the economic interest of participants to either refuse or withdraw. The analogue of refusal in the OHIE was not returning the Medicaid application form after winning the lottery or returning the form but being ineligible for Medicaid. As mentioned above, this effectively meant that 75% of those randomised to the

treatment group were refusals. In addition, attrition both among the 25% who were actually enrolled in Medicaid as well as among the control group was an issue. If a control group household obtained insurance, for example by taking a job that offered employment-related insurance, it effectively ceased to be a useful observation. Similarly, Oregon recertified eligibility for Medicaid every six months. If because of increases in income or assets a household no longer qualified for Medicaid, it also effectively ceased to be a useful observation.

Most of the research funds for the OHIE went to a one-time effort to collect biomarkers for such indicators as blood pressure and measures of diabetes and cholesterol. Because of the loss of sample from both the treatment and control groups over time, these measures were collected at 18 to 24 months after the start of enrolment rather than the three-or five-year period of the RHIE. In addition, the state offered Medicaid to the control group after this period, so no additional data collection was carried out.

A third weakness came from the OHIE's timing. It was beginning in the field at approximately the same time as the analysis group formed to determine the information to collect; as a result, there was no time to carry out a proper baseline interview. Although this did not cause bias, it did cause a loss of power, especially for the biomarker results. Over a short period most biomarkers are reasonably stable absent medical intervention; as a result, a baseline observation on their values is especially useful in improving power. Perhaps for this reason none of the biomarkers showed a statistically significant change in the Medicaid population, although there were substantial increases in the treatment group in those diagnosed with diabetes and high cholesterol and also substantial increases in the number of (self-reported) individuals on anti-diabetic and lipid lowering (anti-cholesterol) medication.

The OHIE, consistent with the RHIE, showed that having insurance increased utilisation. Having Medicaid rather than being uninsured also favourably affected depression. (There was no biomarker for depression, only screening questions.) Consistent with the improved depression score among those with Medicaid, the point estimate implied a substantial increase in the use of anti-depressants among the Medicaid group, but the *p*-value for rejecting the null hypothesis of no difference between the groups in the use of anti-depressants was only 0.07. Some of the salient findings from the OHIE are shown in Table 1.

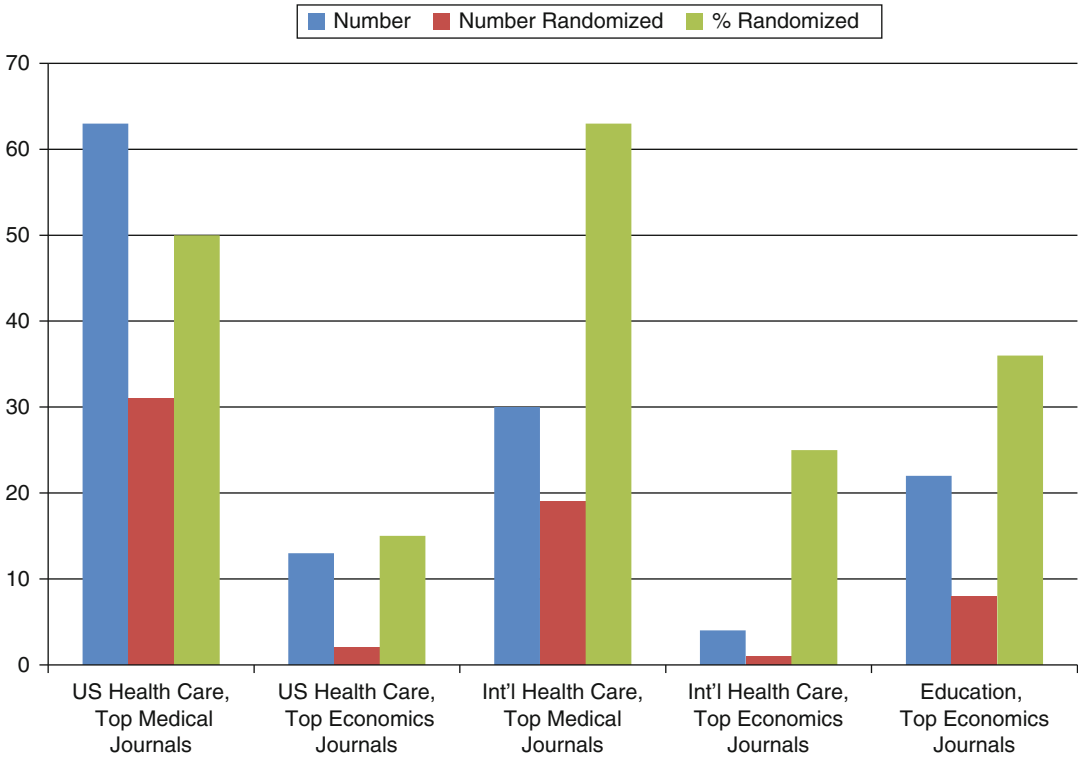
Conclusion

Experiments with the delivery and financing of health care, especially when the investigator controls the design of the experiment, as was the case in the RHIE, offer an exciting opportunity to establish the effects of various programs or

Use of Experiments in Health Care, Table 1 Selected findings of the effect of Medicaid at 18–24 months compared with being uninsured

	Mean of control group (uninsured)	Change in Medicaid group (95% CI in parentheses)
Current number of prescription drugs	1.8	0.66 (0.21, 1.11)
Physician office visits/ year (no.)	5.5	2.7 (0.91, 4.49)
Had a usual source of care (%)	46.1	23.75 (15.44, 32.06)
Out-of-pocket spending ($)	553	−215 (−409,−22)
Catastrophic spending (%)	5.5	−4.48 (−8.26,−0.69)
Depression (%)	30	−9.15 (−16.7,−1.6)
Current use of medication for depression	16.8	5.5 (−0.46, 11.45)

Notes: Catastrophic spending was defined as annual out-of-pocket spending that exceeded 30% of annual household income. The depression measure is from a questionnaire commonly used to screen for depression; consistent with the depression result, the results also showed an improvement in questions that formed the mental health subscale of a general self-rated health measure
Source: Baicker et al. (2013)



Use of Experiments in Health Care, Fig. 2 Frequency of randomised trials in health reported in leading journals, 2009–2013 (Source: Finkelstein and Taubman (2015). The top medical journals were defined as the *New England Journal of Medicine*, *JAMA*, *Annals of Internal Medicine and PLoS Medicine. BMJ* and *Lancet* were excluded because they published no studies of US health care delivery, but if they had been included the number of international studies would surely have been higher. The top economics journals were defined as the *American Economic Review*, *Quarterly Journal of Economics*, *Journal of Political Economy* and *Econometrica*.)

incentives on behaviour. But how often are experiments used?

To answer that question Amy Finkelstein and Sarah Taubman tabulated the total number of studies of health care delivery and financing in top medical and top economic journals over the five-year period 2009–2013, as well as the number of those studies that were randomised (Finkelstein and Taubman 2015). Their results are shown in Fig. 2. There were 76 studies on the US health care delivery system published in that time period in the top medical and economic journals; 33 of them were randomised. Almost all of those 33 randomised studies (31 of them) were published in medical journals, not economics journals. In the international context there were only about half as many studies of health care delivery and financing published in the same journals (34 studies), but a higher proportion of them were randomised than was the case with the American studies (20 of the 34). Like the American studies, however, almost all of the international studies were published in medical journals; only one was published in an economics journal. Interestingly, the economics of education made more use of randomised study designs than health economics; there were 22 studies of education in top economics journals, and 8 of them (36%) were randomised.

Not surprisingly, the top medical journals published many more studies of medical treatment than of medical organisation and financing in both the US and the international context, 176 and 177, respectively (not shown in Fig. 2).

Furthermore, a much higher proportion of the studies of medical treatment than the studies of financing and delivery were randomised, 79% and 77% in the American and international contexts, respectively. In sum, the randomised controlled trial is the dominant method for evaluating medical treatments throughout the world, but randomised trials or experiments in the domain of delivery and financing remain a minority of the published studies in top journals, especially top economics journals.

This paucity of randomised experiments, however, appears to be ending. Although it is not likely that anything as large as the RHIE will be carried out any time soon, there are numerous small-scale experiments being undertaken in the USA and elsewhere under the auspices of J-PAL North America. A description of these projects can be found at http://www.povertyactionlab.org/health.

See Also

- ▶ Health Behaviours, Economics of
- ▶ Health Econometrics
- ▶ Health Economics
- ▶ Health Insurance, Economics of
- ▶ Health Outcomes (Economic Determinants)
- ▶ Health State Evaluation and Utility Theory
- ▶ Population Health, Economic Implications of
- ▶ Risk Adjustment
- ▶ Statistical Analysis of Clinical Trial Data for Resource Allocation Decisions

Bibliography

Ansah, E.K., S. Narh-Bana, S. Asiamah, et al. 2009. Effect of removing direct payment for health care on utilisation and health outcomes in Ghanaian children: A randomised controlled trial. *PLoS Medicine* 6(1): 48–57.

Aron-Dine, A., L. Einav, and A. Finkelstein. 2013. The RAND Health Insurance Experiment, three decades later. *Journal of Economic Perspectives* 27(1): 197–222.

Arrow, K.J. 1963. Uncertainty and the welfare economics of medical care. *American Economic Review* 53(5): 941–73.

Baicker, K., and D.P. Goldman. 2011. Patient cost sharing and health care spending growth. *Journal of Economic Perspectives* 25(2): 47–68.

Baicker, K., S.L. Taubman, H.L. Allen, et al. 2013. The Oregon Experiment – Effects of Medicaid on clinical outcomes. *New England Journal of Medicine* 368(18): 1713–22.

Chandra, A., J. Gruber, and R. McKnight. 2010. Patient cost-sharing and hospitalization offsets in the elderly. *American Economic Review* 100(1): 193–213.

Chandra, A., J. Gruber, and R. McKnight. 2014. The impact of patient cost sharing on the poor: Evidence from Massachusetts. *Journal of Health Economics* 33: 57–66.

Choudhry, N.K., J. Avorn, R.J. Glynn, et al. 2011. Full coverage for preventive medications after myocardial infarction. *New England Journal of Medicine* 365(22): 2088–97.

Choudhry, N.K., K. Bykov, W.H. Shrank, et al. 2014. Eliminating medication copayments reduces disparities in cardiovascular care. *Health Affairs* 33(5): 863–70.

Congressional Budget Office. 2006. *Consumer-directed health plans: Potential effects on health care spending and outcomes*. Washington, DC: Congressional Budget Office.

Congressional Budget Office. 2010. *Selected CBO publications related to health care legislation, 2009–2010*. Washington, DC: Congressional Budget Office.

Congressional Budget Office. 2013. *Health-related options for reducing the deficit: 2014 to 2023*. Washington, DC: Congressional Budget Office.

Cookson, R., and K. Claxton (eds.). 2012. *The humble economist: Tony Culyer on health, health care, and social decision making*. York: York Publishing Services.

Duan, N., W.G. Manning Jr., C.N. Morris, et al. 1983. A comparison of alternative models of the demand for medical care. *Journal of Business and Economic Statistics* 1(2): 115–25.

Evans, R.G. 1984. *Strained Mercy: The economics of Canadian health care*. Toronto: Butterworths.

Finkelstein, A., and S.L. Taubman. 2015. Randomize evaluations to improve health care delivery. *Science* 347(6223): 720–2.

Finkelstein, A., S. Taubman, B. Wright, et al. 2011. *The Oregon Health Insurance Experiment: Evidence from the first year*. Cambridge: National Bureau of Economic Research.

Institute of Medicine. 1999. *To err is human*. Washington, DC: National Academy Press.

Institute of Medicine. 2001. *Crossing the quality chasm: A new health system for the 21st century*. Washington, DC: National Academy Press.

Manning, W.G., J.P. Newhouse, N. Duan, et al. 1987. Health insurance and the demand for medical care: Results from a randomized experiment. *American Economic Review* 77(3): 251–77.

Michalopoulos, C., D. Wittenburg, D.A.R. Israel, et al. 2011. *The accelerated benefits demonstration and evaluation project: Impacts on health and employment at twelve months*. New York: Manpower Demonstration Research Corporation.

Newhouse, J.P., and The Insurance Experiment Group. 1993. *Free for all? Lessons from the RAND Health Insurance Experiment*. Cambridge: Harvard University Press.
Newhouse, J.P., R.H. Brook, N. Duan, et al. 2008. Attrition in the RAND Health Insurance Experiment: A response to Nyman. *Journal of Health Politics, Policy, and Law* 33(2): 295–308.
Nyman, J.A. 2007. American health policy: Cracks in the foundation? *Journal of Health Politics, Policy, and Law* 32(5): 759–83.
Olken, B.A., J. Onishi, and S. Wong. 2011. *Indonesia's PNPM Generasi program: Final impact evaluation report*. Washington, DC: The World Bank.
Pauly, M.V. 1968. Comment. *American Economic Review* 58(3): 531–7.
Powell-Jackson, T., K. Hanson, C.J.M. Whitty, et al. 2014. Who benefits from free healthcare? Evidence from a randomized experiment in Ghana. *Journal of Development Economics* 107: 305–19.
Rice, T.H. 1998. *The economics of health reconsidered*. Chicago: Health Administration Press.
Robins, P.K. 1985. A comparison of the labor supply findings from the four negative income tax experiments. *Journal of Human Resources* 20(4): 567–82.
Rothschild, M., and J. Stiglitz. 1976. Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. *Quarterly Journal of Economics* 90(4): 629–50.
Sine, J.J. 1994. *Demand for episodes of care in the China Health Insurance Experiment*. Santa Monica: Rand. Pardee RAND Graduate School.
Taubman, S.L., H.L. Allen, B.J. Wright, et al. 2014. Medicaid increases emergency department use: Evidence from Oregon's Health Insurance Experiment. *Science* 343(6168): 263–8.

User Cost

Paul Davidson

John Maynard Keynes developed the concept of user cost as a significant component of the supply price of any business enterprise. By introducing user cost as an expectational variable, Keynes hoped to bring the existing unrealistic economic theory back 'to reality' (Keynes 1936, p. 146). Keynes believed that the concept of user cost had 'an importance . . . for the theory of value which has been overlooked' (ibid., p. 66).

Keynes's *General Theory* was based on Marshall's micro (value) theory foundations enlarged by Keynes's chapter 5 argument that entrepreneurial expectations determine output and employment. Accordingly, for theoretical completeness, Keynes had to augment Marshall's analysis of value theory with the concept of user costs, to show how profit-seeking entrepreneurs, in an uncertain world, would have 'no choice but to be guided by these expectations' (Keynes 1936, p. 46) in deciding today's employment hiring and production flow schedules.

While borrowing the name 'user cost' from Alfred Marshall (1890), Keynes's user cost notion involved components of cost different from those in Marshall's concept. Marshall believed that user cost was simply the additional 'wear and tear of plant' caused by current use of equipment compared to leaving it unused. Keynes, on the other hand, defined user cost as 'the reduction in value of the equipment due to using it as compared to not using it, *after allowing for the cost of the maintenance and improvements*' (1936), (p. 70; italics added). For Keynes, therefore, user cost was based on the idea of entrepreneur's intertemporal (profit) opportunity costs; that is, the sacrifice of expected future profits due to using equipment today rather than in the future.

Calendar time is a device which prevents everything from happening at once. Production takes time and hence profit-maximizing enterprises must make current production decisions based on expectations of future outcomes. The firm uses long-lived durable equipment in its production process; the present value of this equipment depends on expectations of the costs of future production flows *and* future sales from this equipment. The flow of production undertaken in time period t_1 will affect both the future ability of the firm to produce *and* the future market conditions it will face, and hence its ability to make profits in time period t_2 (as well as in periods further in the future).

Normally, use of any equipment in t_1 will impair its ability to render service in future periods, thereby raising future costs of production and/or affecting future investment decisions in new plant and equipment. More importantly, any rate of production and sales in the present period can often be expected to affect market demands

(and hence profit opportunities) in the future. Current profit-maximizing production decisions, in an intertemporal setting, will therefore not only involve estimates of current market conditions and current prime labour and materials costs, but they must also involve potential expected changes in future costs and market demands and hence future profit opportunities vis-à-vis leaving the plant and equipment idle.

For Keynes, user cost involved these expected changes in future profit opportunities arising from the current use of equipment in the production process. In a world where the economic horizon extends beyond a single future period, the user cost attributable to any current period production flow will be equal to the discounted (present) value of the expected greatest potential profit change due to using equipment compared to leaving it idle.

The additional 'wear and tear of plant' caused by current use equipment compared to leaving it unused was first identified, by Marshall, as *the* user cost which was associated with the prime or short-run marginal costs of production. Pigou (1933, p. 42), however, expressedly assumed that the differences in wear and tear suffered by equipment by being used in the production process as compared to being left idle could be ignored as being of 'secondary importance', so that one could assume that disinvestment in equipment through use, as opposed to time depreciation, was zero. Most economists writing in the 1920s and 1930s either followed Pigou's lead in assuming the cost of use depreciation to be negligible, or else argued that the costs of additional wear and tear would be equal to the additional maintenance costs necessary to restore the equipment to its original pre-use condition. In the latter case, marginal maintenance costs encompassed Marshall's user cost concept.

Keynes, on the other hand, believed that user cost or 'the marginal disinvestment in the firm's own equipment involved in producing marginal output' (Keynes 1936, p. 67) could differ substantially from marginal maintenance costs. Hence user cost could affect employment and production decisions more than what would be expected merely from correctly accounting for the components of maintenance expenditures into (1) those overhead expenditures for maintenance which had to be made only if the machine was idle and (2) those maintenance costs incurred only if the machine was used. Even if the potential maintenance costs incurred only if equipment is idle might offset the marginal maintenance cost of using equipment (where the latter was included in variable costs), Keynes (ibid., p. 67) insisted that it was 'illegitimate' to assume marginal user cost was zero and hence would not affect entrepreneurial production and employment decisions.

To drive home his point that marginal user costs differed from marginal maintenance costs, Keynes used what he believed was an obvious example – namely, the mining of mineral raw materials. The example used had been worked out in detail in his earlier *Treatise on Money* (1930, vol. 2, ch. 29), where Keynes demonstrated that, in a recession, the production of material goods today depended on the entrepreneurial expectations of when the market's surplus stock of materials will disappear. In the case of the production of raw materials, as Keynes put it, 'if a ton of copper is used today, it cannot be used tomorrow and the value which the copper would have for the purposes of tomorrow must clearly be reckoned as part of the [today's] marginal costs' (Keynes 1936, p. 73).

The user costs associated with the mining or production of any mineral is, Keynes (ibid., p. 73) noted, but 'an extreme case' of user costs associated with any current production flows using existing durables.

Keynes's concept of user costs highlighted the fact that when the same equipment can be used either in today's or tomorrow's production flows, then *prime production costs as well as market demand conditions are (or at least may be expected to be) time interdependent*. Hence expectations regarding this intertemporal interdependence will affect today's employment and production decisions.

Furthermore, when future economic outcomes are due to a nonergodic stochastic process so that the future is uncertain (i.e. future events are not statistically predictable based on the historical evidence), then expected changes in future profit

opportunities on the basis of expected future costs and/or future demand conditions (the user costs of utilizing equipment today) can only be guessed at. Hence, forward-looking profit-maximizing employment decisions in a nonergodic, calendar time setting, must necessarily be subjective and uncertain – and the uncertainty (nonstatistical predictability) of future economic events is one of the essential characteristics that made Keynes's monetary system operate differently from a 'real exchange' system where money was a veil. Thus, for Keynes, the user cost construction was an essential aspect for bringing economic theory 'back to reality' (Keynes 1936, p. 146) where the nonneutrality of money in an uncertain environment dominated economic decision-making processes.

If different firms foresee different future situations, even if they currently possess identical equipment, there will not be any simple, uniform profit-maximizing production decision amongst the competing firms. In such a world of nonergodic uncertainty, therefore, where different agents can, with the same historical information, perceive different futures, there can be no such thing as a unique path of optimal resource allocation over future time.

Keynes's user cost analysis, as developed in *The General Theory*, was one way Keynes attempted to introduce the reality of 'the bundle of vague and more various possibilities' (Keynes 1936, p. 24) which are the basis of entrepreneurial expectations in a statistically nonpredictable, nonergodic world; where these expectations are the determinant on today's production decisions. Keynes (1936, p. 146) argued that the assumptions of orthodox economic theory regarding the predictability of the future (what today's neoclassical economists call rational expectations, i.e. the absence of systematic errors in entrepreneurial expectations) introduced 'a large element of unreality'. By introducing the concepts of user cost and the marginal efficiency of capital (both based on expectations in an uncertain, non-ergodic world), Keynes hoped to bring microeconomic theory 'back to reality, whilst reducing to a minimum the necessary degree of adaption' (1936, p. 146).

Since Keynes's 1936 analysis, user costs have been mainly developed and applied to the question of the intertemporal production flows from depletable resources by many economists (e.g. Adelman 1972; Bain 1937; Davidson 1963; Davidson et al. 1974; Neal 1942; Scott 1953; Weintraub 1949). In general however, economists have not significantly developed the concept of user costs as an intertemporal opportunity cost in the traditional analysis of nonmineral production processes using durable equipment – despite Keynes's suggestion. Nor have many economists used the concept for the analysis of intertemporally related demand conditions.

Weintraub is the only economist who has developed the user cost concept of across-time profit opportunities with intertemporally related demands of monopolistically competitive firms (especially in a world where consumers buy often on the basis of habit, follow fashion trends or purchase replacements for their existing stock of durables). For example, production and sales this period at 'special' prices can affect future profit opportunities by altering future demands relative to future production costs. Thus, although not often recognized as such, negative user cost considerations are involved in 'introductory offer' situations where future demands are complementary to current sales (Weintraub 1949, p. 381).

See Also

► Aggregate Supply Function

Bibliography

Adelman, M.A. 1972. *The world petroleum market*. Washington, DC: Resources for the Future.

Bain, J.S. 1937. Depression pricing and the depreciation function. *Quarterly Journal of Economics* 51: 705–715.

Davidson, P. 1963. Public problems of the domestic crude oil industry. *American Economic Review* 53: 85–108.

Davidson, P., et al. 1974. Oil: its time allocation and project independence. *Brookings Papers on Economic Activity* 2: 411–448.

Keynes, J.M. 1930. *A treatise on money*, vol. 2. London: Macmillan.

Keynes, J.M. 1936. *The general theory of employment, interest and money*. New York: Harcourt, Brace.
Marshall, A. 1890. *Principles of economics*. London: Macmillan.
Neal, A.C. 1942. *Industrial concentration and price inflexibility*. Washington, DC: American Council on Public Affairs.
Pigou, A.C. 1933. *The theory of unemployment*. London: Macmillan.
Scott, A.D. 1953. Notes on user cost. *Economic Journal* 63: 368–384.
Weintraub, S. 1949. *Price theory*. New York: Pitman.

User Fees

Edwin S. Mills

Keywords
Benefit taxes; Mills, E. S.; Tiebout hypothesis; User charges: *see* user fees; User fees

JEL Classifications
H4

The genealogy of the term 'user fees' (or, synonymously, 'user charges') is neither long nor coherent. Neither Marshall nor Pigou appears to have used the term. The term was in common usage in the USA during the early post-World War II years; see e.g. Stockfisch (1960). Throughout its short history, the term seems to have been employed much more frequently in the United States than elsewhere. Since about 1970, the term has appeared in the indexes of most US public finance textbooks.

No writer has provided a careful definition of the term or distinguished it from similar terms. Rosen (1985) defines a user fee as a price charged for a commodity or service produced by a government. Some writers appear to restrict the term to charges for services produced by governments. To add confusion, many writers apply the term 'user fee' to charges levied by a government for the discharge of wastes to the air and water environment. In this usage, the term is synonymous with 'effluent fee'. Although most economists believe that governments should protect the environment by fees or regulations, since the environment has the characteristics of a public good, the environment is not in any reasonable sense a commodity or service produced by a government.

