O

O'Brien, George (1892–1973)

J. Meenan

O'Brien was born and died in Dublin. He turned to Political Economy when ill-health obliged him to retire from the Irish Bar. From 1926 to 1961 he was Professor of National Economics, then of Political Economy, at University College, Dublin.

Throughout his professorship he was at pains to follow developments in economic theory: typically, he lectured fully on the *General Theory* within months of its publication. In general economics his approach was derived from Mill and Marshall. He held that political economy, law and philosophy shared a common root and that no one of them should be separated from the other two. This approach informed his lectures and writing, which displayed a clarity and precision derived from his legal training.

He obtained his chair when the new Irish State was fashioning its economic policies. By membership of a series of Commissions and by articles in informed journals he clarified for the public the issues involved. His insistence on the importance of priorities became less acceptable, but he always wielded influence through his students (many of whom rose to high office), the Statistical Society (President, 1942–6) and the Economic and Social Research Institute (Chairman, 1961–73).

The essay on medieval economic teaching traced the development of the concept of interest from the *Ethics* of Aristotle to the Schoolmen. His notes on profit insisted on its residual quality and its function as the reward of risk-bearing.

He encouraged the young Geoffrey Crowther to write his *Outline of Money* (1940, Preface), and he communicated the discovery of the lost Ricardo–Mill letters (*Economica,* November 1943).

Selected Works

1918. *The economic history of Ireland in the eighteenth century*. Dublin: Maunsel.
1919. *The economic history of Ireland in the seventeenth century*. Dublin: Maunsel.
1920. *Essay on medieval economic teaching*. London: Longmans.
1921. *The economic history of Ireland from the union to the famine*. London: Longmans.
1923. *Essay on the economic effects of the reformation*. London: Burns, Oates and Washbourne.
1929. *Agricultural economics*. London: Longmans.
1929. *Notes on the theory of profit*. Dublin: Hodges Figgis.
1942. *Economic relativity*. Dublin: Statistical & Social Inquiry Society of Ireland.
1948. *The phantom of plenty: Reflections on economic progress*. Dublin: Clonmore and Reynolds.

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

Observational Learning

Lones Smith and Peter Norman Sørensen

Abstract

Observational learning occurs when privately informed individuals sequentially choose among finitely many actions after seeing predecessors' choices. We summarise the general theory of this paradigm: *belief convergence* forces *action convergence*; specifically, copycat 'herds' arise. Also, beliefs converge to a point mass on the truth exactly when the private information is not uniformly bounded. This subsumes two key findings of the original herding literature: With multinomial signals, *cascades* occur, where individuals rationally ignore their private signals, and incorrect herds start with positive probability. The framework is flexible – some individuals may be committed to an action, or individuals may have divergent cardinal or even ordinal preferences.

Keywords

Action herd; Experimentation; Information aggregation; Informational cascade; Informational herding; Limit cascade; Markov process; Martingale; Observational learning; Social learning; Stochastic difference equation

JEL Classifications

D8; D83

Observational Learning

Suppose that an infinite number of individuals each must make an irreversible choice among finitely many actions – encumbered solely by uncertainty about the state of the world. If preferences are identical, there are no congestion effects or network externalities, and information is complete and symmetric, then all ideally wish to make the same decision.

Observational learning occurs specifically when the individuals must decide sequentially, all in some preordained order. Each may condition his decision both on his endowed private signal about the state of the world and on all his predecessors' decisions, but *not* their hidden private signals. This article summarizes the general framework for the herding model that subsumes all signals, and establishes the correct conclusions. The framework is flexible – e.g., some individuals may be committed to an action, or individuals may have divergent preferences.

Banerjee (1992) and Bikhchandani et al. (1992) (hereafter, BHW) both introduced this framework. Ottaviani and Sørensen (2006) later noted that the same mechanism drives expert herding behaviour in the earlier model of Scharfstein and Stein (1990), after dropping their assumption that private signals are conditionally correlated. In BHW's logic, *cascades* eventually start, in which individuals rationally ignore their private signals. Copycat action herds therefore arise *ipso facto*. Also, despite the surfeit of available information, a herd develops on an incorrect action with positive probability: after some point, everyone might just settle on the identical less profitable decision. This result sparked a welcome renaissance in informational economics. Observational learning explains correlation of human behaviour in environments without network externalities where one might otherwise expect greater independence. Various twists on the herding phenomenon have been applied in a host of settings from finance to organisational theory, and even lately into experimental and behavioural work.

In this article, we develop and flesh out the general theory of how Bayes-rational individuals sequentially learn from the actions of posterity, as developed in Smith and Sørensen (2000). Our logical structure is to deduce that almost sure *belief convergence* occurs, which in turn forces *action convergence*, or the action herds. Also, beliefs converge to a point mass on the correct state exactly when the private signal likelihood ratios are not uniformly bounded. For instance, incorrect herds arose in the original herding

papers since they assumed finite multinomial signals. We hereby correct a claim by Bikhchandani et al. (2008), which unfortunately concludes, 'In other words, in a continuous signals setting herds tend to form in which an individual follows the behaviour of his predecessor with high probability, even though this action is not necessarily correct. Thus, the welfare inefficiencies of the discrete cascades model are also present in continuous settings'.