How is a user fee distinguished from two related concepts, a benefit tax and a price? There is a legal and constitutional difference between fees and taxes levied by governments, but the issue here is the economic content of the terms.

A benefit tax is any tax levied proportionately to benefits received by the taxpayer from a commodity or service provided by a government. The appropriate distinction is that a fee is paid only if the consumer decides freely to consume the commodity or service, whereas the taxpayer may be forced to pay a benefit tax even though he or she is not free to decide whether to consume the commodity or service. If the term 'benefit tax' is restricted to taxes whose amounts are no greater than the value to the taxpayer of the commodity or service consumed, the important distinction disappears. No rational consumer would refuse to pay a tax which is less than the benefit which the consumer receives from commodities or services financed by the tax. Thus, the distinction between voluntary and involuntary payment becomes unimportant. In practice, governments levy many taxes in the name of benefits even though they are larger than the benefits derived from the commodity or service provided. Most ostensibly benefit taxes are only approximations to fees for the commodity or service consumed. A gasoline tax is an approximation to a fee for road consumption or congestion and pollution externalities. The Tiebout theory (1956) implies that all local government taxes can be viewed as benefit taxes.

There seems to be no important distinction between a user fee and a price except that the term 'user fee' is used when government is the supplier. Public finance economists frequently use the term 'user fee' when reference is to a service, such as electricity, provided by government, even though the service is sometimes provided by private suppliers and the charge is then referred to as a price.

Why make the distinction between a user fee and a price? There seems to be no justification

except to identify the supplier. Yet that typically is, and always could be, clear from the context. It appears to be unjustified to coin a different, and indeed clumsy, term merely to identify the supplier. One suspects that some intellectual product differentiation is behind the distinction.

See Also

- ▶ Environmental Economics
- ▶ Externalities

Bibliography

Rosen, H. 1985. *Public finance*. Homewood: Richard D. Irwin.

Stockfisch, J.A. 1960. Fees and service charges as a source of city revenue: A case study of Los Angeles. *National Tax Journal* 13 (2): 97–121.

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

Usher, Abbot Payson (1884–1965)

William N. Parker

Keywords

Clapham, J. H.; Economic history; Inventions; Kondratieff cycles; Measurement; Schumpeter, J. A.; Technical change; Usher, A. P.

JEL Classifications

B31

Usher occupied the chair of European economic history at Harvard University from 1936 to 1949 and was surely the most productive and original scholar to occupy this post. For economists of later decades, his most significant book was *A History of Mechanical Inventions* (1st edn, 1929; 2nd edn, 1954). In it he identified invention as a four-stage process in which the individual inventor, being seized of a problem in the presence of the intellectual and physical elements for a solution, achieves the primary insight (called by Usher the 'saltatory act' and by his students the 'ah-ha!' or 'Eureka' moment) and completes the invention through a stage of 'critical revision'. Usher's work here became noticed by economists when it was taken up by J.A. Schumpeter to form the historical basis of his descriptive and theoretical work on *Business Cycles* (1939) and also through its relation to the Kondratieff 'long waves' based on the clustering of a few major inventions at discrete points in and around the 19th century (1770–80, 1840–60, 1890–1910). At a time when economists treated technological change as an element as exogenous to economics as physical geography, Usher alone thought it worth examining as a complex socio-economic 'thread' in history. In this he was the forerunner of such modern students as Schmookler, Mansfield, Ruttan, Nelson and Rosenberg, though his book largely emphasized the technical (supply-side) aspects of the process.

The identification of Usher with the study of technological change is unfortunate, since his many monographs and articles, his two textbooks, and his classroom teaching reveal a comprehensive grasp of the experience of the West in economic life and organization in all its major aspects. It is perhaps fair to say that his mind was fascinated with those points where societies face nature. Population growth, geographical resource patterns, transport, industrial location, technology, physical costs and physical constraints on social action and organization were the themes around which his view of economic history was organized. His insights were those of the engineer, not those of the sociologist.

This bias in Usher's work undoubtedly derived from a deep-seated liberal ideology which regulated both his topics and his methods of research. He looked on economic history not simply as an interesting outlet for scientific curiosity but as an instrument to help societies achieve a rational control over their environment. But, himself the most modest and unpolitical of men, he evidently saw little use to studies where the social control to which they could lead impinged on the individual's personal and private life and values.

And methodologically Usher was a committed empiricist. He evinced, and often reiterated, a deep distrust of what he called the idealistic formulations of Marx, Weber and Parsons. Yet his admiration of the British school typified by Clapham was moderated by an uneasiness over its commitment to a purely literary or descriptive methodology. He was most at home in the study of specific limited topics in which a quantifiable trend could be observed over a long period and where measurement and economic theory of a Marshallian variety could be employed. His concrete applications of the German theories and models of industrial location were particularly powerful, and inspired the later work of E. Hoover, W. Isard and others. He must be accounted, along with S. Kuznets and A. Gerschenkron as a patriarch of the so-called 'new' economic history in the United States, and of those three, Usher's grasp of the relation between theory, measurement and the phenomena of historical change must be accounted to have been the most philosophical and careful, and the best exemplified in concrete historical studies.

Selected Works

A bibliography of Usher's writings is contained in Lambie (1956). This volume also contains an essay on Usher's thought and writings. See also the article by John Dales in the *International Encyclopedia of the Social Sciences,* vol. 16 (1968), and the generous memorial tribute by A. Gerschenkron, retained in the files of the Harvard Department of Economics.

Among Usher's books and articles, *A History of Mechanical Inventions* (1929; 1954), and his two advanced level texts, *An Introduction to the Industrial History of England* (1920) and (with W. Bowden and M. Karpovich) *An Economic History of Europe since 1750* (1937), were the most durable and highly valued by students. His major contributions to early modern European economic history are *The History of the Grain Trade in France, 1400–1710* (1913) and *The Early History of Deposit Banking in Mediterranean Europe* (1943). Usher's attitudes toward economic history and methodology are best stated in three articles, (1932, 1949 and 1951) and in Chapter 4 of *A History of Mechanical Inventions*. His attitude toward economics and economic policy is well stated in his 1934 address to the American Economic Association.

1913. *The history of the grain trade in France, 1400–1710.* Cambridge, MA: Harvard University Press.

1920. *An introduction to the industrial history of England.* Boston: Houghton Miffin.

1929. *A history of mechanical inventions,* 2nd ed. Cambridge, MA: Harvard University Press, 1954.

1932. The application of the quantitative method to economic history. *Journal of Political Economy* 40: 186–209.

1934. A liberal theory of constructive statecraft (Address to the American Economic Association). *American Economic Review, Papers and Proceedings* 24: 1–10.

1937. (With W. Bowden and M. Karpovich.) *An economic history of Europe since 1750.* New York: American Book Company.

1943. *The early history of deposit banking in Mediterranean Europe.* Cambridge, MA: Harvard University Press.

1949. The significance of modern empiricism for history and economics. *Journal of Economic History* 9: 137–155.

1951. Sir John Howard Clapham and the empirical reaction in economic history. *Journal of Economic History* 11: 148–153.

Bibliography

Dales, J. 1968. Usher, Abbot Payson. In *International encyclopedia of the social sciences*, vol. 16. New York: Macmillan.

Gerschenkron, A. 1965. Abbot Payson Usher: A memorial tribute. Files of the Department of Economics, Harvard University.

Lambie, J., ed. 1956. *Architects and Craftsmen in History. Festschrift für Abbot Payson Usher.* Veröffentlichungen der List Gesellschaft, vol. 2. Tübingen: J.C.B. Mohr (Paul Siebeck).

Schumpeter, J.A. 1939. *Business cycles.* 2 vols. New York: McGraw-Hill.

Usury

Henry W. Spiegel

Usury, in the scholastic economic thought of the Middle Ages, referred to a lender's intention to obtain more in return than the principal amount of the loan. As a general rule this meant that any interest-taking was usurious and forbidden, whereas in modern parlance only exorbitant interest is considered usurious. Usury was outlawed by lay and clerical authorities, who addressed the prohibition at first only to the clergy but expanded it later to lay persons as well and repeated it frequently and in strong terms.

In the age of faith during which scholastic economic thought flourished, the authorities that outlawed interest would justify their view by reference to the Bible, several passages of which are critical of interest-taking. Another consideration, later fortified by the thought of Aristotle, was the view that money was barren – which, of course, it is if kept in a strongbox or under a mattress. Still another consideration looked at interest as a payment for the passage of time, something considered not to be the private property of the creditor. References were also made to the Roman-law distinction between fungible and non-fungible goods, the former being moveable goods that are measured by number or weight and consumed by use, such as food or fuel. Fungibles are repaid by being returned in their species rather than individually. In varying formulations and for various reasons, the scholastic authorities forbade interest on the loan of fungibles or certain fungibles. Some stressed that the borrower bears the risk of the loss of the good and is obliged to return its equivalent even if the original amount that was borrowed has been stolen or lost. Others emphasized that in the case of fungibles use and consumption coalesce and that a separate charge for use in addition to the claim for return would require payment twice for the same thing. Attention was drawn to the evil effects of usury on the community and it was held that usury violated the commands of charity, justice and natural law.

The enforcement of the canonical prohibition of usury largely relied on the conscience of the faithful, who would make restitution or abstain from interest-taking rather than die in sin and be refused a Christian burial. The frequent reiteration of the usury prohibition points to the fact that many could not resist temptation. There was even a wolf in sheep's clothing who urged Saint Bernardin of Siena on to preach more insistently against usury. Little did the Saint know that the person in question was the town's most notorious usurer, who was eager to discourage his competitors.

Under the primitive economic conditions that prevailed during the early Middle Ages the typical loan may have been a consumption loan, where the potential for the exploitation of the lender is stronger than it is in the case of production loans. During the later Middle Ages, when flourishing cities were replete with commercial activities, ways were found to secure the lender of funds a return over and above the principal. As time went on, these devices to avoid the effects of the usury prohibition became so numerous and potent as to leave the prohibition an empty shell.

To begin with, it had for long been allowed to employ interest as a means of economic warfare by charging it to political enemies such as the Saracens during the Crusades. Second, and of greater practical significance, it became an established rule that the loan contract might include a provision arranging for a conventional penalty to be paid by the borrower if he failed to return the principal at the appointed time, that is, in the case of default. Third, default itself came to constitute in time a justification for charging interest. By means of these provisions the parties to the loan contract, by arranging for very short loans and simulating default, could make interest look respectable. Fourth, it became recognized that if the creditor would suffer damage on account of the loan, having perhaps himself to borrow from others at usurious terms, he could claim compensation for the damage. Fifth, more haltingly, it was also allowed that the lender be compensated for the gain that escaped him because he granted the

loan. Thus a lender who used capital in his business could claim compensation under this title, which would legitimize a wide range of financial transactions. Sixth, although many loans were secured by pledges, an element of risk-taking was inherent in virtually all of them. In view of this consideration, interest was allowed as a risk-premium. Seventh, since the legal form in which the financial transaction was clothed was of crucial importance, the owner and prospective user of loan funds, instead of arranging for a loan, might form a partnership, with profit and loss divided among them in various ways. Eighth, persons reluctant to assume the burden of entrepreneurship could obtain a return on their money by investing it in annuities. They would turn over their funds to a private or public agency that promised to deliver them the annual return from a productive asset. Ninth, a banker might accept deposits without expressedly promising interest but rewarding the depositors with payments ostensibly in the nature of gifts. This was indeed a characteristic feature of early deposit-banking. Tenth, the parties might engage in a credit transaction involving a bill of exchange. As the name of this credit instruments implies, it used originally to be drawn on a foreign locality, with the opportunity of employing a foreign-exchange rate that would favour the creditor and yield him a return in excess of the principal. Such was indeed the origin of the bill of exchange, later as often used in domestic transactions as in foreign ones. In England, the first legal case dealing with an inland bill of exchange occurred in 1663.

The Christian society of the time persecuted and discriminated against the Jews in many ways. With their opportunities for making a living severely restricted, and with the canonical injunction not addressed to them, many were driven into money-lending at interest. Some were busy in lowly pawnshops, others served as bankers to princes and popes.

With the coming of the age of individualism and *laisser faire* the usury doctrine fell into disuse. When in the 1820s and 1830s inquiries were made with the ecclesiastical authorities in Rome as to what should be done in cases where the faithful had charged interest as allowed by the law of the land, the response invariably was that they were not to be troubled. An Irish priest, Fr. Jeremiah O'Callaghan, who insisted on the application of the original usury rule in all its strictness, was suspended from office by his bishop. In the twentieth century, the Code of Canon Law of 1917 allowed a creditor to accept the legal rate of interest and under certain circumstances even more. The Code of Canon Law of 1983 goes still farther by imposing a *duty* to pay interest when due on an administrator of ecclesiastical goods who has incurred a debt.

Secular legislation became permissive during the sixteenth century. For example, in England after the break with Rome interest up to 10% was allowed by law in 1546. After some wavering this rule was confirmed in 1571. The legal maximum was gradually reduced, but in 1854 the usury laws were abolished altogether.

The economists' reaction to the usury rule mirrored the temper of their time. Turgot, in his *Memorial on Money Loans* of 1769, poked fun at the casuistry of the scholastics and insisted that Christ in no way had intended to condemn all lending at interest. A few years later, in 1787, Bentham published his *Defence of Usury*, in which he took Adam Smith to task for endorsing legislation that put a ceiling on interest rates and in which he made a strong plea for absolute liberty in setting up the terms of loans. It is not known whether Bentham converted Smith, who, however, did not appear to be offended and sent Bentham a gift shortly before Smith's death in 1790. In our own time, the usury doctrine was defended by Keynes, who himself favoured low rates of interest. In the *General Theory*, Keynes praised the scholastics for having attempted to keep the schedule of the marginal efficiency of capital high, while keeping down the rate of interest (p. 352). However much these opinions vary, it is likely that the usury rule had the important effect of channeling funds into equity investments rather than loans. Thereby the usury rule helped to nourish a spirit of enterprise that eased the march into capitalism, the same capitalism which, in turn, brought about the usury's rule downfall.

See Also

- ▶ Aquinas, St Thomas (1225–1274)
- ▶ Just Price
- ▶ Scholastic Economic Thought

Bibliography

Baldwin, J.W. 1970. *Masters, princes and merchants*, vol. 2. Princeton: Princeton University Press, Part IV.

Consult Spiegel (1983, pp. 63–9, 696–70, with ample bibliography); Noonan (1957), the work of a legal historian; Nelson (1969), a sociological study inspired by the ideas of Max Weber; Baldwin (1970, Part IV), an historical study of the views of 12th-century churchmen, and the other works cited below. Langholm (1984) offers a new interpretation of the scholastic theory of usury on the basis of recently discovered medieval treatises.

Langholm, O. 1984. *The Aristotelian analysis of usury*. Bergen: Universitetsforlaget; distributed in the USA by Columbia University Press, New York.

Nelson, B.N. 1969. *The idea of usury*. 2nd edn, enlarged. Chicago: University of Chicago Press.

Noonan Jr., J.T. 1957. *The scholastic analysis of usury*. Cambridge, MA: Harvard University Press.

Poliakov, L. 1965. *Jewish bankers and the Holy See from the thirteenth to the seventeenth century*. Trans. M. Kochan, London: Routledge/Kegan Paul, 1977

Spiegel, H.W. 1983. *The Growth of Economic Thought*. Revised and expanded edn, Durham, North Carolina: Duke University Press.

Viner, J. 1978. Four articles on religious thought and economic society. *History of political economy* 10(1), Spring, 9–45; 46–113; 114–50; 151–89. Also available as *Religious thought and economic society: Four chapters of an unfinished work by Jacob Viner*, ed. J. Melitz and D. Winch, Durham: Duke University Press, 1978.

Utilitarianism

C. Welch

Intense, long, certain, speedy, fruitful, pure–

Such marks in *pleasures* and in *pains* endure.
Such pleasures seek if *private* be thy end;
If it be *public*, wide let them *extend*.
Such *pains* avoid, whichever be thy view;
If pains *must* come, let them *extend* to few.

Jeremy Bentham added these 'memoriter verses' to a revised edition of *An Introduction to the Principles of Morals and Legislation* to fix in the reader's mind those points 'on which the whole fabric of morals and legislation may be seen to rest' (Bentham 1789, p. 38). And indeed, although his formulation equates utility with pleasure in a way that many contemporary utilitarians would reject, Bentham does implicitly identify the central propositions that continue to inform philosophical utilitarianism today: i.e. (1) individual well-being ought to be the end of moral action; (2) each individual is to 'count for one and no more than one'; and (3) the object of social action should be to maximize general utility (or, in Bentham's phrase, to promote the greatest happiness of the greatest number).

This moral position was not, of course, original to Bentham. It was held in some form by a wide array of 18th-century writers – the English theologians Brown, Tucker and Paley, as well as the French *philosophes* Helvetius and Holbach. The distinctive doctrine associated with Bentham and James Mill, however, was first labelled *utilitarianism*. Originally coined by Bentham, and subsequently rediscovered by John Stuart Mill in a novel by Galt, the term entered the general lexicon in the 1820s. It connoted a systematic ideology composed of sensationalist psychology, ethical hedonism, classical economics, and democratic politics. Early utilitarianism – also known as Philosophical Radicalism – inspired an influential movement of reform in English law and politics during the early 19th century. But more important, the philosophy of utility as articulated by Bentham and revised by his successors has retained a central place in the theoretical debates that have dominated economics, sociology, and moral and political philosophy into the 20th century.

Bentham's Theory of Utility

Bentham's theoretical innovations were not striking; like earlier utilitarians he stated both that men are in fact pleasure-seeking creatures and that the promotion of general pleasure or happiness should be the criterion of moral goodness.

But Bentham's utilitarianism aspired to be both scientific and systematic. It derived these scientific pretensions from three tendencies that were particularly pronounced in his thought. First, he held a reductionist version of the empiricist theory of mind in which ideas – born of sensations – were formed by mental associations prompted by the urges of pleasure and pain. Bentham assumed that there was a correct association of ideas that would yield a correspondingly rationalized language. He believed that this rationalization of language was a necessary prerequisite to the proper calculation of selfinterest, and always held to the Enlightenment hope that moral language could be made scientific by purging it of irrationalities and illusions. Second, Bentham stated unequivocally that pleasure is homogeneous and thus quantifiable. He used mathematical 'metaphors' – the felicific calculus, axioms of mental pathology, the table of the springs of action – images that suggested concreteness and precision. Finally, he gave detailed and systematic attention to 'sanctions', i.e. painful disincentives to action. Unlike the theological utilitarians, he neglected the godly sanction and concentrated on those earthly penalties of public opinion and legal punishment that could be placed under the influence or control of the legislator.

Bentham's importance lay not in these refinements of utilitarianism, except insofar as they apparently strengthened its claim to certainty, but rather in his lucid and single-minded application of the doctrine to criticize the 'fallacies' of English public discourse. In this crusade he attacked both the authority of custom and the 'anarchical' philosophy of natural rights. Bentham's rhetorical assault on the French Declarations of Rights was occasioned by his recoil from the Terror, but his arguments against the language of rights remained consistent throughout his life. He makes two powerful claims: (1) rights are not anterior to political society but are created by law; hence an inalienable or non-legal right is a self-contradictory notion; and (2) a philosophy of natural rights offers no way to adjudicate the competing claims of such rights to priority; a non-legal moral right is a 'criterionless notion' (Hart 1982, p. 82). This distinction between law and morals is further developed by Austin and is fundamental to the legal positivist tradition, as well as to contemporary criticisms of rights-based moral theories.

If natural rights offered no clear theory to guide moral or social choice, utility, according to Bentham, did offer such guidance. The main body of his work lay in substituting utility for alleged logical fictions as a rationale for legislation. In his extensive writings on penal law, for example, he attempted to provide a 'calculus of harm' to facilitate the legislator's task of imposing the minimum sanction that would deter certain undesirable actions. Because Bentham's reformist ambitions encompassed civil and constitutional law, his work also touched directly on contentious public issues, such as abolition of the corn laws and reform of the suffrage. Bentham was a Smithian in economics and became a radical democrat in politics, but the logic of the original connections between utilitarianism and economic and political reform become clearer by considering the contributions of James Mill.

James Mill and Philosophical Radicalism

According to J.S. Mill, 'it was my father's opinions which gave the distinguishing character to the Benthamic or utilitarian propagandism of the time' (Mill 1873, p. 72). This propagandism was energetically carried out by a small group of self-styled Philosophical Radicals, including Francis Place, Joseph Hume, George Grote, Arthur Roebuck, Charles Buller, Sir William Molesworth, and – most important – John Stuart Mill. In a series of articles in the *Westminster Review* (beginning in 1824), they launched a political movement to begin the radical revitalization of English public life.

Bentham's work on sanctions (and some of his theoretical statements) suggest that individual interests would have to be associated 'artificially' through the manipulation of legal penalties. At the same time his faith in the general harmony between individual interests and the public interest implies that interests are harmonized 'spontaneously' (see Halevy 1903). Among Bentham's

political disciples, and largely through the influence of James Mill, this tension was resolved decisively in favour of the latter conception. Underlying the Philosophical Radicals' programme lay a dogmatic belief that the sum of enlightened self-interests would yield the general interest, in both economics and politics. It was the scientific reformer's job to attack the systematic distortions of self-interest that were charged to the account of 'King and Company', i.e. to the crown, the aristocracy, and the church.

In economics, the Philosophical Radicals endorsed the 'system of natural liberty' and the classical economic programme of competition, minimal state interference, free trade, and the abolition of monopolies. Given the rule of law necessary to produce a sense of individual security, men would be spurred to productive labour and to a rational pursuit of their interests by the operation of the natural sanctions of hunger and desire for satisfaction. Self-interested exchanges would then lead to the establishment of ever-wider markets and eventually to the production of the greatest possible satisfaction of wants. The principle of 'utility' was thus linked to an economic programme; however, the central problem of theoretical economics, i.e. the notion of 'value', was not conceptualized directly in utilitarian terms.

One could argue that there is an inherently democratic and critical dimension to the politics of utilitarianism because of the assumption that every man is the best judge of his interest, and because of the perception that individual freedom is necessary to recognize and formulate 'rational' interests. But the democratic logic of the original utilitarian radicals, put forward most forcefully in James Mill's *Essay on Government* (1820) was tailored closely to the historical problem of reforming the British aristocratic polity. James Mill argued that government is by definition rule by some group that is less than the whole 'people'. The circumstances of power, however, tempt these rulers to aggrandize themselves in a fashion neither in their own nor the people's long-term interests. They develop corporate, or in Bentham's terms, 'sinister' interests. This aristocratic corruption can be checked only through democratic representative institutions. Philosophical Radicals insisted on breaking the hold of Britian's aristocratic elite through education of the electorate, extension of the suffrage, frequent Parliaments, and the secret ballot. This sort of radicalism was distinguished from that of other democrats by its appeal to a science of politics rather than to the rights – natural or prescriptive – of Englishmen, and from that of liberal Whigs by its ahistorical and doctrinaire view of that 'science'. In the wake of the highly charged but inconclusive debates of the French revolutionary period, the appeal of a rational arbiter in politics was very attractive, especially to Britain's small emerging 'intelligentsia'. The Radicals' endorsement of the neutral standard of utility had strong affinities with the views of certain continental radicals who attempted to exorcize the terrors of the French Revolution by repudiating its language while retaining the substance of moderate republicanism (Welch 1984). In both cases, however, the reformers overestimated the attractions of their programme for the middle classes, and underestimated the possibility of the growth of a distinctively working-class consciousness. In England, the Philosophical Radicals never achieved their goal of creating a fundamental political realignment, although they clearly had an ideological impact much greater than their immediate political one.