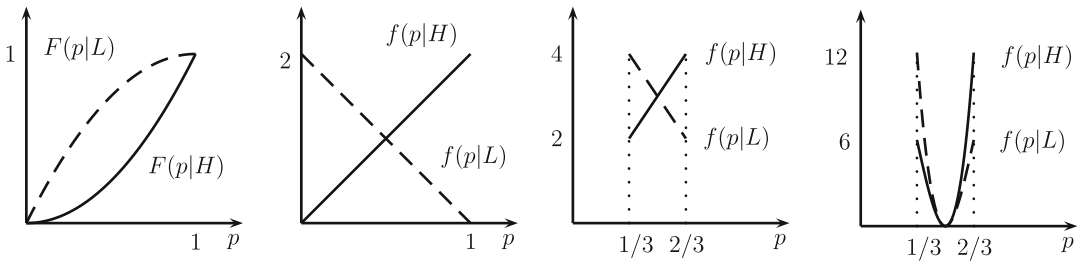
Multinomial signals also violate a log-concavity condition, and for this reason yield the rather strong form of belief convergence that is a cascade. One recent lesson is the extent to which cascades are the exception rather than rule.

The Model

Assume a completely ordered sequence of individuals 1, 2,... Each faces an identical binary choice decision problem, choosing an action $a \in \{1, 2\}$. Individual n's payoff $u(a_n, \omega)$ depends on the realisation of a state of the world, $\omega \in \{H, L\}$, common across n. The high action pays more in the high state: $u(1, L) > u(2, L)$ and $u(1, H) < u(2, H)$. Individuals act as Bayesian expected utility maximisers, choosing action $a = 2$ above a threshold posterior belief $\overline{r}$, and otherwise action $a = 1$. All share a common prior $q_0 = P(\omega = H)$, and for simplicity, $q_0 = 1/2$.

The decision-making here is partially informed. For exogenous reasons, each individual n privately observes the realisation of a noisy signal σ_n, whose distribution depends on the state ω. Conditional on ω, signals are independently and identically distributed. Observational learning is modelled via the assumption that individual i can observe the full history of actions $h_n = (a_1, \ldots, a_{n-1})$. While predecessors' private signals cannot be observed directly, they may be partially inferred. The interesting properties of observational learning follow because the private signals are filtered by coarse public action observations.

The private observation of signal realisation σ_n, with no other information, yields an updated *private belief* $p_n \in [0, 1]$ in the state of the world $\omega = H$. The private belief p_n is a sufficient statistic for the private signal σ_n in the nth individual's decision problem. Its cumulative distribution $F(p|\omega)$ in state ω is a key primitive of the model. Define the unconditional cumulative distribution $F(p) = [F(p|H) + F(p|L)]/2$. The theory is valid for arbitrary signal distributions, having a combination of discrete and continuous portions. But to simplify the exposition, we assume a continuous distribution with density f. The state-conditional densities $f(p|\omega)$ obey the Bayesian relation $p = (1/2)f(p|H)/f(p)$ with $f(p) = [f(p|H) + f(p|L)]/2$, implying $f(p|H) = 2pf(p)$ and $f(p|L) = 2(1-p)f(p)$. The equality $f(p|H)/f(p|L) = p/(1-p)$ can be usefully reinterpreted as a *no introspection condition*: understanding the model likelihood ratio of one's private belief p does not allow any further inference about the state. This special ratio ordering implies that the conditional distributions share the same support, but that $F(p|H) < F(p|L)$ for all private beliefs strictly inside the support (Fig. 1).



Observational Learning, Fig. 1 Private belief distributions. At *left* are generic private belief distributions in the states L, H, illustrating the stochastic dominance of $F(\cdot|H) \succ F(\cdot|L)$. The three other panels depict the specific densities for the unbounded and bounded private belief signal distributions discussed in the text

Private beliefs are said to be *bounded* if there exist p', $p'' \in (0, 1)$ with $F(p') = 0$ and $F(p'') = 1$, and *unbounded* if $F(p) \in (0, 1)$ for all $p \in (0, 1)$. For instance, a uniform density $f(p) \equiv 1$ results in the unbounded private belief distributions $F(p|H) = p^2 < 2p - p^2 = F(p|L)$. But if $f(p) \equiv 3$ on the support [1/3, 2/3], then the bounded private belief distributions are $F(p|H) = (3p - 1)(1 + 3p)/3 < (3p - 1)(5 - 3p)/3 = F(p|L)$.

Analysis via Stochastic Processes

Because only the actions are publicly observed with observational learning, the *public belief* q_n in state H is based on the observed history of the first n–1 actions alone. The associated *likelihood ratio* of state L to state H is then $\ell_n = (1 - q_n)/q_n$. And if so desired, we can recover public beliefs from the likelihood ratios using $q_n = 1/(1 + \ell_n)$. Incorporating the most recent private belief p_n yields the posterior belief $r_n = p_n/(p_n)\ \ell_n(1 - p_n))$ in state H. So indifference prevails at the *private belief threshold* $\overline{p}(\ell)$ defined by

$$\overline{r} \equiv \frac{\overline{p}(\ell)}{\overline{p}(\ell) + \ell(1 - \overline{p}(\ell))} \tag{1}$$

Individual n chooses action $a = 1$ for all private beliefs $p_n \leq \overline{p}(\ell)$, and otherwise picks $a = 2$. Since higher public beliefs (i.e., lower likelihood ratios) compensate for lower private beliefs in Bayes Rule, the threshold is monotone $\overline{p}'(\ell) > 0$.

We now construct the public stochastic process. Given the likelihood ratio ℓ, action $a = 1, 2$ happens with chance $\rho(a|\ell, \omega)$ in state $\omega \in \{H, L\}$, where

$$\rho(1|\,\ell, \omega) \equiv F(\overline{p}(\ell)|\,\omega) \equiv 1 - \rho(2|\,\ell, \omega) \tag{2}$$

When individual n takes action a_n, the updated public likelihood ratio is

$$\ell_{n+1} = \varphi(a_n, \ell_n) \equiv \ell_n \frac{\rho(a_n|\,\ell_n, L)}{\rho(a_n|\,\ell_n, H)} \tag{3}$$

since Bayes' Rule reduces to multiplication in likelihood ratio space due to the conditional independence of private signals. But in light of our stochastic ordering, the binary action choices are informative of the state of the world:

$$\begin{aligned} \rho(1|\,\ell_n, L) > \rho(1|\,\ell_n, H) \quad \text{and} \quad \rho(2|\,\ell_n, L) \\ < \rho(2|\,\ell_n, H) \end{aligned}$$

Observe what has just happened. Choices have been automated, and what remains is a stochastic process (ℓ_n) that is a *martingale*, conditional on state H.

$$E[\ell_{n+1}|\,\ell_1, \ldots \ell_n, H] = \sum_m \rho(m|\,\ell_n, H)\ell_n \frac{\rho(m|\,\ell_n, L)}{\rho(m|\,\ell_n, H)}$$
$$= \ell_n$$

Because the stochastic process (ℓ_n) is a non-negative martingale in state H, the Martingale Convergence Theorem applies. Namely, (ℓ_n) converges almost surely to the (random variable) limit $\ell_\infty = \lim_{n\to\infty} \ell_n$, namely having (finite) values in $[0,\infty)$. The support of ℓ_∞ contains all candidate limit likelihood ratios. Among the most immediate of implications, *learning cannot result in a fully erroneous belief* $\ell = \infty$ *with positive probability.* Just as well, this follows from Fatou's Lemma in measure theory, for $E[\lim\ \inf_{n\to\infty} \ell_n|H] \leq \lim\ \inf_{n\to\infty} E[\ell_n|H] = \ell_0$.

Let's continue to trace this logic, by next observing that the sequence of pairs of actions and likelihood ratios (a_n, ℓ_n) is also a *Markov process* on the domain $\{1, 2\} \times [0,\infty)$. For we can see that each new pair only depends on the last:

$$\begin{aligned} (a_n, \ell_n) \mapsto (a_{n+1}, \varphi(a_{n+1}, \ell_n)) \quad \text{with chance} \\ \rho(a_{n+1}|\,\ell_n, H) \end{aligned}$$

The big gun for Markov processes is the stationarity condition. While our two-dimensional process (a_n, ℓ_n) is clearly non-standard, Smith and Sørensen (2000) prove the following version of the Markov stationarity condition: *If the transition functions* ρ *and* φ *are continuous in* ℓ*, then for any* $\hat{\ell}$ *in the support of* ℓ_∞ *and for all m, we have either* $\rho\left(m|\,H, \hat{\ell}\right) = 0$ *or* $\varphi\left(m, \hat{\ell}\right) = \hat{\ell}$. In other words, either an action does not occur, or it yields no new information, or both.

The stationary points of the (a_n, ℓ_n) process are therefore the *cascade sets*, namely, those sets of likelihood ratios ℓ indexed by actions m that almost surely repeat action m, namely, $\overline{J}_m = \{\ell | \rho(m|\ell, H) = 1\}$. With *bounded private beliefs*, there must exist some high (low) enough likelihood ratios ℓ that pull all private beliefs below (above) the threshold posterior belief $\overline{r}$. In this case, the cascade sets $\overline{J}_1, \overline{J}_2$ for the two actions are both non-empty. When private beliefs are unbounded, the cascade sets collapse to the extreme points, $\overline{J}_1 = \{\infty\}$ and $\overline{J}_2 = \{0\}$. And since we have seen that $\ell = \infty$ cannot arise with positive probability, we must converge to a point mass on the truth (or $\ell = 0$).

Next, we claim that convergence of beliefs implies convergence of actions. Whenever someone optimally chooses action m, any successor must optimally follow suit if he bases his decision just on public information. Individual $n-1$ solves the same decision problem as n faces, but with more information, $(a_1, \ldots, a_{n-2})$ and σ_{n-1}. Contrary actions completely 'overturn' the weight of the entire action history, however long. By this *Overturning Principle*, an infinite subsequence of contrary actions precludes belief convergence. By the Martingale Convergence Theorem, this almost surely cannot happen. By the last paragraph, we conclude that *with unbounded private beliefs, a correct herd eventually arises.*

When Only Correct Herds Arise

Consider an illustrative example, with individuals deciding whether to 'invest' in or 'decline' an investment project of uncertain value. Investing (action 2) is risky, paying $u > 1$ in state H and -1 in state L, declining (action 1) is a neutral action with zero payoff in both states. Indifference prevails at the posterior belief $\overline{r} = 1/(1+u)$. Then Eq. 1 yields the private belief threshold $\overline{p}(\ell) = \ell/(u+\ell)$.

Assume first the earlier unbounded private beliefs example. Then transition chances are $\rho(1|\ell, H) = \ell^2/(u+\ell)^2$ and $\rho(2|\ell, L) = \ell(\ell + 2u)/(u+\ell)^2$, and continuations

$$\varphi(1,\ell) = \frac{u\ell}{u+2\ell} < \ell < \ell)2u \equiv \varphi(2,\ell)$$

by Eqs. 2–3. In other words, the likelihood ratio sequence constitutes a stochastic difference equation. Figure 2 shows how $\overline{J}_2 = \{0\}$ is the only stationary finite likelihood ratio in state H: The limit ℓ_∞ is thus concentrated on 0, the truth.