J.S. Mill

The most famous proselytizer of Philosophical Radicalism, and its most notable apostate, was John Stuart Mill. Although Henry Sidgwick has often been called the last 'classical' utilitarian, the name can better be applied to Mill in the sense that he was the last thinker to attempt to integrate a utilitarian moral and social theory with a full-blown psychology and a theory of politics. In politics Mill came to question the iron-clad logic of his father's *Essay*, to distrust the tendency to uniformity that he perceived in democracy, and to seek a theory of counterpoise and leadership. In economics he was both the last important thinker in the classical tradition and a sharp critic of existing capitalism. But his intent in all of his

writings was, as he said, to modify the structure of his beliefs without totally abandoning the foundations.

An important discussion of the moral foundations of those beliefs can be found in *Utilitarianism* (1861). The argument here rests, inauspiciously enough, on the 'naturalistic fallacy' that underlay the work of Bentham and so many other 18th-century moralists; Mill's case for the moral worth of happiness rests on the 'fact' that people desire it:

> . . . the sole evidence it is possible to produce that anything is desirable, is that people do actually desire it . . . No reason can be given why the general happiness is desirable except that each person, so far as he believes it to be attainable, desires his own happiness (p. 44).

By 'desirable' Mill clearly seems to mean 'ought to be desired' rather than the less problematical 'can be desired'. Mill, then, was not unduly troubled by Bentham's psychological hedonism, which he largely shared, or by the derivation of ethical hedonism from this descriptive theory. Rather what bothered Mill was the suggestion that this psychological theory implied (1) a narrow materialistic view of pleasure, and (2) *egoistic* hedonism (i.e. the notion that every person ought to maximize his own pleasure). Egoistic hedonism, Mill correctly intuited, is not an ethical theory at all. To meet the first problem Mill proposed his notorious defence of qualitative differences in pleasure, a defence that only contributed to the common view that *Utilitarianism* is a casebook of logical blunders. For if there are higher and lower pleasures, it has often been pointed out, another standard than pleasure is clearly implied as the criterion of judgement between them. This tension between 'utility' and some notion of 'moral perfection' runs unresolved through most of Mill's mature works, and reappears in his defences of liberty and of democracy. To meet the second objection, Mill is careful to state, more clearly than his predecessors, that utilitarianism is a system of *ethical* hedonism, i.e. that the criterion applied to individual moral action is general happiness not individual interest. The difficult question, of course, is how to account for the motivation to moral action, given the psychological assumption that people act only to increase their own satisfactions. Mill moves away from Bentham's tendency to see the problem as one of 'conditioning' the agent to recognize the general interest as his self-interest, and offers a more sophisticated theory (reminiscent of Hume) of sympathy or disinterested altruism and its empirical connections with a sense of justice.

The power of Philosophical Radicalism as it entered the ideological arena (in a time when seismic political and industrial change had unsettled forms of social intercourse) was that it fused psychology, economics, and moral and political theory into a compelling 'fit', just how compelling a study of J.S. Mill's intellectual development would confirm. But this synthesis soon began to unravel in the hands of both friends and critics.

Utilitarianism: Reconstructions and Influence

If utilitarianism were only the doctrine of an unsuccessful 19th-century sect of reformers, it would hardly be of much contemporary interest. But as the exemplar of a 'type' of analysis, a type often held to be radically defective, it has served and continues to serve as a point of departure in discussions of economic, social and moral theory.

Utilitarianism and Economics

Utilitarianism has overtly triumphed in only one area of what were once termed the moral sciences, namely, economics. Indeed, the idea of welfare economics, i.e. of determining a 'welfare function', is irreducibly utilitarian in the sense that it seeks to measure individual wantsatisfaction and to construct indices of utility. The principle of decreasing marginal utility, which was to give a decisive turn to the evolution of modern economics when applied to the determination of value, was clearly stated by Bentham for the case of money (*Principles of the Civil Code*, 1802). Paradoxically, however, the roots of the marginalist revolution cannot be traced to the formulations of the original utilitarians in any straightforward

way. The technical innovations of Gossen, Jevons, Menger and Walras seem to have come at least in part from a greater sensitivity to the market position of consumers. Since then, the increasingly sophisticated mathematical structure of utility theory has generated many of the innovations that have dominated debates within the field.

The early marginalists, however, continued to think of utility in terms of the pleasurable sensations associated with consuming a good. They generally defended the cardinal measurability of utility; some even dreamed of a 'hedomiter' to measure it. The important theoretical break with the classical tradition was to abandon this notion of pleasure as a quality inherent in a good that could be measured in favour of a theory of choice based on the possibility of ranked individual preferences. However, although the problem underlying welfare economics is today construed differently – not as measurement of pleasure but as ranking of preferences – the analysis is still fundamentally akin to Bentham's calculus. Indeed, insofar as economists have addressed the larger issue of intellectual debts and affinities, they have acknowledged the formative influence of the classic utilitarians (see Harsanyi 1977). The issue that philosophically-inclined economists must address is that of the reach of this sort of analysis. Deep divisions remain about what sorts of issues a utilitarian theory of social choice can illuminate, and about whether the attempted solutions are morally compelling. The former issue has been posed most trenchantly by sociologists; the latter by moral philosophers.

Utilitarianism and Sociology

If the hypothesis of the rational economic maximizer has been retained in economics because of its heuristic strength in addressing a range of econometric questions, it was abandoned by the earliest of 'sociologists' because of its perceived heuristic weakness. From the beginning of the 19th century, social theorists have criticized methodological 'individualism' as incapable of generating insights into social life because such a view does not attribute constitutive power to social forces, but rather takes individual desires, purposes, and aspirations as the starting point of social analysis.

Sociology was born of the perceived problematic status of order in societies that had, at least in theory, repudiated the ties of 'tradition'. From St. Simon and Comte through Durkheim to Talcott Parsons, sociologists have singled out utilitarianism as singularly incapable of illuminating this problem. For these theorists, utilitarianism represents the notion of society conceived as a set of competing egoisms; this notion is thought to be peculiarly congenial to the English-speaking world and is often loosely and simplistically equated with liberalism. On this view, the utilitarian pedigree includes Hobbes, Locke and Smith, and its progeny the evolutionary utilitarianism of Spencer and McDougall. Durkheim's attack on Herbert Spencer (in *The Division of Labor*, 1893) can be taken as paradigmatic of the sociological critique.

Spencer was greatly attracted by organic analogies, but he applied them to social analysis in a way that radically maintained the notion that consciousness exists only in the individual 'parts' of society. He developed a strict utilitarian theory of ethics, which described the moral ideal as the individual pursuit of long-term pleasures (a calculation that involved cooperation with others through self-interested exchanges). The relative predominance of this sort of calculus over one in which individuals sought immediate gratification distinguished advanced from primitive societies. Durkheim argues that Spencer, and by extension individualist social theory, is not only inadequate but incoherent conceptually in its reliance on the notion of exchange to comprehend the patterning of social life. Formalized exchange makes sense only against the background of a culture that has internalized a particular set of social norms. Talcott Parsons takes up the theme insistently in *The Structure of Social Action* (1937). He argues that any theory which postulates the 'randomness of ends' cannot account for the ultimate reconciliation of those ends in society except by unacknowledged assumptions, sleight of hand, or a providential *deus ex machina*. Thus, on Parsons' view, an analogous function is served

by the Leviathan (for Hobbes); God and natural law (for Locke), the invisible hand (for Smith); and the necessities of evolution (for Spencer). Bentham's utilitarian policy oscillates uneasily between Leviathan and the prior assumption of a natural harmony.

According to many sociologists, then, utilitarianism as the quintessential 'individualist' social theory is fundamentally wrongheaded because individuals are defined, shaped, and constrained within social structures. Nevertheless, a reconstructed and simplified 'utilitarianism' remains the indispensable foil from which they delineate and justify the contributions of their own discipline.

Utilitarianism and Philosophy

The debate engaged between utilitarians and sociologists is between an intentionalist versus a structuralist theory of action, between a theory that heuristically treats individual preferences as random and one that emphasizes the determining constraints on those preferences. The moral philosopher engaged with utilitarianism – either as advocate or critic – has a rather different perspective and set of questions, although the philosophic criticism, especially those of 'communitarian' critics, sometimes overlap with those of sociologists. In general, however, the debates within moral philosophy take place within the camp of liberal 'individualism', in the sense that they have focused on the problems of individual moral agency. Philosophers do not ask how we can understand social order, but rather how we can judge the rightness or wrongness of individual action. The utilitarian answer (i.e. by the goodness or badness of the action's consequences) can be taken as the starting point for constructing both an analysis of moral judgements and a system of normative ethics. The utilitarian tradition of the philosophers, however, differs from that of the sociologists; it harks back to Hume and Shaftesbury rather than to Hobbes, and forward to Sidgwick, Edgeworth and Moore, rather than to Spencer. In an attempt to give a general account of moral thinking, modern philosophers have drawn on this tradition to refine ever more subtle versions of utilitarianism.

Much of this literature focuses on the arena of personal ethics. However, the public dimension – so obvious among the Philosophical Radicals who employed utility principally as an argument for or against public rules, institutions and policies – has always been implicit. Contemporary discussion of the issue occurs largely within an overlapping group of practically minded philosophers and philosophically minded welfare economists. A utilitarian theory of social justice has been explicitly argued for in the works of such thinkers as R.M. Hare, J.J.C. Smart, P. Singer and J. Harsanyi. They endorse utility, as did the classical thinkers, as the only reasonable criterion of justice in a secular society.

Philosophical Utilitarianism

There are three separate but related issues that have been crucial in the evolution of utilitarian moral theories. The first, that of justifying the imperatives of utility, has produced a measure of agreement among contemporary utilitarians and at least some of their critics. The second and third, how to decide what is a good consequence, and how to determine the right way to assess these consequences, have spawned a host of subtle distinctions that continue to preoccupy and provoke theoretical argument.

The problem of justification in utilitarianism is best approached through the work of Henry Sidgwick (*The Methods of Ethics*, 1874). Unlike Bentham or J.S. Mill, Sidgwick did not base his utilitarianism on the psychological theory that individuals always act to obtain their own good. He does argue that desirable or pleasant states of consciousness are the only intrinsic good, and that an act is objectively right only if it produces more good than any other alternative act open to the agent, but he presents these principles as moral imperatives, implicit in common sense morality, not descriptions of actual behaviour. They come to us through a sort of moral intuition that is self-evident and not susceptible of further analysis. Sidgwick narrowed the focus of utilitarianism to

a theory of moral choice, theoretically separable from any particular metaphysical doctrine, psychological theory, or political and institutional programme. He distanced himself not only from the sensationalist psychology of the earlier radicals, but also from their democratic reformism. This narrowed field is still characteristic of much, though not all, contemporary utilitarian theory. However, the arguments advanced for why we should accept utilitarian moral precepts have changed. Although he clarified the problem of justification by recognizing the illegitimacy of the slide from 'is' to 'ought', Sidgwick's own theory of moral intuitions proved extremely vulnerable.

The 20th-century analytic movement in philosophy has tended to discredit the notion of a proof of normative ethics altogether, and to disregard 'intuition' as vague and arbitrary. Nevertheless, the analytic philosopher's preoccupation with the meaning of moral language and the types of moral reasoning that are valid has led to a widespread belief that, even in the absence of epistemological certainty, good moral arguments can be distinguished from bad ones, fallacious statements from true ones. It is on this basis (greater plausibility or reasonableness) that arguments for utility are generally defended. Given some ultimate attitude that is acknowledged to be shared (usually generalized benevolence) the utilitarian hopes to convince others that his system of ethics is more plausible, that is, less prone to conceptual confusions and more coherent, than either unreflective moral sentiments, or some alternative general account of these sentiments. Insofar as some moral critics share the desire to apply to moral argument the established canons of rationality, there is common ground for discussion of the utilitarian viewpoint. John Rawls's *Theory of Justice* (1971) is developed largely through an antagonistic dialogue with utilitarianism on just this common ground.

A second issue that has been important in debates within the utilitarian moral tradition is the problem of how consequences are to be defined. A 'consequentialist' moral theory is one in which the results of action, not the motives to action, are the objects of rational assessment. Bentham, for example, stated that 'there is no such thing as any sort of motive that is in itself a bad one' (1789, p. 100). The classic discussion of this issue took place within the rubric of hedonism; pleasure – in narrow or more expansive senses – was the desired end of moral action. G.E. Moore (*Principia Ethica*, 1903), building on the dissatisfactions already expressed by J.S. Mill, offered a theory of 'ideal' utilitarianism that was consequentialist, but not hedonistic. Moore argued that pleasure was but one of many desirable goods, among which he included truth and beauty. Another answer to this question arises from the attempt to accommodate the common sense moral judgement that it is better to relieve suffering than to promote pleasure. Hence the so-called 'negative' utilitarianism attributed to Karl Popper, which argues that moral experience is uniquely concerned with the prevention of harm to others.

Among many contemporary thinkers the problem of defining the good is thought to be obviated by considering the good in terms of maximizing 'preferences'. The power of legitimation falls, in this view, on the process of choice, not on what is chosen; it is 'topic neutral'. Despite the intuitive appeal and apparent methodological advantages of this reformulation, the constraints imposed by the process of 'sum-ranking' and by the theory of rationality, as well as by common empirical assumptions about what people *do* in fact choose, lead choice-based utilitarianism inexorably back to the notion of maximizing 'well-being' or 'interest'.

The most important distinction developed within modern utilitarianism is that between 'act' and 'rule' utilitarianism, or 'unrestricted' and 'restricted' utilitarianism. This distinction has to do with the proper procedure for determining consequences. The modern statement of the problem dates back to R.F. Harrod (1936), but the intuitive sense of the distinction is quite old and is certainly present in the classical thinkers, who are usually classed as act utilitarians.

An act utilitarian assesses the rightness of an action directly by its consequences, i.e. he judges that action A is to be chosen because the total happiness expected to be produced by A exceeds

that of any alternative action open to the agent. This position has been criticized in a number of ways (for instance, it is said to hold the agent to an impossibly exigent standard of behaviour), but the most serious objections have centred on the possibility that the course of action that would be chosen on act utilitarian principles would clash violently with common sense moral judgements. Two examples, separated by two centuries, bring out the nature of this objection.

The utilitarian William Godwin (1793) argued that, if given a choice between saving one's mother from a burning building or saving a great man whose works were more likely to benefit mankind, one ought to save the great man and leave one's mother to the fire. A modern critic of utilitarianism, H.J. McCloskey (1963), offers one version of a familiar example involving not personal but public ethics. A small-town sheriff would be able to prevent serious public disturbances (in which hundreds would surely die) if he were to execute an innocent person as a scapegoat. (One could present the case, McCloskey argues, in such a way that the sheriff is certain both that his act will not be found out and that the riots will occur.) A strict utilitarian would have to recognize that, on his principles, the correct moral choice would be to kill an innocent person. Or at least he would have to recognize that such a judgement was theoretically possible. Utilitarianism, then, seems to commit one to the possibility of acting in ways abhorrent to the common sense of domestic obligation and justice. To avoid these implications, many have proposed differing versions of rule utilitarianism.

A rule utilitarian assesses the rightness of an action by asking whether it would have good consequences if it became part of general practice. Thus general rules, like 'promises must be kept', are given moral status indirectly through their role in fostering long-term utility. All utilitarians have recognized the indirect utility of rules like promise-keeping, if only as short-cuts ('rules of thumb') to the process of calculating consequences. Bentham and Mill, for example, distinguished between first-order harm and the second-order evil that comes from the example of law-breaking. However, attempts to defend a distinctive rule utilitarian position have proved problematical. Either rule utilitarianism collapses into act utilitarianism in disputed cases (e.g. when general rules conflict), or it departs from the particular utilitarian viewpoint by asserting that some rules are so necessary as to become good in themselves. Many have attempted to gain a foothold on the slippery slope between these two possibilities and the issue has generated a substantial literature.

Criticisms

One line of criticism of moral utilitarianism has always been 'technical', i.e. it has referred to the impossibility of inter-personal comparisons of utility. In 1879 a now-forgotten professor of jurisprudence argued:

> There is an illusive semblance of simplicity in the Utilitarian formula . . . it assumes an unreal concord about the constituents of happiness and an unreal homogeneity of human minds in point of sensibility to different pains and pleasures . . . Nor is it possible to weigh bodily and mental pleasures and pains one against the other; no single man can pronounce with certainty about their relative intensity even for himself, far less for all his fellows (T.E. Cliffe Leslie 1879, 45–6).

The idea that utility is cardinally measurable was basic to Bentham's enterprise, and has always been criticized on the grounds that pleasures are incommensurable. Far from resolving these problems, the economic theory of social choice has merely transposed them into different terms. Many versions of the theory depend heavily on a system of cardinalization derived from the work of von Neumann and Morgenstern on decisions taken under uncertainty. Yet these arguments have always encountered great scepticism (Georgescu-Roegen 1954). At issue is the notion of the substitutability of satisfactions. Many would argue that altruistic preferences, or preferences that are 'public' cannot be translated into preference schedules. And a persistent problem is the inability to deal in a satisfactory way with equity in distribution.

A related but more fundamental line of criticism asserts that utilitarians radically misconstrue the moral experience. If sociologists are

concerned with the alleged poverty of social insight that a theory of utility-maximizing individuals offers, moral philosophers have been haunted by the unnecessary impoverishment of those individuals, and by the narrowing and distorting of individual moral judgement. When Themistocles proposed to burn the ships of Athens' allies in order to secure Athenian supremacy, Aristides is supposed to have answered, 'The project would be expedient but it is unjust.' The fundamental insight that expedience and justice are at some level qualitatively distinct forms the essence of this critical perspective. 19thcentury critics focused on the inability of utilitarians to comprehend duties to God and country, and hence emphasized the virtues of 'excellence', 'reverence', 'nobility' and 'honour'. 20thcentury critics focus on the lack of understanding of the moral person and of duties to oneself (hence their emphasis on 'integrity', 'commitment', and 'self-respect'). Implicit in both these views is the judgement that the psychological assumptions that utilitarianism must make are so narrow and implausible as to render the theory either inadequate, or positively pernicious.

Finally, there is the problem of the cultural and institutional correlates that might accompany the adoption of utilitarianism as the criterion of social justice. Utilitarianism as a practical movement was wedded to a particular theory of politics. Yet this connection between utilitarianism and liberal democracy was largely historical and fortuitous rather than logical. The institutional implications of preference utilitarianism have not been extensively discussed, but they have aroused numerous fears and doubts among its critics. One approach to the problem is to consider again the ambiguity present in Bentham's use of the concept of interests. On the one hand, he takes interests 'as they are'. On the other, he distinguishes between existing interests, and interests that are 'well-understood'. Both conceptions have led to misgivings about the institutional implications of utilitarianism.

The idea of giving people what they happen to desire, or what they 'prefer', has much to recommend it; it seems both benevolent and non-intrusive. Yet, as social theorists have long pointed out, what grounds do we have for accepting the 'givenness of wants'? Within debates over social choice this issue has reemerged in the form of the question 'why should individual want satisfaction be the criterion of justice and social choice when individual wants themselves may be shaped by a process that preempts the choice?' (Elster 1982, p. 219). The use of existing preferences – especially given the severe restrictions on the types of preferences that can usefully be considered – may be a way of predetermining certain outcomes, of reinforcing what people regard as likely or possible in their present situation. Or so argue many critics who have seen in utilitarianism a complacent one-dimensional defense of the status quo.

Yet the concept of 'well-understood interests' (or the analogous 'true preferences') raises the question of the conditions under which these interests and preferences are revealed to be rational or true. One image that has reappeared – especially in the literature on private ethics – is the notion of the rational utilitarian floating in a sea of traditional moralists. Because the notion of a social utility function seems to imply the need for a central directing agency – an assumption itself often challenged from a pluralist perspective – the elitism implicit in the preceding image has often suggested the idea of a manipulating elite, or at best of a benevolent despotism.

Conclusion

Utilitarianism began and continues to be developed on the premise that intuitions of the divine, of tradition, or of natural law and rights have been discredited beyond rehabilitation as criterions of moral choice in a secular world shorn of metaphysics. Yet this view has always been challenged, and is today sharply contested by a resurgence of 'discredited' views. Insights into the underlying structure of social life are again sought in 'contract', 'rights' or 'community' by thinkers (one might mention such different theorists as Rawls, Nozick, McIntyre and Walzer) who argue that other traditions of thought correspond

better to the articulation of the dilemmas of moral and public life. These same theorists, however, share a preoccupation with disposing of the claims of utilitarianism as a necessary prelude to developing their own positions. Indeed, utilitarianism apparently has a special status in the evolution of modern social inquiry, not just because well-being is the modern obsession, or because the model of the 'science' of economics is seductive in an age of science, but because utilitarians claim to offer a criterion of neutrality among competing conceptions of the good life in a pluralistic and antagonistic world. Thus, to many, some version of the theory of utility has a compelling claim on our intellectual attention. If it is ultimately rejected, the imagery is nevertheless that of a 'journey away from' or 'beyond' utilitarianism (see Sen and Williams 1982). Utilitarianism has achieved a paradoxical status; it dominates the landscape of contemporary thought in the social sciences not due to its own commanding presence, but because it has been necessary to create and recreate it in order to map out the relevant terrain. Its critics claim to look forward to the day when 'we hear no more of it' (Williams, in Smart and Williams 1973, p. 150), yet it continues to figure as the alter-ego of much modern moral and social inquiry.

See Also

- ▶ Bentham, Jeremy (1748–1832)
- ▶ Chadwick, Edwin (1800–1890)
- ▶ Edgeworth, Francis Ysidro (1845–1926)
- ▶ Hedonism
- ▶ Mill, John Stuart (1806–1873)
- ▶ Pleasure and Pain
- ▶ Sidgwick, Henry (1838–1900)

Bibliography

Bentham, J. 1789. In *An introduction to the principles of morals and legislation*, ed. J.H. Burns and H.L.A. Hart. London: Athlone Press, 1970.

Bentham, J. 1802. *Principles of the civil code*, vol. 1. In *The complete works of Jeremy Bentham*, 11 vols, ed. J. Bowring. New York: Russel & Russel, 1962.

Cliffe Leslie, T.E. 1879. *Essays in political and moral philosophy*. London: Longmans, Green.

Durkheim, E. 1893. *The division of labour in society*, 1964. New York: Free Press.

Elster, J. 1982. Sour grapes – utilitarianism and the genesis of wants. In Sen and Williams (1982).

Georgescu-Roegen, N. 1954. Choice, expectations, and measurability. *Quarterly Journal of Economics* 68: 503–534. Reprinted in N. Georgescu-Rogen, *Analytical economics: Issues and problems*. Cambridge, MA: Harvard University Press, 1966.

Halévy, E. 1903. *The growth of philosophical radicalism*. Trans. M. Morris. Boston: Beacon, 1960.

Hamburger, J. 1965. *Intellectuals in politics: John Stuart Mill and the philosophical radicals*. New Haven: Yale University Press.

Harrod, R.F. 1936. Utilitarianism revised. *Mind* 45: 137–156.

Harsanyi, J.C. 1977. Morality and the theory of rational behaviour. *Social Research*, Winter. Reprinted in Sen and Williams (1982).

Hart, H.L.A. 1982. *Essays on Bentham: Studies in jurisprudence and political theory*. Oxford: Clarendon Press.

Lyons, D. 1965. *Forms and limits of utilitarianism*. Oxford: Clarendon Press.

McCloskey, H.J. 1963. A note on utilitarian punishment. *Mind* 72: 599.

Mill, J.S. 1861. *Utilitarianism*. New York: Bobbs-Merrill, 1957.

Mill, J.S. 1873. *Autobiography*. New York: Columbia University Press, 1944.

Moore, G.E. 1903. *Principia ethica*. Cambridge: Cambridge University Press.

Parsons, T. 1937. *The structure of social action: A study in social theory with special reference to a group of recent European writers*. Glencoe: Free Press.

Plamenatz, J. 1970. *The english utilitarians*. Oxford: Blackwell.

Rawls, J. 1971. *A theory of justice*. Cambridge, MA: Harvard University Press.

Ryan, A. 1974. *J.S. Mill.*. London: Routledge.

Sen, A.K. 1970. *Collective choice and social welfare*. San Francisco: Holden-Day.

Sen, A.K., and B. Williams (eds.). 1982. *Utilitarianism and beyond*. Cambridge: Cambridge University Press.

Sidgwick, H. 1874. *The methods of ethics*, 7th ed. London: Macmillan, 1907.

Smart, J.J.C., and B. Williams. 1973. *Utilitarianism: For and against*. Cambridge: Cambridge University Press.

Stephen, L. 1900. *The english utilitarians*. New York: Peter Smith, 1950.

Welch, C. 1984. *Liberty and utility: The french ideologues and the transformation of liberalism*. New York: Columbia University Press.

Utilitarianism and Economic Theory

Jonathan Riley

Abstract

Utilitarianism is a family of moral and political philosophies according to which general utility or social welfare is ultimately the sole ethical value or good to be maximized. Normative economics endorsed a hedonistic version of utilitarianism from the latter part of the 18th century well into the 20th century. Despite the ordinalist revolution, some version of utilitarianism continues implicitly to serve as the ethical basis for economic policy judgements. While there are signs that this may be changing, economic theory has not yet moved decisively beyond utilitarianism, nor is it clear that it should.

Keywords

Abundance; Bargaining; Bentham, J.; Binmore, K.; Cardinal utility; Checks and balances; Classical economics; Consequentialism; Contract curve; Contractualism; Distributive justice; Edgeworth, F. Y.; Enlightened self-interest; Equality; Ethical pluralism; Harsanyi, J. C.; Hedonistic utilitarianism; Helvetius, C. A.; Hume, D.; Individuality; Interpersonal utility comparisons; Jevons, W. S.; Leximin; Liberal democracy; Majority rule; Marshall, A.; Maximin; Mill, J. S.; Mirrlees, J.; Moore, G. E.; Natural justice; Naturalistic fallacy; Normative economics; Optimal taxation; Ordinal utility; Outcome utilitarianism; Proportional bargaining; Rational choice utilitarianism; Rawls, J.; Representative democracy; Revealed preference theory; Security; Sen, A.; Sidgwick, H.; Social contract; Spencer, H.; Subsistence; Utilitarianism; Utilitarianism and economic theory; Utility; Value; Veil of ignorance; Von Neumann and Morgenstern; Welfarism

JEL Classifications

D7

Utilitarianism is a family of moral and political philosophies according to which general utility or social welfare is ultimately the sole ethical value or good to be maximized.

Amartya Sen (1979) suggests that members of the family typically combine 'outcome utilitarianism', which, at least if population is not a variable, identifies the goodness of an outcome with the sum of individual utilities at that outcome, and some version of 'consequentialism', which focuses on selected means of achieving outcomes, such as acts, rules, dispositions, or some combination of these, and identifies the right means in any situation as those which result in an optimal feasible outcome according to outcome utilitarianism. (For analysis of the issues raised when population is a variable, see Blackorby et al. 2005). Outcomes can be defined to include the acts, rules, or other means that lead to them. Thus, for any set of possible outcomes, a typical version of utilitarianism calculates social welfare $W(x)$ at any outcome $\times$ by adding together the individual utilities at x:

$$\text{for all } x : W(x) = \sum u_i(x). \qquad (1)$$

The doctrine then prescribes as best an outcome x^* at which the sum of utilities is maximized, that is, $W(x^*) = \max \sum u_i(x)$.

As Sen also points out, outcome utilitarianism can be factorized into 'sum-ranking', the claim that the proper way to aggregate individual utilities is to add them, and 'welfarism', the principle that the goodness of an outcome is an increasing function of the set of individual utilities so that non-utility information relating to any outcome is of no ethical import. Welfarism implies 'Paretianism', the claim that one outcome is better than another if (but not only if) at least somebody has more utility whereas nobody has less utility in the one outcome than in the other. Full axiomatizations of outcome utilitarianism have been

provided by d'Aspremont and Gevers (1977) and Maskin (1978).

Different versions of utilitarianism may vary not only in terms of their consequentialist structures but also in terms of their conceptions of utility. Hedonistic utilitarianism conceives of utility as pleasure including freedom from pain, for example, whereas rational choice utilitarianism conceives of utility as a numerical representation of a preference revealed by consistent choice behaviour which is in principle observable. Sen himself arguably takes a quite restricted view of the utilitarian family because he seems to take for granted that utility, although capable of bearing plural interpretations, can be seen only as a simple phenomenon that remains the same whatever its sources and intended objects. As a result, welfarism and Paretianism cannot accommodate the idea of different aspects or kinds of utility, some of which may be intrinsically more valuable than others, where each kind is inseparably associated with a distinctive mix of human capacities, resources and/or institutional activities. Thus, as an alternative to utilitarianism, Sen (1980) introduces the family of 'utility-supported moralities', in which the goodness of an outcome still depends on utility but non-utility information is needed to distinguish among different kinds of utility and assign them distinct intrinsic values. Sen's move, although reflecting conventional wisdom, appears to be unnecessary and to that extent encourages an overly hasty dismissal of utilitarianism's potential normative appeal (Edwards 1979; Riley 1988, 2001, 2006a, 2007b, c, 2009).

Normative economics, or that portion of economic theory that evaluates institutions such as markets and policies such as tax proposals with a view to offering prescriptions for society, is not necessarily linked to any version of utilitarianism. Nevertheless, as an historical matter, economists generally endorsed hedonistic utilitarianism from the latter part of the 18th century, when both classical economics (or political economy) and classical hedonistic utilitarianism first began to emerge as systematic bodies of thought, well into the 20th century, long after the marginalist revolution of the 1870s when classical economics evolved into neoclassical economics. Even today, economists typically appeal to some version of utilitarianism, although utility is rarely defined in hedonistic terms. As Kenneth Arrow says: 'The implicit ethical basis of economic policy judgement is some version of utilitarianism' (1973, p. 246).

Jeremy Bentham (1789) and his followers, including, among others, James Mill, David Ricardo, and, for a time in his youth, John Stuart Mill, constituted the original school whose hedonistic version of utilitarianism remained dominant for so long within economics, although the doctrine was modified to some extent and combined with other ingredients largely as a result of Henry Sidgwick's influence on the early neoclassical economists such as W. Stanley Jevons, Alfred Marshall and, especially, Francis Y. Edgeworth. But the Benthamites were certainly not the first to preach a doctrine of general utility. Sidgwick traced the doctrine to Richard Cumberland's *De Legibus Naturae* (1672), for example, although Cumberland was not a hedonist. Bentham admitted that his hedonistic utilitarianism owed much to the writings of Claude Adrien Helvétius and David Hume.

Hume, like his contemporaries Francis Hutcheson and Adam Smith, asserted that human sentiments are approved of as prudent, virtuous or just only if they are seen as useful or agreeable either to the individual or to others. To be sure, none of these thinkers claimed that moral sentiments such as the sentiment of justice originate in the individual's idea of general utility. Despite various differences in their accounts, they seem to agree that an innate moral sense immediately feels a peculiar pleasure at the propriety of virtuous acts and dispositions towards other people, and that this pleasure can be made a powerful motive by suitably encouraging and educating the individual's natural sympathy for other human beings. But Bentham and Mill did not claim that moral sentiments originated in the individual's conception of general welfare either, although they rejected the notion of an innate moral sense. Bentham seems to have believed that people are predominantly motivated by self-love but that self-interest can be made to harmonize with the general welfare by giving the egoist

incentives to comply with a utilitarian legal code. The sentiment of justice reduces to cooperating with others in terms of the code, which any rational egoist will do provided he is threatened with sufficient punishment by others for non-compliance. Mill did not rely exclusively on enlightened self-interest and external punishment but instead argued (rather as Hume did, especially in the *Enquiry Concerning the Principles of Morals*, 1751) that natural sympathy for fellow human beings can be raised to such a pitch through moral education that the individual comes to feel far more pleasure when he sacrifices his self-love to the extent required for conscientious mutual cooperation in terms of utilitarian laws and customs that distribute reciprocal rights and duties.

Smith, Hume and Hutcheson may not have been fully fledged hedonistic utilitarians like Bentham and Mill, who were prepared to assert that general welfare can demand the radical reform of rules commonly alleged to be dictates of an innate moral sense. But they may be fairly depicted as what Sidgwick called 'conservative' utilitarians, who sought to explain that common sense morality is generally compatible with public utility. In any case, Hume and Hutcheson appear to have supplied crucial ingredients for Mill's utilitarianism, the one through his account of the moral sentiment of justice, the other through his distinction among different kinds of pleasure, some of which are more intrinsically valuable than others.

Benthamite Utilitarianism

Bentham's assumption that individuals are primarily self-interested fits well with economic theory. He took for granted that pleasure, including freedom from pain, is a single kind of agreeable feeling whose properties are invariant across different sentient beings. Any competent person can estimate the amount of pleasure which he or she experiences in any situation, he argued, by considering factors such as intensity, duration, certainty, propinquity, fecundity and purity. Yet he apparently did not intend to claim that much precision is possible. Indeed, as Mitchell (1918) documents, Bentham frankly admitted at times that intensity of pleasure cannot be measured and that interpersonal comparisons of pleasure cannot be based on the facts. Nevertheless, he insisted that alternative institutions and policies must be evaluated in terms of their consequences for each and every sentient being's pleasure and pain. Thus, rough estimates of aggregate net pleasure must be constructed before moral and political reasoning can even get started.

To obtain such estimates, Bentham seems to have assumed that everyone shares certain vital concerns, including security of expectations, subsistence, abundance and equality, and that different persons ought to be counted as if they are the same person when society chooses how to distribute the means of attaining these vital ingredients of anyone's happiness. The upshot is that the maximization of aggregate net pleasure becomes inseparably associated with a legal code that distributes equal rights and correlative duties for the purpose of providing the greatest total amount of security (in effect, equal marginal security) for every individual's vital concerns (Rosen 1983; Kelly 1990). These legal rights must include, among others, a right to subsistence and, consistently with that, rights to keep or trade the fruits of one's own labour and saving so as to foster abundance. Bentham's implicit premise, apparently, is that such legal rights are any rational egoist's source of his greatest amount of net pleasure, especially when duration and fecundity are taken into account, and that, to ensure universal compliance with the law, the egoist will endorse external punishment for violations of the correlative duties. Other animals can also be afforded some rights to protect their vital interests to some degree, although the power to exercise such claims must rest with humans.

At the same time, Bentham argued that the laws of property should be designed to promote an egalitarian distribution of income and wealth so far as is consistent with maximizing general security under an optimal code of equal rights. He gave priority to security, including a guarantee of subsistence as well as rights to own the means of production, over perfect equality of income and wealth as a source of pleasure for any rational

individual. But he also endorsed the assumption that any individual experiences declining increments of pleasure from additional units of money or other material assets after subsistence is assured.

According to Benthamite utilitarianism, rational hedonistic egoists must be given incentives to act as they would act if they were aiming to maximize the general good conceived in terms of security, subsistence, abundance and equality. Admittedly, egoists will not face the threat of legal penalties with respect to private conduct that is left unregulated by the public authorities. Even so, as Sidgwick (1877, 1886) concluded, Bentham seems to have maintained that enlightened self-interest is always in harmony with virtue, not only in an ideal world but also in the world of actual experience, so that vicious conduct always involves a miscalculation. If Sidgwick's reading is correct, then Bentham prescribed external sanctions to discourage miscalculations of genuine self-interest, apparently on the assumption that such mistakes are likely to be observed to any considerable degree only when the egoist holds power over others, inadequate checks exist against the abuse of power, and the unchecked power corrupts the egoist's judgement.

J.S. Mill's Qualitative Utilitarianism

Although he said that he found much of permanent value in Bentham's doctrine, Mill made clear that he wished to 'enlarge' Benthamite utilitarianism by making room for higher moral and aesthetic sentiments which Bentham had largely ignored as a result of his focus on self-interest. Unlike Bentham, Mill argued that human nature is highly plastic and that many people in civil societies are observed to have developed characters in which sympathy for others and a conscientious desire to do right are powerful motives that would restrain self-interest even in the absence of any threat of external punishment. He agreed that pleasure is the only thing of intrinsic value but maintained that there are plural kinds such that higher kinds are intrinsically more valuable than lower kinds, with the implication that competent people who experience both will not give up even a bit of the higher for any amount of the lower 'which their nature is capable of ' (1861a, p. 211).

Mill apparently classified among the highest kinds of pleasurable feeling the complex moral sentiment of justice as he understood it. As he explains in the final chapter of *Utilitarianism* (1861a), this complex feeling of security (or what others might call the feeling of freedom) can be fully experienced only by cooperating with others in terms of an optimal code that distributes equal rights and correlative duties to all individuals. In his view, this higher kind of pleasure can become such a powerful motive that the just individual rarely if ever even considers pursuing his self-interest to the point of violating others' rights. If so, Mill's pluralistic utilitarian doctrine is able to provide a more stable foundation than Bentham's purely quantitative doctrine can provide for a liberal system of weighty rights and duties. At the same time, individuals can still be permitted to freely pursue their selfish concerns in competitive markets, provided they comply with the code of justice.

Despite his qualitative gloss, which was probably suggested to him by Hutcheson's discussion of different kinds of pleasure in *A System of Moral Philosophy* (1755), Mill's doctrine remains rather similar to Bentham's, at least in broad outline. To be sure, there are important differences. In *On Liberty* (1859), Mill emphasized the importance of individuality as an ingredient of happiness, for instance, and he argued that a right to complete liberty of 'purely self-regarding' conduct is essential to promote individuality. He also went beyond Bentham by taking seriously the possibility of a decentralized socialism. Yet they agree on the need to maximize security by giving suitable priority to an optimal code of equal rights, and they also agree that institutions and policies should promote an egalitarian distribution of income and wealth so far as is consistent with the rights distributed by the code.

Moreover, like Bentham, Mill seems to have despaired of the possibility of ever acquiring data sufficiently precise to permit factual calculations of aggregate net pleasure. Rather, Mill argued that the test of quantity as well as quality of any two

pleasures or pains was the unanimous judgement of competent persons who had experienced both, or, in case of disagreement, the majority judgement of such persons, where judgements are apparently assumed to be nothing more than preference orderings defined over the relevant net pleasures. By implication, for any pair of possible outcomes x and y, any competent individual i who estimates that the quantity of net pleasure $u_i(x)$ which he will experience at x is at least as great as the quantity of the same kind of net pleasure $u_i(y)$ which he can expect at y, forms a weak judgement or preference R_i defined over pleasures which then, by virtue of hedonism, determines his preference over outcomes:

$$\text{for all } x, y : u_i(x) R_i u_i(y) \rightarrow x R_i y, \qquad (2)$$

where R_i includes an asymmetric factor P_i denoting that i forms a strict preference if he estimates that $u_i(x) > u_i(y)$, and a symmetric factor I_i denoting that he is indifferent when he estimates that $u_i(x) = u_i(y)$.

The information about quantities of pleasure is no more precise than that contained in the fallible individuals' estimates, even if any competent person makes use of Bentham's guidelines to consider intensity, duration and so forth, and also accepts such common psychological generalizations as declining marginal pleasure of income beyond some threshold of subsistence. Given the Benthamite dictum that 'everybody is to count for one', aggregating over the individual judgements largely boils down to majority rule, which is consistent with the strongly majoritarian forms of representative democracy defended by Bentham in his *Constitutional Code* (1830) and by James Mill in his *Essay on Government* (1820). Yet majority rule does not rely on interpersonal comparisons of pleasure. Strictly speaking, each person's judgement might be counted as one by being represented by the same interpersonally comparable ordinal utility function as everyone else's in the modern economic sense of utility. In this case, even though any positive monotonic transformation of the single utility function used to represent each person's estimates of pleasure is permissible, adding the utility numbers to form hedonistic utilitarian judgements will usually (but not always) select majority winners if such outcomes exist and always resolve majority preference cycles when they occur (Riley 2007c).

J.S. Mill also argued that any competent individual who judges that the kind of pleasure which he will experience at x is higher in quality than the kind of pleasure which he can expect at y, would never sacrifice even a bit of $u_i(x)$ for any amount of $u_i(y)$. In effect, the two kinds are incommensurable and cannot be traded off against one another in terms of the same scale of value. If the pleasure of the moral sentiment of justice is higher in quality than the enjoyable feelings of ordinary expediency, for instance, then a competent individual refuses to trade off even a little of his or anyone else's security of equal rights for, say, any quantity of enjoyment associated with ill-gotten income. The kind of painful insecurity associated with violating rights always outweighs the amount of merely expedient pleasure that might be gained by means of the violation. Because he feared that the popular majority might not be sufficiently educated to be competently acquainted with the pleasures of equal justice, Mill was less inclined than James Mill, his father, and Bentham were to support majoritarian democracy. Indeed, he seems to have been impressed by Thomas Macauley's objections that the Benthamites' methodology was flawed in so far as they attempted to deduce political conclusions on the assumption that individuals are rational self-interested agents without emotional ties to existing institutions (Lively and Rees 1978). A more historical approach was needed, one that recognized the danger of majority tyranny and did not ignore as irrational those traditional institutions, attitudes and practices that might help to combat it. Thus, in *Considerations on Representative Government* (1861b), Mill argued for a limited form of representative democracy in which a distinctive system of checks and balances is designed to give more power to highly educated minorities, whose special expertise is generally acquired in traditional ways outside the democratic political system, to promote competent government and security of equal rights (Riley 2007a).

Nevertheless, Mill's qualitative hedonism was generally dismissed as incoherent, and his

model of liberal democracy was rejected as undemocratic. Under the influence of leading philosophers and economists, utilitarianism took a turn towards false quantitative precision, and the links with democracy were obscured.

Sidgwick and the Early Neoclassical Economists

Sidgwick (1874), who billed himself as a hedonistic utilitarian, played a central role in all of this. He was in fact a rational intuitionist who argued that utilitarianism presupposed certain rational intuitions such as the axiom that an ethical agent must be impartial between one person's pleasures and another's. His argument muddied the hedonistic utilitarian tradition, which held with Hume that intrinsic value is not a function of a priori reason but rather of feelings of pleasure wherever located, whose origins and properties can in principle be inferred by reason only on the basis of experience. Moreover, although he claimed to be a strong admirer of J.S. Mill, Sidgwick asserted that a consistent hedonism must be purely quantitative because qualitative hedonism implicitly relies on some intrinsic value besides pleasure to distinguish among different qualities of pleasure. His assertion was apparently accepted as gospel by his friends Jevons (1874), Edgeworth (1877, 1881) and Marshall (1884). His student Moore later repeated it in the course of accusing Mill of spreading 'contemptible nonsense' (1903, p. 72). Yet Sidgwick and his followers merely beg the question against Mill because they assume that pleasure is a simple agreeable feeling that always exhibits the same properties. They never consider the possibility that pleasure is a term that covers a family of agreeable feelings, some of which are intrinsically more valuable than others.

Sidgwick also raised serious doubts about the rationality of a purely quantitative hedonistic utilitarianism by arguing that practical reason is 'divided against itself' because it cannot resolve basic conflicts between rational egoism and rational benevolence in some situations. This 'dualism of practical reason' is said to be manifested when a reasonable wish to preserve one's own life or other vital interests that ought to be protected by rights apparently collides with reasonable utilitarian duties to promote others' welfare by sacrificing one's own life and enjoyments. Some such ethical dualism or pluralism is now a staple of the philosophical literature. Many argue that genuine maximizing utilitarianism conflicts with individual rights so fundamentally that an impartial rational resolution of the conflict is impossible.

Moore went even further than Sidgwick and insisted that a 'naturalistic fallacy' is committed if intrinsic goodness is defined to be synonymous with pleasure or desire-satisfaction or any other natural property. Rather, goodness is said to be an indefinable non-natural quality that emanates in mysterious fashion from certain complex 'organic wholes' consisting of plural natural ingredients including, perhaps, pleasure as an essential ingredient, although Moore struggled to make up on his mind on this point (see, for example, Edwards 1979). But Sidgwick rejected Moore's anti-hedonistic view that ideals of goodness are directly intuited by anyone capable of appreciating the relevant 'organic unities'. Indeed, without defining goodness to mean pleasure, Sidgwick accepted that pleasure including freedom from pain might ultimately be found to be the only thing of intrinsic value. He endorsed Bentham's 'empirical method' of ascertaining quantities of pleasure and pain, albeit reluctantly given the difficulties which he saw in applying that method. He also argued that quantitative hedonistic utilitarian reasoning accords with the bulk of common moral intuitions, upon which it can thus legitimately rely as rules for calculating right and wrong actions, and called for formal models of a utilitarian calculus under ideal conditions to help clarify its implications.

Early neoclassical economists, especially Edgeworth, answered the call and used the tools of mathematical calculus to formulate ideal versions of quantitative hedonistic utilitarianism. Unlike Bentham or Mill, Edgeworth assumed that natural units ('just perceivable increments') of pleasure and pain can in principle be ascertained and aggregated over varying populations and time horizons. Rather than consider individuals' judgements of the quantities of net pleasure to be expected from

feasible options as in Eq. (2) above, he imagined an ideal situation in which the quantities are definitely known such that, for any individual i, a unique natural utility function can be specified which, by virtue of hedonism, determines what i's preferences over outcomes ought to be:

$$\text{for all } x, y : u_i(x) \geq u_i(y) \rightarrow xR_iy. \qquad (3)$$

There is no longer any need for i's estimates of pleasure because the utility numbers are assumed to be already known with fantastic precision. There is not even any need for i to form or express his preferences over outcomes. Person i's natural utility function indicates the exact amount of interpersonally comparable natural pleasure which i can reasonably expect to experience at any given option, whether or not i appreciates this fact. Any competent 'impersonal observer' with the requisite individual utility information can then determine a best option by adding up the unique individual utility numbers at each option to see which option has the greatest sum total of pleasure net of pain. A utilitarian calculus is thus no longer necessarily linked to majoritarian aggregation procedures as it was with Bentham and Mill. Indeed, an authoritarian elite might perform and enforce the utilitarian calculations. Sidgwick and Edgeworth seem to have been surprisingly open to the possibility of some such utilitarian elite.

The marginalists also found more or less ingenious ways to overcome Sidgwick's 'dualism of practical reason'. They imported the idea of an evolutionary process as they understood it, whereby ignorant and selfish individuals might eventually evolve into intelligent and virtuous ones through cultural and (as Herbert Spencer suggested) even biological transmission of the relevant concepts and dispositions. Jevons went so far as to endorse the Hegelian notion that this evolutionary solution was under divine direction. Edgeworth did not go so far. Yet, like Jevons, he took for granted that the more highly evolved members of a refined minority have greater capacities than the masses do for pleasure (the minority can enjoy 'higher pleasures' in the sense of larger quantities of pleasure viewed as a single kind of agreeable feeling), and he emphasized that utilitarianism does not necessarily imply equal distribution of the 'means of pleasure' or of the work required to produce the means.