Whenever action 2 is taken, the new likelihood ratio is $\ell_n \geq 2u$. This can only happen finitely many times.[1] So belief convergence implies action convergence, namely, a herd. This example precisely illustrates the logic for one main result: interestingly, a herd arises despite the fact that a cascade never does, since at each and every stage, a contrary action was possible. Since convergence occurs towards the cascade set but forever lies outside, this is called a *limit cascade*.

When Incorrect Herds Must Sometimes Arise

When private beliefs are bounded, public beliefs still converge, and they result in copycat herds. The main difference now is the positive probability of incorrect herds. Indeed, adjust the last example for the bounded beliefs family. Given the private belief threshold $\overline{p}(\ell) = \ell/(u+\ell)$, the laws of motion (2)–(3) yield transitions

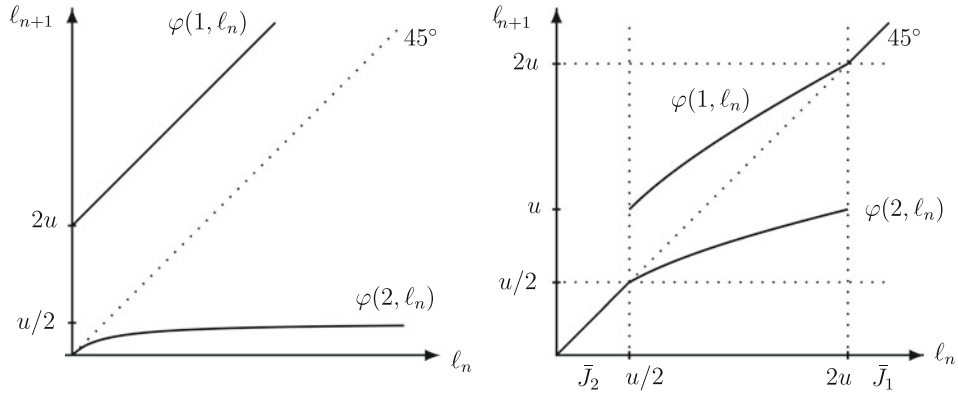
$$\varphi(1,\ell) \equiv \ell\frac{\ell+4u}{5\ell+2u} < \ell < \ell\frac{2\ell+5u}{4\ell+\mathrm{u}} \equiv \varphi(2,\ell)$$

[1] Still, it helps to introspect on exactly *why* no such contrarian can arise. Let the chance that the kth individual breaks the herd be pk, given the state. If these chances vanish fast enough that they are summable, then their tail sum can be made as small as desired. Then by conditional independence, the chance that no one among 1, 2, … , k breaks the herd is positive:

$$(1-p_1)\cdots(1-p_k) >$$

$$1 - p_1 - p_2 - \cdots - p_k > 0.$$

O



Observational Learning, Fig. 2 Transitions and cascade sets. Transition functions for the examples: unbounded private beliefs (*left*), and bounded private beliefs (*right*). By the martingale property, the expected continuation in state *H* lies on the *diagonal*. The *stationary points* are where both arms hit the diagonal, or where one arm is taken with zero chance ($\ell = 0$ in the *left panel*, $\ell \leq 2u/3$ or $\ell \geq 2u$ in the *right panel*)

with probabilities

$$\rho(1|H,\ell) = \frac{(4\ell+u)(2\ell-u)}{3(u+\ell)^2} \quad \text{and}$$
$$\rho(2|L,\ell) = \frac{(\ell+4u)(2u-\ell)}{3(u+\ell)^2}$$

for likelihood ratios $\ell \in (u/2,\ 2u)$. As seen in Fig. 2 (left panel), a cascade can never start after the first individual decides. But since the likelihood ratio must converge, a limit cascade starts, towards one of the cascade sets $\bar{J}_1$ or $\bar{J}_2$. A herd on the corresponding action must then start eventually, lest beliefs fail to converge.

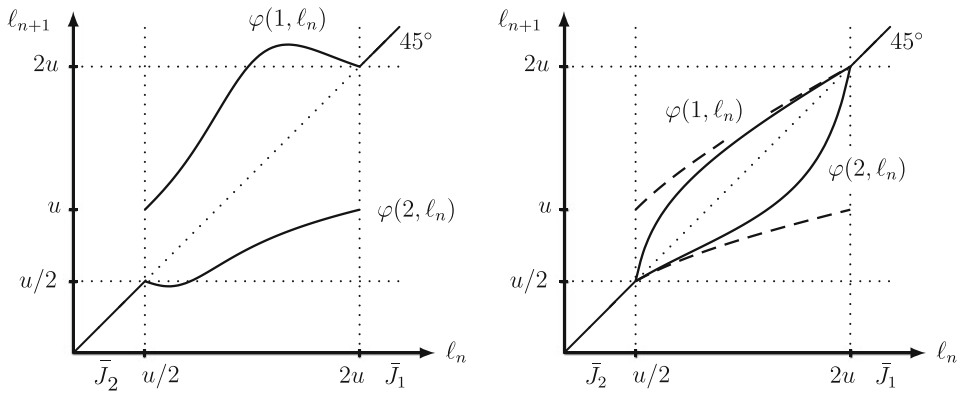
We now explore the easy logic for why *an incorrect herd occurs with strictly positive probability given bounded beliefs.* Again, we appeal to a big gun from measure theory. For if we start at some public likelihood ratio $\ell_0 \in (u/2,\ 2u)$, then by Fig. 2, dynamics are trapped in $(u/2,\ 2u)$. Since $0 \leq \ell_n \leq 2u$, Lebesgue's Dominated Convergence Theorem allows us to swap the expectation and limit operations, and thus conclude that $E[\ell_\infty \mid H\ ] = \lim_{n\to\infty} E[\ell_n \mid H\ ] = \ell_0$. Write $\ell_0 = \pi(u/2) + (1-\pi)(2u)$, where $0 < \pi < 1$ whenever $u/2 < \ell_0 < 2u$. Then the random variable ℓ_∞ places weight π on $u/2$ and weight $1-\pi$ on $2u$. So in state *H*, a herd arises with chance π on action 2, and with chance $1-\pi$ on action 1.