Moreover, Edgeworth was specific about the way in which Sidgwick's 'dualism of practical reason' could be overcome in the economic arena. After calculating the 'contract curve' of Pareto-efficient allocations that self-interested traders could willingly negotiate and enforce as contracts, and showing that the contract curve converges on a perfectly competitive equilibrium only as the number of bargainers goes to infinity, he stressed the indeterminacy faced by any finite number of bargainers in selecting among efficient contracts. He then suggested that under certain conditions, including equal prior probabilities of any particular efficient contract being selected, the selfish bargainers would agree to accept a utilitarian contract as a just compromise because it is one of the efficient options on the contract curve, ignoring instances where the utilitarian bargain might make an individual worse off than his initial endowment (Creedy 1986, pp. 79–92; Newman 2003, pp. xxxvii–xlvii). This contractualist argument for utilitarianism is interesting not only because it anticipates John Harsanyi (1977, 1992) but also because Edgeworth was well aware that a utilitarian bargain is typically distinct from a competitive equilibrium and thus calls for redistributive measures.

The Ordinalist Revolution

By the turn of the 20th century, hedonism was under siege in both psychology and philosophy as evolutionary psychology and behaviourism came to the fore along with philosophical idealism, pragmatism and ethical pluralism. Hedonism was largely abandoned within economics during the ordinalist revolution of the 1930s and 1940s, whose leading figures included Lionel Robbins, John Hicks, R.G.D. Allen, Abram Bergson and Paul Samuelson. The ordinalists, who registered doubts about meaningful interpersonal comparability as well as cardinal measurability of utility, recognized that the analysis of efficient allocations does not require either a hedonistic theory

of motivation or rich interpersonally comparable cardinal utility information. They redefined the concept of utility to denote not pleasure but rather a formal numerical representation of any preference ordering revealed by consistent choice behaviour, without reference to the motivations or reasons underlying the revealed preference. In effect, utility merely represents what the agent is disposed to choose, independently of any psychological explanation or ethical justification for the given dispositions. Thus, in contrast to Eqs. (2) or (3) above, any individual *i*'s utility function is independent of hedonism and simply recapitulates the information which is contained in *i*'s revealed preference ordering over outcomes:

$$\text{for all } x, y : u_i(x) \geq u_i(y) \leftrightarrow xR_iy. \qquad (4)$$

Moreover, any positive monotonic transformation of the utility function is also an admissible utility function because such transformations preserve the information in the preference ordering which is being represented.

Given the restriction to such impoverished individual utility information, purely ordinal and bereft of any ethical standards, it is hardly surprising that there was an impulse to push aside issues of distributive justice, relegating them to other disciplines such as political philosophy. Normative economics could then concentrate on relatively uncontroversial Pareto efficiency considerations, including clarification of the conditions under which a general competitive equilibrium is efficient, for instance, and of the different conditions under which an efficient allocation can be achieved as a competitive equilibrium (cf. the 'fundamental theorems' of welfare economics).

Arrow's (1951) 'general impossibility theorem' may initially have reinforced the impulse to sidestep distributive issues because the theorem shows in effect that a social welfare function must reflect the preferences of a dictator if it is required to rely exclusively on purely ordinal utility information to generate rational social or moral choices from any set of distinct individual preferences. Yet normative economics cannot plausibly ignore distributive justice and the consequent need for interpersonal comparisons. In this regard, Arrow's negative result has also stimulated a large literature in which social choice theorists and game theorists have clarified many different forms of social decision functions and games, including their various informational requirements, and thereby clarified alternative theories of justice and morality which might be employed within normative economics. In addition to Arrow himself, Amartya Sen (1970, 1982, 2002, 2007) has played a leading role in this literature. But numerous others, including John C. Harsanyi (1955, 1977, 1992) and Kenneth Binmore (1994, 1998, 2005), have made noteworthy contributions.

It remains unclear what impact this ongoing literature will ultimately have on normative economics. Perhaps, as has been suggested (Mongin 2006; see also Mongin and d'Aspremont 1998), another revolution is under way, what might be called a non-utility revolution in so far as the thrust of it is to argue that a social welfare procedure ought to rely at least in part on information about the outcomes which is not reflected in individual preferences. Yet economists continue to defend and employ versions of utilitarianism. Indeed, Ng (1975, 1985) has made a case for redeploying within normative economics a so-called Benthamite doctrine that brings back Edgeworth's notion of 'just perceivable units' of pleasurable feeling. But economists have tended to adopt rational choice versions of utilitarianism which, unlike the forms of utilitarianism typically discussed by modern philosophers such as Hare (1981), Ng and Singer (1981), Gibbard (1987, 1990) and Brandt (1992), do not tie the idea of utility to pleasure, desire-satisfaction, norm-expression, or any other motivation. Harsanyi's doctrine is a leading example (for a cogent summary of its main features, see Hammond 1987).

Harsanyi's Rational Choice Utilitarianism

Harsanyi defines utility as merely a numerical indicator of any preference revealed by rational choice behaviour but he also argues that utility functions can be viewed as cardinal and interpersonally comparable rather than purely ordinal. He builds various conditions into his idea of what

constitutes fully rational and moral behaviour so that meaningful interpersonal comparisons can be made of the gains or losses of utility that represent relative intensities of revealed preferences, that is, how much one outcome is preferred or opposed relative to another. Technically, this implies that a person's utility function can be subjected to any positive linear (but not nonlinear) transformations but, if any person's utility function is transformed, then every other person's must also be transformed in the same way. Such interpersonally comparable cardinal utility information is needed to operate a rational choice utilitarian calculus in anything like the manner imagined by neoclassical economists such as Edgeworth.

Like John Rawls (1971, 1993), Harsanyi assumes that rational moral agents will imagine themselves in an original position under a veil of uncertainty about their particular social circumstances in order to calculate a fair social contract, that is, a conception of justice in terms of which they mutually agree to cooperate despite their different personal preferences. Unlike Rawls, however, Harsanyi assumes that rational choice behaviour under risk and uncertainty conforms to standard expected utility theory. He argues that individual attitudes towards risk should be used to infer personal preference intensities over outcomes, in which case the von Neumann–Morgenstern method of cardinalization becomes appropriate and a cardinal expected utility function represents an individual's revealed preferences under risk and uncertainty. Also, a moral agent must become an impersonal observer by forgetting his actual circumstances and imagining that he has an equal chance of occupying any person's social position with that person's preference intensities over outcomes. To avoid double counting, any impersonal observer must also ignore what Ronald Dworkin (1977, p. 234) calls 'other-oriented' preferences defined over other persons' positions. All impersonal observers are guaranteed to make identical interpersonal comparisons and moral choices because human beings supposedly share a fundamentally similar psychology, which is left unspecified, though hedonism is rejected as naive. Finally, impersonal observers will jointly choose to constrain themselves, Harsanyi claims, by making a binding commitment to the same optimal code of moral rules. A person whose revealed preferences satisfy all these conditions is behaving as if he were a rule utilitarian, whatever his desires, feelings or other motivations might be.

Harsanyi's theory is vulnerable to various serious objections. It is not clear that revealed attitudes towards risk measure intensity of subjective feelings or that they would have much normative significance even if they did, for instance, even when individuals can be assumed to be concerned exclusively with outcomes and to ignore the process of assigning probabilities ('gambling') per se. Also, as Diamond (1967) has pointed out, Harsanyi's claim that moral and social utility must be a linear function of the individual utilities seems overly rigid because it ignores the distribution of the utilities. Rather than define morality so that it always demands the simple addition or averaging of individual utilities, a more appealing approach would insist that the process of moral and social decision-making should give everyone a 'fair shake'. Fairness might be thought, for instance, to require a quasi-Rawlsian concern to maximize the utility of those individuals or groups with the worst utility levels in comparison to others. (For further discussion of this quasi-Rawlsian approach known in the literature as 'leximin', see, for example, Hammond 1976, 1977; and Deschamps and Gevers 1978. Unlike Rawls's maximin theory, which works in terms of 'primary goods' that every rational person is presumed to want, the leximin theory works in terms of utility levels.)

Doubts about the von Neumann–Morgenstern method of cardinalization and the fairness of sum-ranking have prompted some leading economists, including Arrow, Diamond and James Mirrlees, to take seriously an ordinalist variant of rational choice utilitarianism.

Ordinalist Utilitarianism

Ordinalist utilitarianism presupposes that interpersonally comparable ordinal utility information is available. In other words, it must be assumed

that meaningful interpersonal comparisons can be made of the levels of utility which represent revealed preference orderings. Technically, this implies that a person's utility function can be subjected to any positive monotonic transformations, but if any person's utility function is transformed, then every other person's must be similarly transformed.

An ordinalist utilitarian calculus is quite flexible in terms of its functional form. In contrast to Eq. (1) above, social welfare is calculated as the sum of utilities with a particular concave transformation $f\{.\}$ applied to each person's utility function:

$$\text{for all } x: W(x) = \sum f\{u_i(x)\} = f\left\{\sum u_i(x)\right\}. \quad (5)$$

But $f\{.\}$ may be subjected to any positive monotonic transformations without affecting W because the aggregation process is meaningful only for ordinal (rather than cardinal) comparable utility information. Arrow (1973) has shown that this ordinalist utilitarian approach subsumes the quasi-Rawlsian leximin theory of distributive justice as a special case, namely, the case in which the concavity of $f(.)$ is extreme because each person is assumed to be highly risk-averse. Another important application is in the modern theory of optimal taxation (Mirrlees 1971, 1982; Diamond and Mirrlees 1974; Stiglitz 1987). Indeed, optimal tax theory can be viewed as the further development and refinement of a body of thought that includes Edgeworth (1897) as well as Mill's and even Bentham's recommendation that a tax system ought to satisfy a principle of equal marginal sacrifice above some threshold of subsistence guaranteed for all.

Nevertheless, ordinalist utilitarianism also seems vulnerable to serious objections. Arrow remains reluctant to accept interpersonal comparisons of ordinal utility, for example, because he fears that they imply a denial of individuality: 'the autonomy of individuals, an element of mutual incommensurability among people seems denied by the possibility of interpersonal comparisons' (1977, p. 225). Individual rights to freely pursue one's own good in one's own way, of the sort defended by Mill, seem to be of peculiar moral importance, and should not be overridden by considerations of general utility based on putative interpersonal comparisons. Moreover, as Arrow also remarks, it is disappointing that, even if meaningful ordinalist comparisons are assumed possible, seemingly mild conditions (including a weak equity axiom) confine us to the quasi-Rawlsian 'leximin' theory of justice. But leximin is arguably too extreme because it absolutely forbids institutions and policies that fail to maximize the utility level of the worst-off, even if those measures result in massive utility gains for everyone else. Finally, as Gibbard (1987) suggests, ordinalist utilitarianism, like any other version of rational choice utilitarianism, needs an 'empirically adequate' psychology as well as a convincing theory of ethical deliberation, even if hedonism continues to be rejected as both a psychology and an ethical theory. Grounds for right action cannot simply be equated with what people are disposed to choose. Rather, utility, and thus welfarism and Paretianism, must be tied to a normative theory that identifies any individual's morally significant interests in any given social context and also justifies which choice dispositions ought to be formed to achieve the relevant interests in this or that situation. Such a normative theory must in turn be tied to a psychology that supplies empirically adequate psychic concepts and explains how dispositions to choose are formed in terms of the concepts.

In light of such objections to utilitarianism even in its ordinalist guise, some leading economists and philosophers see the need for a non-welfarist theory that makes use of non-utility information to evaluate possible outcomes. Although there are many different ways to go beyond utilitarianism and welfarism, two of the most interesting ways have been proposed by Binmore and Sen, respectively.

Beyond Utilitarianism?

Binmore, inspired by his reading of Hume, presents a provocative proportional bargaining theory of 'natural justice', which, like Harsanyi's

utilitarian theory, relies on cardinal comparable utility functions to represent the preferences revealed by moral choices. Unlike Harsanyi, though, Binmore assumes that moral agents remain predominantly selfish and are inclined to cheat on bargains unless they are threatened with sufficient punishment by others for non-compliance. He builds inequalities of bargaining power into his theory of moral behaviour by assuming that even bargainers who place themselves in an original position under a veil of ignorance will rely on cultural standards of interpersonal comparison which are associated with a given Nash bargaining equilibrium. His theory thus relies in part on non-utility information. The relevant Nash bargaining outcome supplies the cultural norms which agents in the original position use to calculate a proportional bargaining equilibrium, for instance, and it has a privileged status as a fallback position if these agents fail to agree to coordinate on the proportional bargaining solution. As a result, Binmore concludes that, in general, utilitarian bargains would not be willingly enforced by rational expected utility maximizers seeking justice.

Binmore apparently assumes that people are predominantly selfish because human behaviour is ultimately constrained in accord with the selfish gene paradigm. But there is no compelling scientific evidence for that paradigm. Rather, human nature appears to be highly plastic. If so, rational agents might eventually be moulded by cultural forces into social and moral actors who effectively believe that they are the same person – no different from anyone else – when it comes to certain vital personal interests that ought to be treated as rights. In this context, a utilitarian bargain, involving some code of justice that distributes equal rights and correlative duties whose content is held by the culture to be extremely valuable for every person's well-being, is an efficient and fair Nash equilibrium point. Even in the absence of any threat of punishment by others, compliance with the rules is enforced by the actor's own conscience, a powerful internal 'judicious spectator' which threatens to inflict harsh punishment in the form of intense feelings of guilt for cheating. Indeed, there is textual evidence that Hume himself took seriously this possibility of utilitarian justice (Riley 2006b).

Sen also advocates the use of non-utility information in the course of proposing a pluralistic theory of ethical evaluation and distributive justice that involves protecting a list of basic human functionings and capabilities or 'freedoms' for each member of society (Sen 1985, 2002, 2007). Moreover, his Paretian liberal impossibility result (Sen 1970) shows that a social welfare function cannot in general simultaneously satisfy utility-based Paretian values and non-utility-based liberal rights as he conceives them. His work contains deep insights into the sort of moral and political philosophy which normative economics seems to require, and it also casts serious doubt on whether any version of welfarism can accommodate such insights.

Nevertheless, it is not entirely obvious that a move beyond utilitarianism is required, especially if the utilitarian family is defined (as it arguably should be) to include what Sen calls 'utility-supported moralities'. A case can be made, for example, that genuine maximizing utilitarianism involves an optimal code that distributes weighty equal rights and correlative duties, as many leading utilitarians, including Bentham and Mill as well as Brandt, Hare, and Harsanyi, have insisted (Riley 2006a). Moreover, it may even be possible to combine welfarism with a normative theory that interprets any individual's morally significant interests in terms of functionings and capabilities, the most vital of which ought to be protected by strong rights. A moral individual is presumably able to form dispositions to make the relevant moral choices, at least after engaging in a process of impartial deliberation about the functionings and capabilities of the different people implicated in any choice situation. Consistent ethical behaviour may reveal preference orderings that can be represented by utility functions. If so, an ethical version of welfarism could perhaps subsume Sen's ethical theory. The more general point is that, by itself, rational choice welfarism is essentially a formal shell that needs to be filled in with a substantive psychology and ethical theory. It can be filled in with various theories besides hedonism, to which it has no essential tie (Riley 2001).

Normative economists evidently face numerous competing candidates when attempting to select a 'reasonable' social welfare procedure that is both efficient and fair. Yet ordinalist utilitarianism perhaps has more appeal than has so far appeared. Objections similar to Arrow's are often voiced against any version of utilitarianism (see, for example, Smart and Williams 1977; Sen and Williams 1982; Scheffler 1982). Yet the objection that even ordinalist utilitarianism cannot accommodate reasonable principles of distributive justice arguably turns on an improper formulation of interpersonal comparisons of utility. As usually formulated in terms of extended sympathy, where one person places himself in another's position and imagines that he experiences the same psychic phenomena as the other experiences in that position, interpersonal comparisons involve the sort of double counting against which Harsanyi and Dworkin caution (see also Gibbard 1987, p. 144). Reformulating interpersonal comparisons so that they do not involve double counting of any person's utility can liberate ordinalist utilitarianism from being chained to 'leximin'. Arrow's other worry – that the possibility of interpersonal comparisons seems to deny individuality and personal integrity – might also be met by limiting the permissible scope of interpersonal comparisons so as to preserve weighty rights of individuality or self-development. Such an ethical limitation may be endorsed by utilitarians if a code that distributes such rights and correlative duties is deemed essential to the maximization of general utility.

Gibbard rightly objects that ordinalist utilitarianism by itself can hardly provide an acceptable ethical theory if utility merely represents given dispositions to choose, independently of any ethical justification for the dispositions. He advises that ethical thinking must rely on 'rough quantitative' estimates of which outcomes are 'more worth wanting' than others: 'we should settle for some vagueness and indecision, epistemological and normative' (1987, p. 148). This is wise advice, although it may be possible to combine ordinalist utilitarianism with an appealing ethical theory such that 'some vagueness and indecision' surrounds the 'psychic magnitudes' used in ethical thinking whereas utility functions represent revealed preferences that reflect the relevant ethical thinking. Indeed, a form of qualitative ordinalist utilitarianism along Millian lines deserves further study (Edwards 1979; Riley 1988, 2007b, c, 2009). Such a qualitative ordinalist utilitarianism might incorporate what is taken to be Mill's theory of ethical thinking, including his suggested order of priority among intrinsically different kinds of ethical judgements, without making any commitment to psychological or ethical hedonism. Eventually, though, the time may come when some version of hedonism is again taken seriously (see, for example, Kahneman et al. 1999; Feldman 2004).

See Also

- ▶ Arrow, Kenneth Joseph (Born 1921)
- ▶ Bentham, Jeremy (1748–1832)
- ▶ Edgeworth, Francis Ysidro (1845–1926)
- ▶ Final Utility
- ▶ Happiness, Economics of
- ▶ Harsanyi, John C. (1920–2000)
- ▶ Interpersonal Utility Comparisons
- ▶ Mill, John Stuart (1806–1873)
- ▶ Mirrlees, James (Born 1936)
- ▶ Optimal Taxation
- ▶ Sen, Amartya (Born 1933)
- ▶ Sidgwick, Henry (1838–1900)
- ▶ Welfare Economics

Bibliography

Arrow, K.J. 1951. *Social choice and individual values*, 2nd rev. edn., New Haven: Yale University Press, 1963.

Arrow, K.J. 1973. Some ordinalist-utilitarian notes on Rawls's theory of justice. *Journal of Philosophy* 70: 245–263.

Arrow, K.J. 1977. Extended sympathy and the possibility of social choice. *American Economic Review* 67: 219–225.

d'Aspremont, C., and L. Gevers. 1977. Equity and the informational basis of collective choice. *Review of Economic Studies* 46: 199–210.

Bentham, J. 1789. In *An introduction to the principles of morals and legislation*, ed. J.H. Burns and H.L.A. Hart, 1970. London: Athlone Press.

Bentham, J. 1830. In *Constitutional code*, ed. F. Rosen and J.H. Burns, Vol. 1, 1983. Oxford: Oxford University Press.
Binmore, K. 1994. *Playing fair*. Cambridge, MA: MIT Press.
Binmore, K. 1998. *Just playing*. Cambridge, MA: MIT Press.
Binmore, K. 2005. *Natural justice*. Oxford: Oxford University Press.
Blackorby, C., W. Bossert, and D. Donaldson. 2005. *Population issues in social choice theory, welfare economics, and ethics*. Cambridge: Cambridge University Press.
Brandt, R. 1992. *Morality, utilitarianism, and rights*. Cambridge: Cambridge University Press.
Creedy, J. 1986. *Edgeworth and the development of neoclassical economics*. Oxford: Blackwell.
Cumberland, R. 1672. In *De Legibus Naturae: A treatise of the laws of the nature*, ed. J. Parkin, 2005. Indianapolis: Liberty Fund.
Deschamps, R., and L. Gevers. 1978. Leximin and utilitarian rules: a joint characterization. *Journal of Economic Theory* 17: 143–163.
Diamond, P.A. 1967. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility: Comment. *Journal of Political Economy* 75: 765–766.
Diamond, P.A., and J.A. Mirrlees. 1974. Optimal taxation and public production, I: production efficiency, and II. tax rules. *American Economic Review* 61(8–27): 261–278.
Dworkin, R. 1977. *Taking rights seriously*. London: Duckworth.
Edgeworth, F.Y. 1877. *New and old methods of ethics, or 'Physical Ethics' and 'Methods of Ethics'*. Oxford and London: James Parker. Repr. in Newman (2003).
Edgeworth, F.Y. 1881. *Mathematical psychics: An essay on the application of mathematics to the moral sciences*. London: Kegan Paul. Repr. in Newman (2003).
Edgeworth, F.Y. 1897. The pure theory of taxation. *Economic Journal* 7: 46–70, 226–38, 550–71.
Edwards, R.B. 1979. *Pleasures and pains: A theory of qualitative hedonism*. Ithaca and London: Cornell University Press.
Feiwel, G.R., eds. 1987. *Arrow and the foundations of the theory of economic policy*. London: Macmillan.
Feldman, F. 2004. *Pleasure and the good life: Concerning the nature, varieties, and plausibility of hedonism*. Oxford: Clarendon Press.
Gibbard, A.F. 1987. Ordinal utilitarianism. In *Feiwel*.
Gibbard, A.F. 1990. *Wise choices, Apt feelings*. Cambridge, MA: Harvard University Press.
Hammond, P.J. 1976. Equity, Arrow's conditions and Rawls' difference principle. *Econometrica* 44: 793–804.
Hammond, P.J. 1977. Dual interpersonal comparisons of utility and the welfare economics of income distribution. *Journal of Public Economics* 7: 51–71.
Hammond, P.J. 1987. On reconciling Arrow's theory of social choice with Harsanyi's fundamental utilitarianism. In *Feiwel*.
Hare, R.M. 1981. *Moral thinking: Its method, levels and point*. Oxford: Clarendon Press.
Harsanyi, J.C. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy* 63: 309–321.
Harsanyi, J.C. 1977. *Rational behavior and bargaining equilibrium in games and social situations*. Cambridge: Cambridge University Press.
Harsanyi, J.C. 1992. Game and decision theoretic models in ethics. In *Handbook of game theory*, ed. R.J. Aumann and S. Hart, Vol. 1. Amsterdam: Elsevier.
Hume, D. 1751. In *An enquiry concerning the principles of morals*, ed. T.L. Beauchamp, 1998. Oxford: Oxford University Press.
Hutcheson, F. 1755. In *A system of moral philosophy*, ed. D. Carey, 2000. London/New York: Continuum.
Jevons, W.S. 1874. Utilitarianism. In *Pure logic and other minor works of W.S. Jevons*, ed. R. Adamson and H.A. Jevons, 1890. London: Macmillan.
Kahneman, D., E. Diener, and N. Schwarz. 1999. *Well-Being: The foundations of hedonic psychology*. New York: Russell Sage Foundation.
Kelly, P.J. 1990. *Utilitarianism and distributive justice*. Oxford: Clarendon Press.
Lively, J., and J. Rees. 1978. *Utilitarian logic and politics*. Oxford: Clarendon Press.
Marshall, A. 1884. On utilitarianism: A summum bonum. In *The early economic writings of Alfred Marshall, 1867–1890*, ed. J.K. Whitaker, Vol. 2, 1975. New York: Free Press.
Maskin, E.S. 1978. A theorem on utilitarianism. *Review of Economic Studies* 45: 93–96.
Mill, J. 1820. An essay on government. In *James Mill: Political writings*, ed. T. Ball, 1992. Cambridge: Cambridge University Press.
Mill, J.S. 1859. On liberty. In *Collected works*, ed. J.M. Robson, Vol. 18, 1977. Toronto/London: University of Toronto Press/Routledge.
Mill, J.S. 1861a. Utilitarianism. In *Collected works*, ed. J.M. Robson, Vol. 10, 1969. Toronto/London: University of Toronto Press/Routledge.
Mill, J.S. 1861b. Considerations on representative government. In *Collected works*, ed. J.M. Robson, Vol. 19, 1977. Toronto/London: University of Toronto Press/Routledge.
Mirrlees, J.A. 1971. An exploration in the theory of optimal income taxation. *Review of Economic Studies* 38: 175–208.
Mirrlees, J.A. 1982. The economic uses of utilitarianism. In *Sen and Williams*.
Mitchell, W.C. 1918. Bentham's felicific calculus. *Political Science Quarterly* 33: 161–183.
Mongin, P. 2006. A concept of progress for normative economics. *Economics and Philosophy* 22: 19–54.