Herds Without Cascades

For an interesting contrast to the discrete signal world of BHW, observe that in Fig. 3 (right panel), if we do not begin in a cascade, we never enter one – even though a herd eventually starts. Indeed, visually, it is clear that $\ell_n \in (u/2,\ 2u)$ for all *n*, provided that initially $\ell_0 \in (u/2,\ 2u)$. So while the analysis in BHW explicitly depended on cascades ending the dynamics in finite time, a somewhat subtler dynamic story emerges here: *Herds must arise even though a contrarian has positive probability at every stage.*

This no-cascades result is robust to changes in both the signal distribution and payoffs, for it arises whenever the continuation functions $\varphi(1,\ell)$, $\varphi(2,\ell)$ are monotone increasing in ℓ. Monotonicity asserts the seemingly plausible condition that a higher prior public belief implies a higher posterior public belief after every action. Yet, despite how intuitive this property may seem, it is violated by any multinomial signal distribution (loosely, because it is 'lumpy').

We have shown in Smith and Sørensen (2008) that the continuation functions are monotone under an easily verifiable regularity condition – namely, that the unconditional density of the log-likelihood ratio $\log(p/(1-p))$ be log-concave. Most popular continuous distributions satisfy this condition, for instance, the Gaussian, uniform or generalised



Observational Learning, Fig. 3 Modified transitions. Transition functions for bounded beliefs with a quadratic density (*left panel*) and uniform bounded beliefs with and without 20% crazy types (*solid* and *dashed* lines in *right panel*). The non-monotonicities of transition functions (*left panel*) imply that a cascade on a starts when a is taken where ℓ_n is sufficiently close to $\bar{J}_a$. The transition function discontinuity in the *right panel* of Fig. 2 vanishes with the addition of crazy types (*right panel*), corresponding to the failure of the overturning principle

exponential. But the analysis in BHW and a vast number of successor papers was based on the multinomial family – namely, the one main signal family for which the regularity condition *fails*. This discussion hereby corrects the claim by Bikhchandani et al. (2008), that 'In some continuous signal settings cascades do not form (Smith and Sørensen 2000)'. On the contrary, one really must view cascades as the informationally rare outcome, a case where a tractable example class proved misleading. The true touchstone of this literature is simply the observed phenomenon of action herding.

Cascades with Smooth Signals

To fully flesh out this picture, we offer an example of a continuous signal distribution that violates the monotonicity result. (This example is based on one included in the original working paper of Smith and Sørensen (2000) found in Sørensen (1996)). To this end, we construct a sufficiently heroic violation of our log-concavity condition. Suppose that private beliefs p have a quadratic density $f(p) = 324(p - 1/2)^2$ over the bounded support [1/3, 2/3]. Then the conditional private belief densities are $f(p|H) = 2pf(p)$ and $f(p|L) = 2(1-p)f(p)$, as depicted in the right panel of Fig. 1. Integration yields the (suppressed) polynomial expressions for $F(p|L)$, $F(p|H)$.

Returning to the running investment payoff example, for all likelihood ratios $\ell \in (u/2, 2u)$, we find the likelihood ratio transitions (left panel of Fig. 3):

$$\varphi(1,\ell) = \ell \frac{23(u+\ell)^3 - 93\ell(u+\ell)^2 + 126\ell^2(u+\ell) - 54\ell^3}{3(u+\ell)^3 + 9\ell(u+\ell)^2 - 54\ell^2(u+\ell) + 54\ell^3},$$

$$\varphi(2,\ell) = \ell \frac{12(u+\ell)^3 - 63\ell(u+\ell)^2 + 108\ell^2(u+\ell) - 54\ell^3}{2(u+\ell)^3 + 3\ell(u+\ell)^2 - 36\ell^2(u+\ell) + 54\ell^3}.$$

A More General Observational Learning Framework

The Overturning Principle may not sound very realistic, *a priori*. Should we expect that a single deviator from an action herd of one million individuals can, entirely by himself, change the course of subsequent play? Is the excessive reliance on the assumption of common knowledge of rationality implicit in the overturning principle reasonable? Experimental results on the informational herding model, e.g., Çelen and Kariv (2004), have cast doubt on this. (The review by Anderson and

Holt (2008) speaks more broadly to such experimental evidence.)

It turns out that our reduction of the model to a stochastic difference equation in the likelihood ratio obeying a martingale property is robust to a wide array of economically inspired modifications that can accommodate deviations from the overturning principle. For instance, suppose that a fraction of 'crazy' individuals randomly choose actions. Figure 3 depicts the modified continuation functions in the right panel, for a case where 10% of individuals are committed to action 1 and 10% are committed to action 2. The remaining population is rational. Since all actions occur with a non-vanishing frequency, none can have drastic effects. Yet the limit beliefs are unaffected by the noise, contrary actions being deemed irrational (and ignored) inside the cascade sets. Of course, the failure of the overturning principle invalidates the argument that limit cascades force herds. But because actions are still informative of beliefs, social learning is productive.

We show more strongly in Smith and Sørensen (2000) that herds nonetheless do arise among all rational (non-crazy) individuals, when beliefs are bounded and have non-zero density near the bounds. Essentially, the public likelihood ratios (ℓ_n) converge so fast that the chance of an infinite string of rational contrarians is zero. (Of course, an outside observer of the action history would hardly be able to detect infrequent rational non-herders, should they occur.)

Alternatively, we may relax the assumption that all individuals solve the same decision problem. Individuals may well have different rational preference types. First, if ordinal preferences are aligned, so that everyone takes action 2 for stronger beliefs in state H, then the limit likelihood ratio ℓ_∞ is focused on the intersection of their respective cascade sets.

Suppose instead that the ordinal preferences differ for some pair of types. Then there arises the possibility of a *confounded learning point*. This is a non-cascade likelihood ratio ℓ^* such that if $\ell_{n-1} = \ell^*$, then individual n's observation of action a_n is non-informative – the probabilities satisfy $\rho(1|H, \ell^*) = \rho(1|L, \ell^*)$. In this case, $\ell_{n+1} = \ell_n$ following either action of individual n. If such a confounding outcome ℓ^* exists, then it is *locally stochastically stable*: there is positive probability that $\ell_\infty = \ell^*$ provided some ℓ_n is ever sufficiently close to ℓ^*.

Conclusion

This model of observational learning explores a modelling framework to analyse imitation of observed behaviour. The model is quite tractable. Public beliefs based on the ever-lengthening action history must converge to a limit, which is among the fixed points of a stochastic difference equation. As long as all ordinal preferences coincide, we eventually settle on an action herd, even though beliefs might never settle down. When private signals sufficiently violate a log-concavity condition, a cascade can arise.

Lee (1993) noted that beliefs can be perfectly revealed when the action space is continuous, just like the belief space. The social learning paradigm instead by and large explores when a coarse action set communicates the private beliefs of decision makers. It may sufficiently frustrates the learning dynamics that an incorrect action herd occurs. If individuals seek to help each other by taking more informative actions, and if this signaling is understood by successors, then any cascade sets shrink, and the welfare of later individuals generally rises. As we show in Smith and Sørensen (2008), the analysis is qualitatively similar to that outlined here, although solving for the new, forward-looking transition chances requires dynamic programming.

A greater message of social learning is the self-defeating nature of learning from others. Moving outside the finite action, sequential entry model into a Gaussian world, Vives (1993) found that social learning is slower than private learning in a market setting where individual decisions are obscured by Gaussian noise.

If observations are not made of an ever-expanding history, such as simply knowing the number but not order of past action choices, then our approach is less useful. The survey by Gale

and Kariv (2008) discusses the problem of learning in networks. In Smith and Sørensen (1994), and Chapter 3 of Sørensen (1996), we identified a case where the stochastic difference equation is a useful tool, even when public beliefs do not follow a martingale.

See Also

- ▶ Information Cascades
- ▶ Information Cascade Experiments
- ▶ Learning and Information Aggregation in Networks

Bibliography

Anderson, L., and C.A. Holt. 2008. Information cascade experiments. In *The new Palgrave dictionary of economics*, ed. S.N. Durlauf and L.E. Blume. New York: Palgrave MacMillan.

Banerjee, A.V. 1992. A simple model of herd behavior. *Quarterly Journal of Economics* 107: 797–817.

Bikhchandani, S., D. Hirshleifer, and I. Welch. 1992. A theory of fads, fashion, custom, and cultural change as information cascades. *Journal of Political Economy* 100: 992–1026.

Bikhchandani, S., D. Hirshleifer, and I. Welch. 2008. Information cascades. In *The new Palgrave dictionary of economics*, ed. S.N. Durlauf and L.E. Blume. New York: Palgrave MacMillan.

Çelen, B., and S. Kariv. 2004. Distinguishing informational cascades from herd behavior in the laboratory. *American Economic Review* 94: 484–498.

Gale, D., and S. Kariv. 2008. Learning and information aggregation in networks. In *The new Palgrave dictionary of economics*, ed. S.N. Durlauf and L.E. Blume. New York: Palgrave MacMillan.

Lee, I.H. 1993. On the convergence of informational cascades. *Journal of Economic Theory* 61: 395–411.

Ottaviani, M., and P.N. Sørensen. 2006. Professional advice. *Journal of Economic Theory* 126: 120–142.

Scharfstein, D.S., and J.C. Stein. 1990. Herd behavior and investment. *American Economic Review* 80: 465–479.

Smith, L., and P. Sørensen. 1994. An example of Non-martingale learning. MIT Working Paper.

Smith, L., and P. Sørensen. 2000. Pathological outcomes of observational learning. *Econometrica* 68: 371–398.

Smith, L., and P. N. Sørensen. 2008. Informational herding and optimal experimentation. University of Copenhagen Working Paper.

Sørensen, P. 1996. Rational social learning. PhD thesis, MIT.

Vives, X. 1993. How fast do rational agents learn? *Review of Economic Studies* 60: 329–347.