Mongin, P., and C. d'Aspremont. 1998. Utility theory and ethics. In *Handbook of utility theory*, ed. S. Barbera, P. Hammond, and C. Seidl, Vol. 1. Dordrecht: Kluwer.
Moore, G.E. 1903. *Principia ethica*, 1959. Cambridge: Cambridge University Press.
Moore, G.E. 1912. *Ethics*, 1958. Oxford: Oxford University Press.
Newman, P. 2003. *F.Y. Edgeworth: Mathematical psychics and further papers on political economy.* Oxford: Oxford University Press.
Ng, Y.-K. 1975. Bentham or Bergson? Finite sensibility, utility functions, and social welfare functions. *Review of Economic Studies* 42: 545–570.
Ng, Y.-K. 1985. Some fundamental issues in social welfare. In *Issues in contemporary microeconomics and welfare*, ed. G.R. Feiwell. London: Macmillan.
Ng, Y.-K., and P. Singer. 1981. An argument for utilitarianism. *Canadian Journal of Philosophy* 11: 229–239.
Pigou, A.C. 1920. *The economics of welfare*, 4th rev. edn. London: Macmillan, 1932.
Rawls, J. 1971. *A theory of justice*. Cambridge, MA: Harvard University Press.
Rawls, J. 1993. *Political liberalism*. 2nd ed, 2002. New York: Columbia University Press.
Riley, J. 1988. *Liberal utilitarianism*. Cambridge: Cambridge University Press.
Riley, J. 2001. Welfare (philosophical aspects). In *International encyclopedia of the social and behavioral sciences*, ed. N.J. Smelser and P.B. Bates, Vol. 25. London: Pergamon.
Riley, J. 2006a. Liberal rights in a Pareto-optimal code. *Utilitas* 18: 61–79.
Riley, J. 2006b. Genes, memes and justice. *Analyse & Kritik* 28: 32–56.
Riley, J. 2007a. Mill's neo-Athenian model of liberal democracy. In *J.S. Mill's political thought: A bicentennial reassessment*, ed. N. Urbinati and A. Zakaras. New York: Cambridge University Press.
Riley, J. 2007b. *Mill's radical liberalism*. London: Routledge.
Riley, J. 2007c. Classical ordinalist utilitarianism. Unpublished working paper.
Riley, J. 2009. The interpretation of maximizing utilitarianism, and distributive justice. *Social Philosophy and Policy* 26(Winter).
Rosen, F. 1983. *Jeremy Bentham and representative democracy: A study of the constitutional code*. Oxford: Clarendon Press.
Scheffler, S. 1982. *The Rejection of consequentialism*, Rev. edn. Oxford: Oxford University Press. 1994.
Sen, A.K. 1970. *Collective choice and social welfare*. San Francisco: Holden-Day.
Sen, A.K. 1979. Utilitarianism and welfarism. *Journal of Philosophy* 76: 463–489.
Sen, A.K. 1980. Plural utility. *Proceedings of the Aristotelian Society* 81: 193–215.
Sen, A.K. 1982. *Choice, welfare and measurement*, 1997. Oxford: Blackwell. Repr. Harvard University Press.
Sen, A.K. 1985. *Commodities and capabilities*, 1999. Amsterdam: North-Holland. Repr. Oxford University Press.
Sen, A.K. 2002. *Rationality and freedom*. Cambridge, MA: Harvard University Press.
Sen, A.K. 2007. *Freedom and justice*. Cambridge, MA: Harvard University Press.
Sen, A.K., and B.A.O. Williams, eds. 1982. *Utilitarianism and beyond*. Cambridge: Cambridge University Press.
Sidgwick, H. 1874. *The methods of ethics*. 7th ed, 1907. London: Macmillan.
Sidgwick, H. 1877. Bentham and Benthamism in politics and ethics. In *Essays on ethics and method: Henry Sidgwick*, ed. M.G. Singer, 2000. Oxford: Clarendon Press.
Sidgwick, H. 1886. *Outlines of the history of ethics*. 5th ed, 1902. London: Macmillan.
Smart, J.J.C., and B.A.O. Williams. 1977. *Utilitarianism: For and against*. Cambridge: Cambridge University Press.
Stiglitz, J.E. 1987. Pareto efficient and optimal taxation and the new welfare economics. In *Handbook of public economics*, ed. A.J. Auerbach and M. Feldstein, Vol. 2. Amsterdam: North-Holland.

Utility

R. D. Collison Black

Keywords

Cardinal utility; Classical theory of value; Consumer surplus; Cost of production; Demand; Diminishing marginal utility; Dupuit, A.-J.-E.J.; Edgeworth, F. Y.; Exchange; Exchange value; Gossen, H. H.; Hicks, J. R.; Interpersonal utility comparisons; Jevons, W. S.; Marginal revolution; Marginal utility; Market price; Marshall, A.; Menger, C.; Natural price; Ordinal utility; Pareto, V.; Pigou, A. C.; Price theory; Revealed preference theory; Use value; Utility; Value; Walras, L.

JEL Classifications

A1

Utility is a term which has a long history in connection with the attempts of philosophers and

political economists to explain the phenomenon of value. It has most frequently been given the connotation of 'desiredness', or the capacity of a good or service to satisfy a want, of whatever kind. Its use with that meaning can be traced back at least to Gershom Carmichael's 1724 edition of Pufendorf's *De Officio Hominis et Civis Iuxta Legam Naturalem*, and arguably came down to him through the medieval schoolmen from Aristotle's *Politics*.

Utility in the sense of desiredness is a purely subjective concept, clearly distinct from usefulness or fitness for a purpose – the more normal everyday sense of the word and the first meaning given for it by the *Oxford English Dictionary*.

While most political economists of the 18th and 19th centuries used the term in this subjective sense, the distinction was not always kept clear, most notably in the writings of Adam Smith. In a famous passage in the *Wealth of Nations* Smith wrote:

> The word VALUE, it is to be observed, has two different meanings, and sometimes expresses the utility of some particular object, and sometimes the power of purchasing other goods which the possession of that object conveys. The one may be called 'value in use'; the other, 'value in exchange'. The things which have the greatest value in use have frequently little or no value in exchange; and, on the contrary, those which have the greatest value in exchange have frequently little or no value in use. Nothing is more useful than water; but it will purchase scarce anything; scarce anything can be had in exchange for it. A diamond, on the contrary, has scarce any value in use; but a very great quantity of other goods may frequently be had in exchange for it. (1776, Book I, ch. IV)

Smith has sometimes been accused, because of the wording of this passage, of falling into the error of claiming that things which have *no* value in use can have value in exchange, which is tantamount to saying that utility is not a necessary condition for a good to have value. It would appear, however, that Smith was not here using the theme 'value in use', or utility, in the subjective sense of desiredness but in the normal objective sense of usefulness (cf. Bowley 1973, p. 137; O'Brien 1975, pp. 80 and 97). Most other classical economists and even Smith himself in his *Lectures on Jurisprudence* used the term in its subjective sense, but the passage in the *Wealth of Nations* gave rise to considerable confusion and misinterpretation. Nor was this the only source of confusion in the early writing on the subject: even those who used the term utility in its subjective sense were not always clear as to whether it should be considered a feeling in the mind of the user or a property of the good or service used. Thomas De Quincey, for example, referred to the 'intrinsic utility' of commodities (*Logic of Political Economy*, 1844, p. 14).

Most classical economists, however, were not greatly concerned with the subtleties of meaning which the term utility might contain. Generally they used it in the broad sense of desiredness, and Ricardo employed it in a typically classical way when he wrote

> Utility then is not the measure of exchangeable value, although it is absolutely essential to it. If a commodity were in no way useful – in other words, if it could in no way contribute to our gratification – it would be destitute of exchangeable value, however scarce it might be, or whatever quantity of labour might be necessary to procure it. (*Principles of Political Economy and Taxation*, 1817, ch. 1, sect. I)

'Useful' here is interpreted as 'contributing to gratification' but the very word carries an echo of Smith's confusion.

For Ricardo and others in the mainstream classical tradition down to J.S. Mill and Cairnes utility became a necessary but not a sufficient condition for a good to possess value. In this context, the utility referred to was generally the total utility of the good to the purchaser, or the utility of a specific quantity which is all that is available in the circumstances of the example – for example, the utility of a single item of food to a starving person.

As a result of this approach it followed, in the words of J.S. Mill, that 'the utility of a thing in the estimation of the purchaser is the extreme limit of its exchange value: higher the value cannot ascend; peculiar circumstances are required to raise it so high' (*Principles of Political Economy*, 1848, Book III, ch. II, §. 1). Classical economists like Mill accepted the view put forward by J.B. Say that 'labour is not creative of objects, but of utilities' but could see the weakness in

Say's contention that price measured utility. Clearly in the case of competitively produced commodities it did not and it was cost of production and not utility (in the total sense) which determined value.

Since the classical economists were mainly interested in 'natural' rather than 'market' price, that is, in long-run normal values which were mainly determined by supply and cost, the fact that they had no theory to explain fully the relationships between utility, demand and market price was not a matter of concern to most of them. Nevertheless in the period from about 1830 to 1870 a number of attempts were made to work out these relationships more fully or to clarify aspects of them. Some of these attempts took place in Britain, within the framework of the classical system, but not surprisingly some of the best work was done at this time in France, where the tradition of demand analysis was stronger.

The full explanation of the relation between utility and demand requires the distinction between total utility and increments of utility, and the recognition of the principle that consumption of successive increments of a commodity yields not equal but diminishing increments of satisfaction or utility to the consumer. A number of writers in the mid-19th century showed an understanding of this point, but only a few stated it explicitly and correctly. Among those in Britain who did so were William Foster Lloyd (*A Lecture on the Notion of Value, delivered before the University of Oxford in Michaelmas Term*, 1833) and Nassau Senior (*Outline of the Science of Political Economy*, 1836), but neither proceeded to develop his insights into a complete theory of the relationship between utility, demand and market values.

The French engineer A.J. Dupuit was the first to present an analysis which clearly explained the concept of marginal utility and related it to a demand curve, in his paper 'On the Measurement of the Utility of Public Works' (*Annales des Ponts et Chaussées*, vol. 8, 1844; English translation in *International Economic Papers*, No. 2, 1952, pp. 83–110). Dupuit also extended his analysis to show that the total area under the demand curve represents the total utility derived from the commodity; deducting from this the total receipts of the producer he arrived at the 'utility remaining to consumers' or what was later to be termed 'consumers' surplus'.

The significance of Dupuit's contribution is now well recognized, but at the time of appearance it had little impact. The same is even more true of the work of Hermann Heinrich Gossen, one of the few German contributors to utility theory in this period. His book *Entwicklung der Gesetze des Menschlichen Verkehrs*, published in 1854, contained not only a statement of the 'law of satiable wants', or diminishing marginal utility, but also of the proposition that to maximize satisfaction from any good capable of satisfying various wants it must be allocated between those uses so as to equalize its marginal utilities in all of them.

Gossen's analysis of the principles of utility maximization was thus more complete than any which had preceded it. Yet his one book, which foreshadowed many features of general equilibrium as well as utility theory, received virtually no attention until 1878, 20 years after the author's death, when Robert Adamson, W.S. Jevons's successor as Professor of Philosophy and Political Economy at Manchester, obtained a copy of it and drew it to the attention of Jevons himself.

By that time the whole character of utility analysis and its place in economic theory had begun to change significantly. This change is usually dated from the very nearly simultaneous publication of Jevons's *Theory of Political Economy* in England and Menger's *Grundsätze der Volkswirtschaftslehre* Austria, both in 1871, and Walras's *Eléments d'économie politique pure* in Switzerland in 1874. All these work contained a treatment of the theory of value in which the analysis of diminishing marginal utility (under a variety of other names) played a considerable part, but each of the three authors seems to have arrived independently at the main ideas of his theory without indebtedness to the others or to the predecessors already mentioned above.

This remarkable example of multiple discovery in the history of ideas has come to be known as 'the Marginal Revolution'. Discussion of its causes and character lies outside the scope of this article, but it is generally accepted that, as

Sir John Hicks has said, 'the essential novelty in the work of these economists was that instead of basing their economics on production and distribution, they based it on exchange' (Hicks 1976, p. 212).

A major element in this 'shift in attention' undoubtedly was a change from the classical concept of value in use, or total utility, as a necessary but not sufficient condition to explain the normal values of freely reproducible commodities, to the concept of what Jevons called 'the degree of utility' and of adjustments in it, through exchange of quantities of goods held or consumed, in order to maximize satisfaction. Marginal analysis can, however, be applied to questions of production and distribution as well as consumption, and hence the 'Marginal Revolution' involved more than a new stage in the development of utility theory.

Although all three pioneers of the Marginal Revolution did contribute to that development they also contributed in other ways to the theory of pricing and exchange. Perhaps only for Jevons was the theory of utility genuinely central to the structure of his economic work. On the opening page of his *Theory of Political Economy* he emphatically asserted that '*value depends entirely upon utility'* and he went on to say that 'Political Economy must be founded upon a full and accurate investigation of the conditions of utility' (1871 p. 46). Jevons indeed appears to have shared with his classical predecessors the view that a theory of value must go beyond the phenomena of demand and supply to some more fundamental explanation which for him was to be found in utility rather than in labour. 'Labour is found often to determine value, but only in an indirect manner, by varying the degree of utility of the commodity through an increase or limitation of supply' (Jevons 1871, p. 2).

Apart from differences of terminology, Walras's treatment of utility in relation to the problem of exchange had substantial similarities with that of Jevons; but Walras saw the problem in a different context.

> His whole attention was focused on market phenomena and not on consumption . . . while the driving force in the theory of exchange is, as Walras saw it, the endeavour of all traders to maximise their several satisfactions, it is marketplace satisfactions rather than dining-room satisfactions which Walras had in mind. (Jaffé 1973, pp. 118–19)

For Menger, as for Walras, the concepts of utility theory formed only a part of a much large analytical structure (concerned in his case not so much with equilibrium as with development), but unlike both Walras and Jevons he refused to state his theories in mathematical terms.

Menger developed a theory of economizing behaviour showing how the individual would seek to satisfy his subjectively felt needs in the most efficient manner. In the process he elaborated the essential propositions of a theory of maximizing behaviour for the consumer, but he expressed them in terms of the satisfaction of needs by the consumption of successive units of goods. In his discussion of this process Menger used the same phrases – use-value and exchange-value – which Smith had used almost a century earlier, and with similar connotations. Use-value he defined as 'the importance that goods acquire for us because they *directly* assure us the satisfaction of needs that would not be provided for if we did not have the goods at our command. Exchange value is the importance that goods acquire for us because their possession assures the same result *indirectly*' (Menger 1871, p. 228). Menger did use the term 'utility', but not as a synonym for use-value; he viewed it as an abstract relation between a species of goods and a human need, akin to the general term 'usefulness'. As such it constituted a prerequisite for a good to have economic character, but had no quantitative relationship to value.

The three pioneers of the Marginal Revolution thus saw the problem of the relationship of utility to exchange value in different contexts and expressed their solutions to it in different ways. Inevitably also their first solutions were incomplete in various respects. For example, the precise relationships between the individual's utility function and demand function, the market demand function and the market price were not clearly specified in some of the earlier formulations; it remained for later contributors such as Marshall, Wicksteed and Edgeworth to deal with these points.

Nevertheless, despite their differences of terminology and approach, the writings of the pioneers did contain a common core which gradually gained wider acceptance, and by 1890 with the appearance of Marshall's *Principles of Economics* it seemed that the new analysis of market values had been effectively integrated with an analysis of supply and cost which served to explain long-run 'normal' values. It did provide some things which the classical system had not contained, among them a consistent theory of consumer behaviour, expressed in terms of utility.

So in 1899 it was possible for Edgeworth to write that:

> the relation of utility to value, which exercised the older economists, is thus simply explained by the mathematical school. The value in use of a certain quantity of commodity corresponds to its total utility; the value in exchange to its marginal utility (multiplied by the quantity). The former is greater than the latter, since the utility of the final increment of commodity is less than that of every other increment. (Edgeworth 1899, p. 602)

At this stage utility analysis appeared to have evolved to something approaching finality, and in 1925 Jacob Viner could still say:

> In its developed form it is to be found sympathetically treated and playing a prominent role in the exposition of value theory in most of the current authoritative treatises on economic theory by American, English, Austrian, and Italian writers.

Yet Viner immediately went on to add:

> In the scientific periodicals, however, in contrast with the standard treatises, sympathetic expositions of the utility theory of value have become somewhat rare. In their stead are found an unintermittent series of slashing criticisms of the utility economics. (Viner 1925, p. 179)

The principal criticisms which Viner noted were the apparent involvement of utility theory with hedonistic psychology and the problems of measuring welfare in terms of utility. In later years questions of the measurement and summation of utility came to trouble economists more and more.

The two basic problems involved here are whether utility can be measured cardinally or simply ordinally, and whether interpersonal comparisons of utility are possible. The pioneers of the Marginal Revolution were not unaware of these problems; Jevons nowhere attempted to define a unit of utility, and indeed said that 'a unit of pleasure or pain is difficult even to conceive', but he went on to say that 'it is from the quantitative effects of the feelings that we must estimate their comparative amounts' (Jevons 1871, p. 14). However they may be estimated, Jevons did not hesitate to refer to 'quantity of utility', and his whole analysis proceeds by treating utility as if it could be measured. The question was not examined in detail by Walras or Menger, but both their analyses treat utility as cardinally measurable.

On the question of interpersonal comparisons of utility, Menger and Walras seemed to find no difficulty, and Walras was prepared to speak of a 'maximum of utility' for society (Walras 1874, p. 256). Jevons on the other hand declared that 'every mind is . . . inscrutable to every other mind, and no common denominator of feeling seems possible' (Jevons 1871, p. 211) – but this did not always prevent him from comparing and aggregating utilities.

In the early editions of his *Principles of Economics* Marshall fully accepted the idea of utility as cardinally measurable and allowed the possibility if not of interpersonal certainly of intergroup comparisons of utility (1890, pp. 151 and 152). In later years he became more reticent and defensive on these points, and he was always more concerned than Jevons with the effects of feelings rather than the feelings themselves; yet cardinal utility always remained the basis of Marshall's demand theory.

Now, as Sir John Hicks said:

> if one starts from a theory of demand like that of Marshall and his contemporaries, it is exceedingly natural to look for a welfare application. If the general aim of the economic system is the satisfaction of consumer wants, and if the satisfaction of individual wants is to be conceived of as a maximising of utility, cannot the aim of the system be itself conceived of as a maximising of utility – universal utility, as Edgeworth called it? If this could be done and some measure of universal utility could be found, the economist's function could be widened out, from the understanding of cause and effect to the judgement of the effects – whether, from the point of view of want- satisfaction, they are to be judged as successful or unsuccessful, good or bad. (Hicks 1956, p. 6)

This was, in effect, the task which was undertaken by Marshall's successor, A.C. Pigou, in his *Economics of Welfare* (1920; earlier version published under the *title Wealth and Welfare*, 1912). Pigou made no attempt to establish a measure of universal utility; instead he took what Marshall had called 'the national dividend', aggregate real income, as the 'objective counterpart' of economic welfare. Pigou argued that economic welfare would be greater when aggregate real income increased, when fluctuations in its amount were reduced, and when it was more equally distributed among persons. It was in the context of this last point that interpersonal utility comparisons were most evident; Pigou argued that

> the old 'law of diminishing utility' thus leads securely to the proposition: Any cause which increases the absolute share of real income in the hands of the poor, provided that it does not lead to a contraction in the size of the national dividend . . . will in general, increase economic welfare. (1920, p. 89)

In the 1930s most economists became increasingly uncomfortable with the idea of measurement and interpersonal or intergroup comparisons of utility. In 1934, in a famous article entitled 'A Reconsideration of the Theory of Value', Hicks and Allen used the technique of indifference curves originated by Edgeworth and developed by Walras's successor at Lausanne, Vilfredo Pareto, in presenting a theory of consumer behaviour involving only ordinal comparisons of satisfaction. A few years later a further step towards eliminating what were now considered dubious psychological assumptions from that theory was taken by treating consumer behaviour solely on the basis of revealed preference.

Accompanying these changes there was a movement away from the type of welfare economics developed on the basis of utility theory by Marshall and Pigou towards that based on Pareto's concept of an economic optimum as a position from which it is impossible to improve anyone's welfare without damaging that of another.

Indifference analysis and revealed preference theory are now standard features of microeconomic theory; but the utility concept has not disappeared; the most widely used introductory economics texts still tend to begin their treatments of household behaviour with an account of utility theory.

See Also

▶ Gossen, Hermann Heinrich (1810–1858)

Bibliography

Bowley, M. 1973. Utility, the paradox of value and 'all that' and classical economics. In *Studies in the history of economic theory before 1870*, ed. M. Bowley. London: Macmillan.

De Quincey, T. 1844. *The logic of political economy.* Edinburgh: Blackwood.

Dupuit, A.J. 1844. On the measurement of the utility of public works. *Annales des Ponts et Chaussées*, 2nd Series, vol. 8. Trans. in *International economic papers*, No. 2, London: Macmillan, 1952.

Edgeworth, F.Y. 1899. Utility. In *Dictionary of political economy*, ed. R.H.I. Palgrave, vol. 3. London: Macmillan.

Gossen, H.H. 1854. *Entwicklung der Gesetze des menschlichen Verkehrs und der daraus Fliessenden Regeln für menschliches Handeln*. Brunswick: Viewig. Translated as *The laws of human relations and the rules of human action derived therefrom.* Cambridge, MA: MIT Press, 1983.

Hicks, J.R. 1956. *A revision of demand theory.* Oxford: Clarendon Press.

Hicks, J.R. 1976. 'Revolutions' in economics. In *Method and appraisal in economics*, ed. S.J. Latsis. Cambridge: Cambridge University Press.

Hicks, J.R., and R.G.D. Allen. 1934. A reconsideration of the theory of value. *Economica* 1 (52–76): 196–219.

Howey, R.S. 1960. *The rise of the marginal utility school, 1870–1889*. Lawrence: University of Kansas Press.

Jaffé, W. 1973. Léon Walras's role in the 'marginal revolution' of the 1870s. In *The marginal revolution in economics*, ed. R.D. Collison Black, A.W. Coats, and C.D. Goodwin. Durham: Duke University Press.

Jevons, W.S. 1871. *The theory of political economy.* London: Macmillan.

Kauder, E. 1965. *A history of marginal utility theory.* Princeton: Princeton University Press.

Lloyd, W.F. 1833. A lecture on the notion of value, delivered before the University of Oxford in Michaelmas Term 1833. Reprinted in *Economic history, supplement to the Economic Journal*, No. 2 (1927), 168–183.

Marshall, A. 1890. *Principles of economics*, 9th (variorum) ed, ed. C.W. Guillebaud. London: Macmillan, 1961.

Menger, C. 1871. *Grundsätze der Volkswirtschaftslehre*. Trans. J. Dingwall and B.F. Hoselitz as *Principles of economics.* Glencoe: Free Press, 1951.
Mill, J.S. 1848. *Principles of political economy,* ed. W. J. Ashley. London: Longmans, 1909.
O'Brien, D.P. 1975. *The classical economists*. Oxford: Clarendon Press.
Pigou, A.C. 1912. Wealth and welfare. Expanded and republished as *The economics of welfare*. London: Macmillan, 1920.
Ricardo, D. 1817. *The principles of political economy and taxation*, vol. 1 of *The works and correspondence of David Ricardo*, ed. P. Sraffa. Cambridge: Cambridge University Press, 1951.
Senior, N.W. 1836. *An outline of the science of political economy.* London: W. Clowes. Reprinted, London: Allen & Unwin, 1938.
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.
Stigler, G.J. 1950. The development of utility theory. Pts. I and II. *Journal of Political Economy* 58: 307–327; 373–396. Reprinted in G.J. Stigler, *Essays in the history of economics.* Chicago: University of Chicago Press, 1965.
Viner, J. 1925. The utility concept in value theory and its critics. *Journal of Political Economy* 33: 369–387. Reprinted in J. Viner, *The long view and the short*. Glencoe: Free Press, 1958.
Walras, L. 1874. *Eléments d'économie politique pure*. Lausanne: Corbaz et Cie. Trans. W. Jaffé as *Elements of pure economics*. London: Allen & Unwin, 1954.

Utility Theory and Decision Theory

Peter C. Fishburn

The conjunction of utility theory and decision theory involves formulations of decision making in which the criteria for choice among competing alternatives are based on numerical representations of the decision agent's preferences and values. Utility theory as such refers to these representations and to assumptions about preferences that correspond to various numerical representations. Although it is a child of decision theory, utility theory has emerged as a subject in its own right as seen, for example, in the contemporary review by Fishburn (*See* ► Representation of Preferences). Readers interested in more detail on representations of preferences should consult that entry.

Our discussion of utility theory and decision theory will follow the useful three-part classification popularized by Luce and Raiffa (1957), namely decision making under certainty, risk, and uncertainty. I give slightly different descriptions than theirs.

Certainty refers to formulations that exclude explicit consideration of chance or uncertainty, including situations in which the outcome of each decision is known beforehand. Most of consumer demand theory falls within this category.

Risk refers to formulations that involve chance in the form of known probabilities or odds, but excludes unquantified uncertainty. Games of chance and insurance decisions with known probabilities for possible outcomes fall within the risk category. Note that 'risk' as used here is only tangentially associated with the common notion that equates risk with the possibility of something bad happening.

Uncertainty refers to formulations in which decision outcomes depend explicitly on events that are not controlled by the decision agent and whose resolutions are known to the agent only after the decision is made. Probabilities of the events are regarded either as meaningless, unknowable, or assessable only with reference to personal judgement. Situations addressed by the theory of noncooperative games and statistical decision theory typically fall under this heading.

A brief history of the subject will provide perspective for our ensuing discussion of the three categories.

Historical Remarks

The first important paper on the subject was written by Daniel Bernoulli (1738) who, in conjunction with Gabriel Cramer, sought to explain why prudent agents often choose among risky options in a manner contrary to expected profit maximization. One example is the choice of a sure $10,000 profit over a risky venture that loses $5000 or gains $30,000, each with probability 1/2. Bernoulli argued that many such choices

could be explained by maximization of the expected utility ('moral worth') of risky options, wherein the utility of wealth increases at a decreasing rate. He thus introduced the idea of decreasing marginal utility of wealth as well as the maximization of expected utility.

Although Bernoulli's ideas were endorsed by Laplace and others, they had little effect on the economics of decision making under risk until quite recently. On the other hand, his notion of decreasing marginal utility became central in consumer economics during the latter part of the 19th century (Stigler 1950), especially in the works of Gossen, Jevons, Menger, Walras and Marshall.

During this early period, utility was often viewed as a measurable psychological magnitude. This notion of intensive measurable utility, which was sometimes represented by the additive form $u_1(x_1) + u_2(x_2) + \cdots + u_n(x_n)$ for commodity bundles $(x_1,x_2,\ldots,x_n)$, was subsequently replaced in the ordinalist revolution of Edgeworth, Fisher, Pareto, and Slutsky by the view that utility represents nothing more than the agent's preference ordering over consumption bundles. A revival of intensive measurable utility occurred after 1920 when Frisch, Lange and Alt axiomatized the notion of comparable preference differences, but it did not regain the prominence it once held.

Bernoulli's long-dormant principle of the maximization of expected utility reappeared with force in the expected utility theory of von Neumann and Morgenstern (1944, 1947). Unlike Bernoulli and Cramer, who favoured an intensive measurable view of utility, von Neumann and Morgenstern showed how the expected-utility form can arise solely from simple preference comparisons between risky options. They thus accomplished for decision making under risk what the ordinalists accomplished for demand theory a generation earlier.

Although little noted at the time, Ramsey (1931), in an essay written in 1926 and published posthumously, attempted something more ambitious than the utility theory for risky decisions of von Neumann and Morgenstern. Ramsey's aim was to show how assumptions about preferences between uncertain decisions imply not only a utility function for outcomes but also a subjective or personal probability distribution over uncertain events such that one uncertain decision is preferred to another precisely when the former has greater subjective (probability) expected utility. Ramsey's outline of a theory of decision making under uncertainty greatly influenced the first complete theory of subjective expected utility, due to Savage (1954). Savage also drew heavily on Bruno de Finetti's seminal ideas on subjective probability, which are similar in ways to views espoused much earlier by Bayes and Laplace.

During the historical period, several unsuccessful proposals were made to replace 'utility' by a term better suited to the concepts it represents. Despite these failures, the terms *ordinal utility* and *cardinal utility*, introduced by Hicks and Allen (1934) to distinguish between the ordinalist viewpoint and the older measurability view of utility as a 'cardinal magnitude', caught on. Present usage adheres to the following measurement theoretic definitions.

Let $\succ$ denote the relation *is preferred to* on a set X of decision alternatives, outcomes, commodity bundles, or whatever. Suppose preferences are ordered and can be represented by a real valued function u on X as

$$x \succ y \Leftrightarrow u(x) > u(y), \tag{1}$$

for all x and y in X. We then say that u is an *ordinal utility function* if it satisfies (1) but is subject to no further restrictions. Then any other real function v that preserves the order of $\succ$, or satisfies (1) in place of u, is also an ordinal utility function, and all such functions for the given $\succ$ are equivalent in the ordinal context. A different preference ordering on X will have a different equivalence class of order-preserving functions. If u is also required to be continuous, we may speak of continuous ordinal utility.

If u satisfies (1) and is restricted by subsidiary conditions in such a way that v also satisfies (1) *and* the subsidiary conditions if and only if there are numbers $a > 0$ and b such that

$$v(x) = au(x) + b, \quad \text{for all } x \text{ in } X, \tag{2}$$

then u is a *cardinal utility function* and is said to be unique up to a positive ($a > 0$) liner transformation. Subsidiary conditions that force (2) under appropriate structure for X include additivity $u(x_1,\ldots,x_n) = u_1(x_1) + \ldots + u_n \quad (x_n)$ with $n \geq 2$, the linearity property $u[\lambda p + (1 - \lambda)q] = \lambda u(p) + (1 - \lambda)u(q)$ of expected utility, and the ordered preference-difference representation $(x, y) \succ^{*} (z, w) \Leftrightarrow u(x) - u(y) > u(z) - u(w)$. Only the last of these, where $(x, y) \succ^{*} (z, w)$ says that the intensity of preference for x over y exceeds the intensity of preference for z over w, involves a view of preference that goes beyond the basic relation $\succ$.

Decisions Under Certainty

Representation (1) is the preeminent utility representation for decision making under certainty. It presumes that the preference relation $\succ$ is

$$\textit{asymmetric:} \text{if } x \succ y \text{ then not } (y \succ x),$$
$$\textit{negatively transitive:} \text{ if } x \succ z \text{ then } x \succ y \text{ or } y \succ z,$$

and, when X is uncountably infinite, that there is a countable subset C_0 in X such that, whenever $x \succ y$, there is a z in C_0 such that $x \succsim z \succsim y$, where $x \succsim z$ means not $(z \succ x)$. An asymmetric and negatively transitive relation is often referred to as a *weak order*, and in this case both $\succ$ and its induced indifference relation $\sim$, defined by

$$x \sim y \text{ if neither } x \succ y \text{ nor } y \succ x,$$

are *transitive*, that is $(x \succ y, y \succ z) \Rightarrow x \succ z$ and $(x \sim y, y \sim z) \Rightarrow x \sim z$. If X is a connected and separable topological space, the countable C_0 condition can be replaced by the assumption that the preferred-to-x set $\{y: y \succ x\}$ and the less-preferred-than-x set $\{y: x \succ y\}$ are open sets in X's topology for every x in X. When this holds, u can be taken to be continuous. If the countable C_0 condition fails when $\succ$ is a weak order, (1) cannot hold and instead we could represent $\succ$ by vectorvalued utilities ordered lexicographically. For details and further references, see Fishburn (1970, 1974).

Economics is often concerned with situations in which any one of a number of subsets of X might arise as the feasible set from which a choice is required. For example, if X is a commodity space $\{(x_1,\ldots,x_n): x_i \geq 0 \text{ for } i = 1,\ldots,n\}$, then the feasible set at price vector $p = (p_1,\ldots, p_n) > (0,\ldots,0)$ and disposable income $m \geq 0$ is the *opportunity set* $\{(x_1,\ldots,x_n): p_1x_1 + \ldots + p_nx_n \leq m\}$ of commodity bundles that can be purchased at p and m. The allure of (1) in such situations is that the same u can be used for choice by maximization of utility for each non-empty feasible set Y so long as the set

$$\max_u Y = \{x \text{ in } Y : u(x) \geq u(y) \text{ for all } y \text{ in } Y\}$$

is not empty. The existence of non-empty $\max_u Y$ is assured if Y is finite or if it is a compact subset of a connected and separable topological space on which u is upper semicontinuous.

When (1) holds, the set

$$\max \succ Y = \{x \text{ in } Y : y \succ x \text{ for no } y \text{ in } Y\}$$

of maximally-preferred elements in Y is identical to $\max_u Y$. On the other hand, $\max \succ Y$ can be non-empty when no utility function satisfies (1). For example, if X is finite, then $\max \succ Y$ is non-empty for every non-empty subset Y of X if, and only if, X contains no preference cycle, that is no $x_1,\ldots,x_m$ such that $x_1 \succ x_2 \succ \ldots \succ x_m \succ x_1$. In this case it is always possible to define u on X so that, for all x and y in X,

$$x \succ y \Rightarrow u(x) \succ u(y). \tag{3}$$

Then choices can still be made by maximization of utility since $\max_u Y$ will be a non-empty subset of $\max \succ Y$. However, if $\succ$ has cycles, then the principle of choice by maximization breaks down.

The situation for infinite X and suitably constrained feasible sets is somewhat different. Sonnenschein (1971) shows for the commodity space setting that $\succ$ can have cycles while every opportunity set Y has a non-empty $\max \succ Y$. His key assumptions are a semicontinuity condition

on $\gtrsim$ and the assumption that every preferred-to-x set is convex. Thus, choice by maximal preference may obtain when $\succ$ can be characterized by neither (1) nor (3).

Max $\succ Y$ for opportunity sets Y in commodity space is the agent's *demand correspondence* (which depends on p and m) or, if each max $\succ Y$ is a singleton, his *demand function*. The revealed preference approach of Samuelson and others represents an attempt to base the theory of consumer demand directly on demand functions without invoking preference as an undefined primitive. If $f(p, m)$ denotes the consumer's unique choice at (p, m) from the opportunity set there, we say that commodity bundle x is *revealed to be preferred to* commodity bundle y if $y \neq x$ and there is a (p, m) at which $x = f(p, m)$ and $p_1y_1 + \cdots + p_ny_n \leq m$. Conditions can then be stated (Uzawa 1960; Houthakker 1961; Hurwicz and Richter 1971) for the revealed preference relation such that there exists a utility function u on X for which $\max_u Y = \{f(p, m)\}$ when Y is the opportunity set at (p, m), for every such Y.

The revealed preference approach in demand theory has stimulated a more general theory of choice functions. A *choice function* C is a mapping from a family of non-empty feasible subsets of X into subsets of X such that, for each feasible Y, $C(Y)$ is a non-empty subset of Y. The *choice set* $C(Y)$ describes the 'best' things in Y. Research in this area has identified conditions on C that allow it to be represented in interesting ways. Examples appear in Fishburn (1973, chapter 15) and Sen (1977). One is the condition.

$$\text{if } Y \subseteq Z \text{ and } Y \cap C(Z) \text{ is non} - \text{empty, then } C(Y) = Y \cap C(Z).$$

When every two-element and three-element subset of X is feasible, this implies that the revealed preference relation $\succ_r$, defined by $x \succ_r y$ if $x \neq y$ and $C(\{x, y\}) = \{x\}$, is a weak order. The weaker condition

$$\text{if } Y \subseteq z \text{ then } Y \cap C(Z) \subseteq C(Y)$$

implies that $\succ_r$ has no cycles when every non-empty finite subset of X is feasible.

Decisions Under Risk

Let P be a convex set of probability measures on an algebra $\mathscr{A}$ of subsets of an outcome set X. Thus, for every p in $\mathscr{P}$, $p(A) \geq 0$ for all $\mathscr{A}$ in $\mathscr{A}$, $p(A) = 1$, and $p(AUB) = p(A) + p(AUB)$ whenever A and B are disjoint sets in $\mathscr{A}$. *Convexity* means that $\lambda p + (1 - \lambda)q$ is in $\mathscr{P}$ whenever p and q are in $\mathscr{P}$ and $0 \leq \lambda \leq 1$. We assume that each $\{x\}$ is in $\mathscr{A}$ and each measure that has $p(\{x\}) = 1$ for some x in X is in. $\mathscr{P}$

The basic expected utility representation is, for all p and q in $\mathscr{P}$,

$$p \succ q \Leftrightarrow \int_X u(x)\mathrm{d}p(x) > \int_X u(x)\mathrm{d}q(x), \qquad (4)$$

where u is a real valued function on X. When u satisfies (4), it is unique up to a positive linear transformation. The expected utility representation follows from the preference axioms of von Neumann and Morgenstern (1947) when each p in $\mathscr{P}$ has $p(A) = 1$ for a finite A in $\mathscr{A}$. Other cases are axiomatized in Fishburn (1970, 1982a). The most important axiom besides weak order is the *independence condition* which says that, for all p, q and r in P and all $0 < \lambda < 1$,

$$p \succ q \Rightarrow \lambda p + (1 - \lambda)r \succ \lambda q + (1 - \lambda)r. \qquad (5)$$

If \$5000 with certainty is preferred to a 50–50 gamble for \$12,000 or \$0, then (5) says that a 50–50 gamble for \$5000 or – \$20,000 will be preferred to a gamble that returns \$12,000 or \$0 or – \$20,000 with probabilities 1/4, 1/4 and 1/2 respectively: $r(-\$20{,}000) = 1$ and $\lambda = 1/2$.

The principle of choice for expected utility says to choose an expected-utility maximizing decision or measure in the feasible subset $\mathscr{L}$ of $\mathscr{P}$ when much a measure exists. Since convex combinations of measures in $\mathscr{L}$ can be formed at little or no cost with the use of random devices, feasible sets are often assumed to be convex. Although this will not create a maximizing combination when none existed prior to convexification under the usual expected utility model, it can create maximally-preferred measures in more general theories that allow cyclic

preferences. Convex feasible sets are also important in the minimax theory of noncooperative games (Nash 1951) and economic equilibrium without ordered preferences (Mas-Colell 1974; Shafer and Sonnenschein 1975).

Expected utility for the special case of monetary outcomes has sired extensive literatures on risk attitudes (Pratt 1964; Arrow 1974) and stochastic dominance (Whitmore and Findlay 1978; Bawa 1982). Risk attitudes involve curvature properties of an increasing utility function ($u'' < 0$ for risk aversion) and their economic consequences for expected-utility maximizing agents. Stochastic dominance relates shape features of u to distribution function comparisons. For example, $\int u\mathrm{d}p \geq \int u\mathrm{d}q$ for all increasing u if and only if $p(\{x{:}x \geq c\}) \geq q(\{x{:}x \geq c\})$ for every real c.

Alternatives to expected utility maximization with monetary outcomes base criteria for choice on distribution function parameters such as the mean, variance, below-target semivariance, and loss probability (Markowitz 1959; Libby and Fishburn 1977). The best known of these are mean (more is better)/variance (more is worse) models developed by Markowitz (1959), Tobin (1965) and others. Whether congruent with expected utility or not (Chipman 1973), such models assume that preferences between distributions depend only on the parameters used.

Recent research in utility/decision theory of risky decisions has been motivated by empirical results (Allais and Hagen 1979; Kahneman and Tversky 1979; Slovic and Lichtenstein 1983) which reveal systematic and persistent violations of the expected utility axioms, including (5) and transitivity. Alternative utility models that weaken (5) but retain weak order have been proposed by Kahneman and Tversky (1979), Machina (1982), and others. A representation which presumes neither (5) nor transitivity is axiomatized by Fishburn (1982b).

Decisions Under Uncertainty

We adopt Savage's (1954) formulation in which each potential decision is characterized by a function f, called an *act*, from a set S of *states* into a set X of *consequences*. The consequence that occurs if f is taken and state s obtains is $f(s)$. Exactly one state will obtain, the agent does not know which it is, and the act chosen will not affect its realization. Examples of states are possible temperatures in central London at 12 noon next 14 July and possible closing prices of a company's stock next Thursday.

Suppose S and the set F of available acts are finite, and that there is a utility function u on X that satisfies (1) and perhaps other conditions. Choice criteria that avoid the question of subjective probabilities on $\mathscr{S}$ (Luce and Raiffa 1957, chapter 13) include

$$\textit{maximum utility}: \text{choose } f \text{ to maximize } \min_s u[f(s)];$$

$$\textit{minimax loss}: \text{choose } f \text{ to minimize } \max_S \left\{ \max_F u[f(s)] - u[f(s)] \right\};$$

$$\textit{Hurwicz}\,\alpha: \text{given } 0 \leq \alpha \leq 1, \text{choose } f \text{ to maximize } \alpha \max_S u[f(s)] + (1-\alpha)\min_S u[f(s)].$$

Maximin, which maximizes the worst that can happen, is very conservative. Minimax loss (or regret), which is less conservative than maximin, minimizes the maximum difference between the best that could happen and what actually happens. Hurwicz α ranges from maximin ($\alpha = 0$) to 'maximax' ($\alpha = 1$).

Another criterion maximizes the average value of $u[f(s)]$ over s. This is tantamount to the subjective expected utility model with equal probability for each state.

Subjective probability as developed by Ramsey, de Finetti and Savage quantifies partial beliefs by the extent to which we are prepared to act on them. If you would rather bet £100 on horse A than on horse B then, *for you*, A has the higher probability of winning. If your beliefs adhere to appropriate axioms for a comparative probability relation $\succ *$ on the algebra of $\mathscr{S}$ subsets of S (Fishburn 1986) then there is a probability measure ρ on $\mathscr{S}$ such that, for all A and B in $\mathscr{S}$, $A \succ^* B \Leftrightarrow \rho(A) > \rho(B)$.

Savage's axioms for $\succ$ on F (Savage 1954; Fishburn 1970, chapter 14) imply the existence of

a bounded utility function u on X and a probability measure C on $\mathscr{S}$ such that, for all f and g in F

$$f \succ g \Leftrightarrow \int_S u[f(s)]\mathrm{d}\rho(s) > \int_S u[g(s)]\mathrm{d}\rho(s), \quad (6)$$

with u unique up to a positive linear transformation and ρ unique. His axioms include weak order, independence axioms that in part yield the preceding representation of $\succ^*$, and a continuity condition. Many other people (Fishburn 1981) have developed alternative axiomatizations of (6) and closely-related representations.

Recent alternatives to Savage's subjective expected utility theory have been motivated by the empirical results cited in the preceding section and by Ellsberg's (1961) challenges to the traditional subjective probability model. Suppose an urn contains 30 red balls and 60 others that are black and yellow in an unknown proportion. One ball is to be drawn at random. Many people are observed to prefer a bet on *red* rather than *black*, and a bet on *black or yellow* rather than *red or yellow*. By the traditional model, the first preference gives $\rho(\text{red}) > \rho(\text{black})$ and the second gives $\rho(\text{black}) > \rho(\text{red})$.

Schmeidler (1984) axiomatizes a utility model that replaces the additive probability measure ρ in (6) by a monotone $[A \subset B \Rightarrow \rho(A) \leq \rho(B)]$ but not necessarily additive measure and argues that his model can accommodate Ellsberg's phenomena. A different model (Loomes and Sugden 1982) uses additive p but accommodates other violations of independence and cyclic preferences.

Maximization of subjective expected utility is the core principle of Bayesian decision theory (Savage 1954; Raiffa and Schlaifer 1961; Winkler 1972). This name, used in distinction to classical methods of statistical analysis pioneered by R.A. Fisher, Jerzy Neyman, Egon Pearson, and Abraham Wald, recognizes the unabashed use of subjective probability and the revision of probabilities in light of new evidence by the basic formula of conditional probability known as Bayes's Theorem.

A typical problem is statistical decision theory is to decide which of several possible experiments, if any, to perform for the purpose of gathering additional information that will be used in a subsequent decision. In the Bayesian approach, the primary states that occasion the need for further information can be enriched to incorporate potential experimental outcomes in such a way that (6) refers to the entire decision process. The problem can then be decomposed, as is usually done in practice, to compute optimal subsequent decisions based on particular experiments and their possible outcomes. Decision functions for each experiment that map outcomes into best subsequent acts can then be compared to determine a best experiment. Various methods of analysis in the Bayesian mode are described and illustrated in Raiffa and Schlaifer (1961).

See Also

- ▶ Decision Theory
- ▶ Expected Utility and Mathematical Expectation
- ▶ Expected Utility Hypothesis

Bibliography

Allais, M., and O. Hagen (eds.). 1979. *Expected utility hypotheses and the Allais paradox*. Dordrecht: Reidel.

Arrow, K.J. 1974. *Essays in the theory of risk bearing*. Amsterdam: North-Holland.

Bawa, V.S. 1982. Stochastic dominance: A research bibliography. *Management Science* 28: 698–712.

Bernoulli, D. 1738. Specimen theoriae novae de mensura sortis. *Commentarii Academiae Scientarium Imperalis Petropolitanae* 5: 175–192. Trans. L. Sommer, *Econometrica* 22: 23–36, (1954).

Chipman, J.S. 1973. The ordering of portfolios in terms of mean and variance. *Review of Economic Studies* 40: 167–190.

Ellsberg, D. 1961. Risk, ambiguity, and the Savage axioms. *Quarterly Journal of Economics* 75: 643–669.

Fishburn, P.C. 1970. *Utility theory for decision making*. New York: Wiley.

Fishburn, P.C. 1973. *The theory for social choice*. Princeton: Princeton University Press.

Fishburn, P.C. 1974. Lexicographic orders, utilities, and decision rules: A survey. *Management Science* 20: 1442–1471.

Fishburn, P.C. 1981. Subjective expected utility: A review of normative theories. *Theory and Decision* 13: 139–199.

Fishburn, P.C. 1982a. *The foundations of expected utility*. Dordrecht: Reidel.