Occam's [Ockham's] Razor

S. Hargreaves-Heap and M. Hollis

Called after William of Ockham or Occam (*c*1285–1349), this is the principle usually stated as 'entities are not to be multiplied beyond necessity' (*entitia non multiplicanda sunt praeter necessitatem*). These words are not Ockham's own, although he does say 'plurality is not to be assumed without necessity' and 'what can be done with less is done in vain with more'. The principle belongs with his radical empiricism, by which only direct experience of particular things and events can be evidence for claims to knowledge, and with his nominalism, by which logical analysis of language can be assured of removing the need for extra-linguistic universals. Nothing is to be assumed in explaining a fact, unless established by experience, by reasoning from experience or by the requirements of Faith. Whatever is real is particular.

The context is an old and subtle dispute about how we recognize different things (trees, for instance) as the same. Realism held, in one form or another, that different trees have a common nature. Ockham declared common natures unknowable, unnecessary and indeed unintelligible. They exist only in the sense that trees are rightly all called trees – a fact about the verbal sign and not about the inner being of the tree. For Ockham, the problem of universals reduces to one of showing how concepts arise and function in relation to experience of particulars. His answer is that a concept is an act of understanding the individual things of which it is the concept. This answer has the merit of not multiplying entities beyond necessity. But the problem of universals is still with us and nominalism has never disposed of its critics.

Two examples should serve to illustrate the application of this principle to economics. Consider the use of ordinal and cardinal utility theory. Ordinal theory requires that agents can say whether they prefer option A to B (or are

indifferent), whereas cardinal utility theory also assumes that agents can say by how much more they prefer A to B. Both theories can be used to justify the so-called law of demand, but ordinalism is often preferred because it assumes less about agents. Likewise, the neo-Ricardian rejection of Marx's labour theory of value is based on a similar thought. The price vector can be determined on the basis of the production technology and the real wage: there is no need to use Marx's labour-values, especially as their calculation also depends on the production technology.

See Also

- ▶ Methodology
- ▶ Models and Theory
- ▶ Philosophy and Economics

Occupational Segregation

Myra H. Strober

Neither men and women nor whites and non-whites are distributed equally across occupations. This inequality by gender or race is termed occupational segregation. Occupational segregation by gender is of greater magnitude and has been more persistent over time. Also, it has been more widely studied.

Occupational segregation is generally measured by the index of segregation, I.S., defined as

$$\text{I.S.} = \frac{1}{2}\sum_{i=1}^{m} |x_i - y_i|$$

where $x =$ the percentage of one group (e.g., women or non-whites) in the ith category of a particular occupation, and $y =$ the percentage of the other group (e.g., men or whites) in that same category (Duncan and Duncan 1955). The index ranges from 0, indicating complete integration, to 100, indicating complete segregation. The value of the index for segregation be gender may be interpreted as the percentage of women (or men) that would have to be redistributed among occupations in order for there to be complete equality of the occupational distribution by gender. The value of the index for segregation by race may be interpreted as the percentage of non-whites (or whites) that would have to be redistributed among occupations in order for there to be complete equality of the occupational distribution by race.

In 1981, in the USA, the index of occupational segregation by race, computed over the eleven major census occupational categories, was 24 for men (comparing white men to nonwhite men) and 17 for women (comparing white women to non-white women) (Reskin and Hartmann 1986). These values reflect a considerable decline that took place during the post World War II period; in 1940 the index for the same categories was 43 for men and 62 for women (Treiman and Terrell 1975).

It is generally agreed that in the USA the index of segregation by gender changed little between 1900 and 1960 although the changes in occupational categories over a sixty-year period make such comparisons difficult to interpret. Between 1940 and 1981, across the 11 major occupational categories, the segregation index by gender fell only slightly for whites (from 46 to 41), though somewhat more for blacks (from 58 to 39) (Treiman and Terrell 1975; Reskin and Hartmann 1986). The persistence of segregation by gender is seen as surprising in light of the marked increase in women's labour force participation rate in the post-World War II period, from 25.8 per cent in 1940 to 52.2 per cent in 1981. (The labour force participation rate for men was 79.1 per cent in 1940 and 77.4 per cent in 1981.)

The magnitude of the segregation index depends in part upon the degree of aggregation of the occupations: the greater the detailed specification of the occupations, the greater the level of measured segregation. For example, in 1980, although the occupational category 'professional' was gender-neutral – women were about one-half

of all professionals – they were not distributed equally across the professions; about one-half of all women professionals were in two occupations, nursing and non-college teaching. For 1981, Jacobs (1983) reported the segregation index by gender at 40.0 when calculated across 10 major occupational categories, 62.7 across 426 occupational categories and 69.6 across 10,000+ categories. Bielby and Baron (1984), in their study of approximately 400 establishments in California, found that in more than 50 per cent of the establishments occupations were completely segregated by gender and that only 20 per cent of the establishments had segregation indices lower than 90.

The gender segregation index declined by about 10 per cent during the 1970s (Beller 1984; Jacobs 1983), mostly as a result of greater integration of occupations rather than through changes in the size of predominantly male or predominantly female occupations. The decline during the 1970s was greatest among those with more than 17 years of education, and among those 25–34 (Jacobs 1983).

Another common way of looking at occupational segregation is to array occupations according to their percentage of female incumbents and then calculate the percentage of the female work force employed in predominantly female occupations. In 1980, about half of all employed women were in occupations that were at least 80 per cent female, while about 70 per cent of all men worked in occupations that were at least 80 per cent male (Reskin and Hartmann 1986). For black women and men these proportions were somewhat lower (Malveaux 1982).