Fishburn, P.C. 1982b. Nontransitive measurable utility. *Journal of Mathematical Psychology* 26: 31–67.
Fishburn, P.C. 1986. The axioms of subjective probability. *Statistical Science* 1: 335–345.
Hicks, J.R., and R.G.D. Allen. 1934. A reconsideration of the theory of value, I, II. *Economica* 1: 52–75, 196–219.
Houthakker, H.S. 1961. The present state of consumption theory. *Econometrica* 29: 704–740.
Hurwicz, L., and M.K. Richter. 1971. Revealed preference without demand continuity assumptions. In *Preferences, utility, and demand*, ed. J.S. Chipman, L. Hurwicz, M.K. Richter, and H.F. Sonnenschein. New York: Harcourt Brace Jovanovich.
Kahneman, D., and A. Tversky. 1979. Prospect theory: An analysis of decision under risk. *Econometrica* 47: 263–291.
Libby, R., and P.C. Fishburn. 1977. Behavioural models of risk taking in business decisions: A survey and evaluation. *Journal of Accounting Research* 15: 272–292.
Loomes, G., and R. Sugden. 1982. Regret theory: An alternative theory of rational choice under uncertainty. *Economic Journal* 92: 805–824.
Luce, R.D., and H. Raiffa. 1957. *Games and decisions*. New York: Wiley.
Machina, M.J. 1982. 'Expected utility' analysis without the independence axiom. *Econometrica* 50: 277–323.
Markowitz, H. 1959. *Portfolio selection*. New York: Wiley.
Mas-Colell, A. 1974. An equilibrium existence theorem without complete or transitive preferences. *Journal of Mathematical Economics* 1: 237–246.
Nash, J. 1951. Non-cooperative games. *Annals of Mathematics* 54: 286–295.
Pratt, J.W. 1964. Risk aversion in the small and in the large. *Econometrica* 32: 122–136.
Raiffa, H., and R. Schlaifer. 1961. *Applied statistical decision theory.* Boston: Division of Research, Graduate School of Business, Harvard University.
Ramsey, F.P. 1931. Truth and probability. In *The foundations of mathematics and other logical essays*, ed. R.B. Braithwaite. New York: Harcourt, Brace. Reprinted in *Studies in subjective probability*, ed. H.E. Kyburg and H.E. Smokler, 61–92. New York: Wiley, 1964.
Savage, L.J. 1954. *The foundations of statistics*. New York: Wiley. 2nd rev. ed, Dover Publications, 1972.
Schmeidler, D. 1984. Subjective probability and expected utility without additivity. Reprint no. 84, Institute for Mathematics and its Application, University of Minnesota.
Sen, A. 1977. Social choice theory: A re-examination. *Econometrica* 45: 53–89.
Shafer, W., and H. Sonnenschein. 1975. Equilibrium in abstract economies without ordered preferences. *Journal of Mathematical Economics* 2: 345–348.
Slovic, P., and S. Lichtenstein. 1983. Preference reversals: A broader perspective. *American Economic Review* 73: 596–605.
Sonnenschein, H.F. 1971. Demand theory without transitive preferences, with applications to the theory of competitive equilibrium. In *Preferences, utility, and demand*, ed. J.S. Chipman, L. Hurwicz, M.K. Richter, and H.F. Sonnenschein, 215–223. New York: Harcourt Brace Jovanovich.
Stigler, G.J. 1950. The development of utility theory: I, II. *Journal of Political Economy* 58: 307–327, 373–396.
Tobin, J. 1965. The theory of portfolio selection. In *The theory of interest rates*, ed. F.H. Hahn and F.P.R. Brechling, 3–51. New York: Macmillan.
Uzawa, H. 1960. Preference and rational choice in the theory of consumption. In *Mathematical methods in the social sciences, 1959*, ed. K.J. Arrow, S. Karlin, and P. Suppes, 129–148. Stanford: Stanford University Press.
von Neumann, J., and O. Morgenstern. 1944. *Theory of games and economic behavior*. Princeton: Princeton University Press. 2nd ed, 1947; 3rd ed, 1953.
Whitmore, G.A., and M.C. Findlay (eds.). 1978. *Stochastic dominance*. Lexington: Heath.
Winkler, R.L. 1972. *An introduction to Bayesian inference and decision*. New York: Holt, Rinehart and Winston.

Utopias

Gregory Claeys

The word 'utopia' is derived from a Greek term meaning 'no place'. A utopia is a fictional account of a perfect or ideal society which in its economic aspect is usually stationary and often includes community of goods. Many proposals for social reform have included elements inspired by utopias, and most utopias at least tacitly plead for social change. There is no single utopian tradition and thus no unilinear relationship between 'utopia' and the history of economic thought. Insofar as the provision of a subsistence for mankind has been the aim of all forms of normative economic thought, however, the mode of thinking about perfect or harmonious societies termed 'utopian' has usually presented itself as the most comprehensive answer to the riddles offered by economic writers. Particularly in the modern period this has involved the use of science and technology to solve economic problems. In turn, the most

ambitious plans to settle all economic difficulties have themselves often verged upon the utopian (in the sense of being particularly fanciful or unachievable). A clarification of this relationship requires distinguishing utopian thought from at least four related modes of speculation. In millenarianism, all social problems are disposed of through divine intervention, often in the form of the Second Coming of Christ, at which time a perfect society is founded. In the medieval English poetic vision described in the 'Land of Cockaygne' and similar works, all forms of scarcity are dissolved in a fantasy of satiety, where desires remain fixed while their means of satisfaction increase without labour and are consumed without effort. In arcadias, a greater stress is given to the satisfaction of 'natural' desires alone and to the equal importance of a spiritual and aesthetic existence. In what has been termed the 'perfect moral community' the necessity for a prior change in human nature and especially in human wants is also assumed and more attention is given to spiritual regeneration as the basis of social harmony.

In all forms of ideal societies the problem of wants or needs is central. The utopian tradition has tended to accept the central tension between limited resources and insatiable appetites, neither ignoring the problem nor assuming any essential change in human nature (Fuz (1952) has termed 'utopias of escape' those which begin with the assumption of plenty, 'utopias of realization' those which presume scarcity as a starting-point). Most utopias attempt instead to control the key forms of social malaise (crime, poverty, vice, war, etc.) which result from human frailty, giving greater stress to the best organization of social institutions rather than idealizing either nature (as in the Land of Cockaygne) or man (as does the perfect moral commonwealth), and relying upon designs fostered by human ingenuity rather than those derived from divine foresight. In economic as well as other aspects, utopias seek the perfection of a completely ordered and detailed social model rather than an interim solution to or partial reform of present disorders. In the imaginative grasp of possibility and presumptive omniscience of exactitude lies the charm and utility as well as the overperfectionist dangers of utopian schemes. Seeking at once to preserve the best of the past and to design an ideal future, utopias have themselves often served as models for judging the adequacy of the present as well as – particularly in the areas of science and technology – its logical development.

As a general rule the economic aspect of the utopian tradition can be understood as moving from a central concern with the maintenance of limited wants and (very often) a community of goods to solve problems of production and distribution, to a greater reliance upon the productive powers provided by science, technology and new forms of economic organization, with less strenuous demands being made for a denial of 'artificial' needs. In this sense the history of utopias mirrors both economic history and the history of economic thought insofar as the latter has legitimized that potential for satisfying greater needs for which scientific and technological development have provided the chief basis. As mainstream liberal political economy came to relinquish the ideal of economic regulation in the eighteenth century, relying instead upon the development of the market to overcome scarcity, utopianism also shifted its emphasis away from the creation of virtue and towards that of organized superfluity and affluence, often in combination with centralized economic planning and organization. Technology has been presumed to have brought a diminution in the amount of socially necessary labour without the necessity for a concomitant reduction in wants. The inevitability of an extreme division of labour has also been supplanted by the vision of alternating forms of more interesting and creative employment in many modern utopias. Contemporary utopianism both builds upon the promises of technology, and remains critical of forms of social organization which fail to develop this potential or to curb its harmful excesses. No longer content to offer a transcendent image of possibility, modern utopianism is moreover committed to the problem of actualizing planned and ideal societies.

Though the utopian genre is usually dated from the publication of Thomas More's *Utopia* (1516), the proposal of a community of goods as a major

element in the solution to economic disorder is much older. An important antecedent was Plato's *Republic* (*c*360 BC), in which the ruling Guardians alone shared their goods in common as a means of ensuring the least conflict between private and public interest. At the end of the second century AD Plutarch wrote his life of the mythical Spartan legislator Lycurgus, who ended avarice, luxury and inequality by an equal division of lands, the replacement of gold and silver by iron coinage, and various sumptuary laws. Though Aristotle was an early and influential critic of Plato's communism, the idea that a community of goods was the ideal state of property survived in various forms in the early Christian era. The very ancient image of a mythical Golden Age of flowing milk and honey which appeared in Hesiod (*c*750 BC), Ovid, and the Stoic-influenced account of the Isles of the Blessed here found a counterpart in the imagery of Paradise and the Garden of Eden, and it was universally assumed that the institution of private property could only have resulted from the Fall and the expulsion of Adam and Eve from Paradise. Some community of goods existed among the Jewish sect of the Essenes, in the early Christian Church as well as later monastic movements, and there was later considerable debate as to whether the Apostles had intended this to hold amongst themselves or for all mankind. But early on the Church offered a robust defence of the naturalness of private property on the grounds that it produced greater peace, order and economic efficiency. Charity, however, and especially the support of the poor in times of necessity, was regarded as the duty accompanying the private ownership of goods on an earth intended by God to be sufficient for the sustenance of all.

This was the tradition which Thomas More, with one eye on Plato and another, perhaps, on the potential of the New World, was to overthrow. In More the possibility of secular, social improvement was revived and now recrafted in a new image of fantasy. Both at this time and later, rapid economic change in Britain was a key reason for the Anglo-centric character of much of the utopian tradition. No doubt angered by the effects of land enclosures on the poor, More gave to the Utopians not only equality but also plenty, six hours' daily work (and more dignity to their activity than had done the ancient utopias), and a rotation of homes every 10 years and of town and country inhabitants more frequently. Public markets made all goods freely available, while public hospitals cared for the sick. National plenty and scarcity were to be balanced by compensatory distribution, while the surplus was in part given away to the poor of other countries and in part sold at moderate rates. Iron was to be esteemed higher than silver or gold, while jewels and pearls were treated as mere baubles fit only for children. Needs were clearly fixed and limited to the level of comforts. With the conquest of the fear of want, greed was largely eliminated, while pomp and excess derived from pride alone were prohibited by law.

The mid-sixteenth century saw a variety of radical Protestant attempts and plans to emulate the purported communism of the early Church (e.g. in the Hutterite Anabaptism of Peter Rideman), and a considerable augmentation to anti-luxury sentiments within a few of the Protestant sects. A preference for agriculture and hostility to luxury typifies most Renaissance utopias, for instance Johann Günzberg's *Wolfaria* (1621), Andreae's *Christianopolis* (1619) (in which a guild model was of some importance), Campanella's *City of the Sun* (1623) (in which slave labour was first abolished in a utopia), and Robert Burton's *Anatomy of Melancholy* (1621) which included a powerful attack upon avarice as well as a national plan for land utilization, the management of economic resources by a bureaucracy, communal granaries, and the public employment of doctors and lawyers. Francis Bacon's *New Atlantis* (1627) was less concerned with the details of economic organization than with the justification of the rule of scientists, and established a paradigmatic attitude towards technology often repeated in later utopias. Bacon also paid some heed to the dangers posed by novelties generally to social order, while Samuel Gott's *Nova Solyma* (1648) was more severe in its condemnation of luxury and intolerance of waste. Of the utopias of the English civil war period, two are particularly worthy of note. Gerrard Winstanley's

The Law of Freedom in a Platform (1652) developed the Diggers' efforts to reclaim common land for the poor into a scheme for the communal ownership of all land which included universal agricultural labour to age 40. Public storehouses were to make all necessary goods freely available as needed, while domestic buying and selling and working for hire were prohibited. Gold and silver were to be used for external trade alone. Better known was James Harrington's *Oceana* (1656), which popularized the proposal for agrarian laws in order to prevent the predominance of the aristocracy and urged a limit upon dowries and inheritance for similar reasons.

The late seventeenth century occasioned a profusion of welfare or full employment utopias in Britain (only in the following century would France see as rich a development of the genre). At this time schemes for practical, immediate social reform and utopias proper were often not far removed. It is in this period, too, that we begin to find a shift away from a concern with a limited demand and the satisfaction of only natural wants towards a conception of maximized production with the full employment of people and resources and a minimization of waste (goals to some extent shared by mainstream Mercantilism). Such aims are evident in, for example, *A Description of the Famous Kingdom of Macaria* (1641), where most legislation is concerned with regulating the production of wealth, Peter Chamberlen's *The Poore Man's Advocate* (1946), which included a detailed scheme for the joint-stock employment of the poor to be supervised by public officials, Peter Plockhoy's *A Way Propounded to Make the Poor in These and Other Nations Happy* (1659), which proposed the resettlement into communities of an elite of artisans, husbandmen and traders, and John Bellers' *Proposals for Raising a Colledge of Industry* (1695), in which the wealthy would help to found communities where the poor were to support them while also providing a decent subsistence for themselves. In such plans, solutions to economic distress tended to focus increasingly upon isolated communities rather than the nation-state, and upon segments of the population rather than, for example, all the poor. It has been suggested (by J.C. Davis 1981) that this implied a waning confidence in the ability of the state to tackle the problem of poverty, and certainly it seems evident that the Act of Settlement of 1662 transferred this burden to individual parishes and away from central government.

The period between 1700 and 1900 marks not only the great age of utopian speculation, but also the period in which economic practice and utopian precept become increasingly intertwined. In addition, it was here that a community of goods ceased to be the *sine qua non* of utopian ideas of property, and that the liberal view of the benefits of private property ownership itself was expressed in utopian form. This entailed a combination of utopian thought and the theory of progress, though in the genre as a whole the two are usually understood as contradictory. In both modern socialism and classical political economy, then, needs are perceived as virtually unlimited, and social harmony is contingent largely upon their fulfilment. The homage to *homo oeconomicus* is usually understood to have begun in Daniel Defoe's *Robinson Crusoe* (1719), and was at its most exalted in Richard Cobden and John Bright's mid-nineteenth-century claims about the universal peace which would be incumbent upon the global extension of free trade. One of its first serious challenges was in John Stuart Mill's acceptance after 1850 of the desirability of a steady-state economy in which further economic development was avoided. Many eighteenth-century utopias were devoted to the notion of progress (e.g. Mercier's *L'An 2440* (1770) and Condorcet's *L'Esquisse d'un Tableau historique des progrès de l'ésprit humain* (1794)). In others the critique of commercial society took various forms, such as Swift's gentle satire in *Gulliver's Travels* (1726), where the Houyhnhnms showed great disdain for shining stones and distributed their produce according to need, or Rousseau's more biting castigation of civilization in his *Discours sur l'orgine de l'inégalité* (1755). Similar criticisms were developed into the foundations of modern communism in the writings of Raynal, Mercier, Mably, Morelly, Babeuf and in Britain, Spence and Godwin. In many of these the Spartan model was of some importance, and luxury seen as a principal source of working class oppression as well as general moral corruption.

Though the entire utopian edifice was severely shaken by the pessimistic prognosis of Malthus' *Essay on Population* (1798), the first half of the nineteenth century witnessed the widespread foundation of small 'utopian socialist' ideal communities which aimed to bring utopian goals into practice, and which could be essentially communistical (Robert Owen, Etienne Cabet) or semi-capitalist (Charles Fourier). Other plans concentrated upon the nation-state and the beneficial development of large-scale industry (Saint-Simon), a pattern which was to become increasingly dominant as the potential role of machinery in creating a new cornucopia became evident. (Some disenchantment with this view occurred later, however, for example in William Morris's *News from Nowhere* (1890), with its preference for rustic and artisanal virtues.) Considerably more attention came to be paid in the early nineteenth century (by Owen and Fourier, e.g.) to the disadvantages of too narrow a division of labour and the benefits of task rotation. At mid-century began the most compelling radical vision of the age in the works of Marx and Engels, whose plans qualify as utopian in the degree to which they inherited overly optimistic assumptions about human nature, technology and social organization in a future society in which private property and alienation were to be superseded. The last 20 years of the century found at least in Britain and America a virtually continuous outpouring of planned economy utopias, of which the best known are Edward Bellamy's *Looking Backward* (1887), which included provisions for the abolition of money, equal wages and credit for all, and an industrial army, W.D. Howells's *A Traveller from Altruria* (1894), and H.G. Wells's *A Modern Utopia* (1905), which made some effort to incorporate a conception of progress into the ideal image of the future, and included a mixed rather than wholly publicly owned economy.

In the twentieth century utopianism has faltered in face of some of the consequences of modernity, and speculation has often taken the form of the negative utopia or dystopia. In the most famous of these, George Orwell's *Nineteen Eighty-Four* (1949), both capitalist aggression and inequality and communist despotism were criticized, with a central thesis of the work being the prevention of the majority enjoying the benefits of mass production via the deliberate destruction of commodities in war. More satirical of the hedonist utopia is Aldous Huxley's *Brave New World* (1932), though Huxley's later *Island* (1962) is a positive utopia which criticises the spiritual impoverishment of an overly-materialistic civilization. Late twentieth century popular utopianism has included some works of science fiction, the libertarian speculation of Murray Rothbard and Robert Nozick (*Anarchy, State, and Utopia*, 1974), and the steady-state environmentalism of Ernest Callenbach's *Ecotopia* (1975). With the progressive extension of both machinery and the welfare state, utopias developing such themes optimistically have declined. To those sated with goods some of the attractions of the consumerist paradise have faded. Technological determinism has often seemingly rendered forms of economic organization unimportant. Two world wars and the spectre of nuclear catastrophe have dented confidence in human perfectibility, while half a century's experimentation with centrally planned communism has lent little credence to the view that this provides the surest path to moral and economic improvement. Nor is 'growth' any longer an uncritically accepted ideal even amongst those who have not yet experienced its effects. Nonetheless the utility of utopias to economic thought is undiminished, for they offer both illumination into important aspects of the history of economic ideas (especially in the areas of welfare and planning), as well as an imaginative leap into possible futures into which more positivist and empirically based thinking fears to wander. If 'progress' can be realized without 'growth', it will likely first persuasively appear in utopian form.

See Also

- ▶ Anarchism
- ▶ Full Communism
- ▶ Individualism
- ▶ Socialism

Bibliography

Adams, R.P. 1949. The social responsibilities of science in *Utopia, New Atlantis* and after. *Journal of the History of Ideas* 10: 374–398.

Armytage, W.H.G. 1984. Utopias: The technological and educational dimension. In *Utopias*, ed. P. Alexander and R. Gill. London: Duckworth.

Boguslaw, R. 1965. *The new Utopians: A study of system design and social change*. Englewood Cliffs: Prentice-Hall.

Bowman, S. 1973. Utopian views of man and the machine. *Studies in the Literary Imagination* 6: 105–120.

Claeys, G. 1986. Industrialism and hedonism in Orwell's literary and political development. *Albion* 18: 219.

Claeys, G. 1987. *Machinery, money and the millennium. From moral economy to socialism*. Oxford: Polity Press.

Dautry, J. 1961. Le pessimisme économique de Babeuf et l'histoire des Utopies. *Annales Historiques de la Révolution Francaise* 33: 215–233.

Davis, J.C. 1981. *Utopia and the ideal society. A study of English Utopian writing 1516–1700*. Cambridge: Cambridge University Press.

Eurich, N. 1967. *Science in Utopia*. Cambridge, MA: Harvard University Press.

Farr, J. 1983. Technology in the digger Utopia. In *Dissent and affirmation: Essays in honor of Mulford Sibley*, ed. A.L. Kalleberg, J.D. Moon, and D. Sabia. Bowling Green: Bowling Green University Popular Press.

Flory, C.R. 1967. *Economic criticism in American fiction, 1798 to 1900*. New York: Russell & Russell.

Fogg, W.L. 1975. Technology and dystopia. In *Utopia/Dystopia?* ed. P.E. Richter. Cambridge, MA: Schenkman.

Fuz, J.K. 1952. *Welfare economics in English Utopias from Francis Bacon to Adam Smith*. The Hague: Martinus Nijhoff.

Gelbart, N. 1978. Science in French enlightenment Utopias. *Proceedings of the Western Society for French History* 6: 120–128.

Goodwin, B. 1984. Economic and social innovation in Utopia. In *Utopias*, ed. P. Alexander and R. Gill. London: Duckworth.

Gusfield, J. 1971. Economic development as a modern utopia. In *Aware of Utopia*, ed. D.W. Plath. Urbana: University of Illinois Press.

Hall, A.R. 1972. Science, technology and utopia in the seventeenth century. In *Science and society 1600–1900*, ed. P. Mathias. Cambridge: Cambridge University Press.

Hont, I., and M. Ignatieff. 1983. Needs and justice in the *Wealth of Nations:* An introductory essay. In *Wealth and virtue: The shaping of political economy in the Scottish enlightenment*, ed. I. Hont and M. Ignatieff. Cambridge: Cambridge University Press.

Hudson, W. 1946. Economic and social thought of Gerrard Winstanley: Was he a seventeenth-century Marxist? *Journal of Modern History* 18: 1–21.

Hymer, S. 1971. Robinson Crusoe and the secret of primitive accumulation. *Monthly Review* 23: 11–36.

King, J.E. 1983. Utopian or scientific? A reconsideration of the Ricardian socialists. *History of Political Economy* 15: 345–373.

Klassen, P.J. 1964. *The economics of Anabaptism 1525–60*. The Hague: Mouton.

Krieger, R. 1980. The economics of Utopia. In *Utopias: The American experience*, ed. G.B. Moment and O.F. Kraushaar. London: Scarecrow Press.

Landa, L. 1943. Swift's economic views and mercantilism. *English Literary History* 10: 310–335.

Leiss, W. 1970. Utopia and technology: Reflections on the conquest of nature. *International Social Science Journal* 22: 576–588.

Levitas, R. 1984. Need, nature and nowhere. In *Utopias*, ed. P. Alexander and R. Gill. London: Duckworth.

MacDonald, W. 1946. Communism in Eden? *New Scholasticism* 20: 101–125.

MacKenzie, D. 1984. Marx and the machine. *Technology and Culture* 25: 473–502.

Manuel, F.E., and F.P. Manuel. 1979. *Utopian thought in the western world*. Oxford: Basil Blackwell.

Mumford, L. 1967. Utopia, the city and the machine. In *Utopias and Utopian thought*, ed. F.E. Manuel. Boston: Beacon.

Novak, M. 1976. *Economics and the fiction of Daniel Defoe*. New York: Russell & Russell.

Perrot, J.-C. 1982. Despotische Verkunft und ökonomische Utopie. In *Utopieforschung. Interdisziplinäre Studien zur neuzeitlichen Utopie*, ed. W. Vosskamp. Stuttgart: J.B. Metzlersche Verlagsbuchhandlung.

Pocock, J.G.A. 1980. The mobility of property and the rise of eighteenth-century sociology. In *Theories of property, Aristotle to the present*, ed. A. Parel and T. Flanagan. Waterloo: Wilfred Laurier University Press.

Sargent, L.T. 1981. Capitalist eutopias in America. In *America as Utopia*, ed. K.M. Roemer. New York: Burt Franklin.

Schlaeger, J. 1982. Die Robinsonade als früjbügerliche 'Eutpia'. In *Utopieforschung. Interdisziplinäre Studien zur neuzeitlichen Utopie*, ed. W. Vosskamp. Stuttgart: J.B. Metzlersche Verlagsbuchhandlung.

Schoeck, R.J. 1956. More, Plutarch, and King Agis: Spartan history and the meaning of *Utopia. Philological Quarterly* 35: 366–375.

Segal, H. 1985. *Technological Utopianism in American culture*. Chicago: Chicago University Press.

Sibley, M.Q. 1973. Utopian thought and technology. *American Journal of Political Science* 17: 255–281.

Soper, K. 1981. *On human needs: Open and closed theories in Marxist perspectives*. London: Harvester Press.

Springborg, P. 1981. *The problem of human needs and the critique of civilisation*. London: George Allen & Unwin.

Steintrager, J. 1969. Plato and more's *Utopia. Social Research* 36: 357–372.

Taylor, W.F. 1942. *The economic novel in America*. Chapel Hill: University of North Carolina Press.

Thompson, N.W. 1985. *The People's science. The popular political economy of exploitation and crisis, 1816–34*. Cambridge: Cambridge University Press.

Welles, C.B. 1948. The economic background of Plato's communism. *Journal of Economic History* 8: 101–114.