Occupational segregation produces several deleterious effects. To the extent that it inhibits men and women from working in jobs that match their talents and skills and instead employs them in occupations that match societal stereotypes, occupational segregation lessens both individual satisfaction and potential economic output. In addition, occupational segregation contributes to the earnings differential between women and men. In 1970, women who worked full-time, year-round, earned approximately 60 per cent of the earnings of men who worked fulltime, year-round. Based on an analysis of 499 detailed occupational categories, Treiman and Hartmann (1981) concluded that about 35–40 per cent of this earnings differential was the result of occupational segregation. (The other 60–65 per cent came from the fact that within occupations men tend to earn more than women.) Gender segregation within occupations, gender segregation by firms and job segregation within firms also contribute to the female/male earnings differential (Reskin and Hartmann 1986). Finally, occupational segregation affects gender differences in occupational prestige and mobility as well as access to on-the-job training, job stress and vulnerability to lay-off and unemployment.

Theories to explain the existence and persistence of occupational segregation are remarkably divergent. Some sociological and psychological theories suggest that women's own behaviour – their values, aspirations, attitudes, and sex-role expectations – are the cause of occupational segregation. Similarly, human capital theory views women's choices about their educational attainment and interrupted work histories as responsible for their occupational designations and low pay rates. Other sociological theories, as well as economic theories of discrimination locate the employer, often aided and abetted by pressure from customers and/or employees or unions, as the source of occupational segregation. Although the world view of dual-labour-market or internal-labour-market theories is much less oriented toward individual choice and market processes than is neoclassical economics, these theories, too, locate the source of occupational segregation in employer behaviour. (For reviews of all of these theories see Reskin 1984 and Reskin and Hartmann 1986). Hartmann (1976) and Strober (1984; Strober and Arnold 1986) have pointed out that in the context of the societal-wide sex-gender system, employers, male employees and female employees *all* play a role in initiating and maintaining occupational segregation.

During the 1960s and 1970s, at both the Federal and State levels, several laws and Executive orders were designed to reduce occupational segregation in employment and in education and training programmes. The laws and orders were enforced with varying degrees of stringency;

indeed, in some cases, enforcement agencies lacked sufficient enforcement powers, funding and personnel. It appears that where the laws were enforced they were effective, although unevenly so: 'In general … positive effects occurred most often for black men, somewhat less so for black women, and were least evident for white women' (Reskin and Hartmann 1986, p. 96).

It may be, however, that occupations that become integrated by gender remain so for only a brief and transient period. Bank-telling, secretarial work and teaching are all examples of formerly all-male occupations that have been resegregated as women's occupations, with concomitant losses in relative earnings and opportunities for upward mobility (Strober and Arnold 1986; Davies 1975, 1982; Tyack and Strober 1981).

See Also

- ▶ Discrimination
- ▶ Gender
- ▶ Labour Market Discrimination

Bibliography

Beller, A.H. 1984. Trends in occupational segregation by sex, 1960–1981. In ed. Reskin.

Bielby, W.T. and Baron, J.N. 1984. A woman's place is with other women: Sex segregation within organizations. In ed. Reskin.

Davies, M.W. 1975. *Women's place is at the typewriter: Office work and office workers, 1870–1930*. Philadelphia: Temple University Press.

Davies, M.W. 1982. Women's place is at the typewriter; the feminization of the clerical labor force. In *Labor market segmentation*, ed. R. Edwards, M. Reich, and D. Gordon. Lexington: D.C. Heath.

Duncan, O.D., and B. Duncan. 1955. A methodological analysis of segregation indexes. *American Sociological Review* 20(2): 210–217.

Hartmann, H.I. 1976. Capitalism, patriarchy and job segregation by sex. *Signs* 1(3): 137–169. Pt II.

Jacobs, J.A. 1983. *The sex segregation of occupations and the career patterns of women*. Ann Arbor: University Microfilm International.

Malveaux, J. 1982. Recent trends in occupational segregation by race and sex. Paper presented at the workshop on job segregation by sex, committee on women's employment and related social issues. Washington, DC: National Research Council.

Reskin, B.F. (ed.). 1984. *Sex segregation in the workplace: trends, explanations, remedies*. Washington: National Academy Press.

Reskin, B.F., and H.I. Hartmann (eds.). 1986. *Women's work, men's work: Sex segregation on the job*. Washington: National Academy Press.

Strober, M.H. 1984. Toward a theory of occupational segregation: The case of public school teaching. In ed. Reskin.

Strober, M.H. and Arnold, C. 1986. The dynamics of occupational segregation by gender: Bank tellers (1950–1980). Stanford University.

Treiman, D.J. and Hartmann, H.I. (eds) 1981. *Women, work, and wages: Equal pay for jobs of equal value*. Report of the committee on occupational classification and analysis. Washington: National Academy Press.

Treiman, D.J., and J. Terrell. 1975. Sex and the process of status attainment: A comparison of working women and men. *American Sociological Review* 40(2): 174–200.

Tyack, D.B., and M.H. Strober. 1981. Jobs and gender: A history of the structuring of educational employment by gender. In *Educational policy and management: Sex differentials*, ed. P. Schmuck and W.W. Charters. New York: Academic Press.

Offer

John Eatwell

The term 'offer' has typically been used to refer to offers for sale from given stocks. The flow of commodities from a production process is typically labelled 'supply'. The distinction is not purely semantic, but bears upon differences in price determination – as between the attainment of equilibrium in the market for a stock, and the balance of supply and demand in production.

Pedagogical exposition of neoclassical theory often proceeds from the analysis of pure exchange, to the analysis of exchange and production by means of 'original', non-producible factor services, to exchange and production which includes original factors and producible means of production. Such, for example, is the structure of Walras's *Elements of Pure Economics* and volume 1 of Wicksell's *Lectures on Political Economy.* The

rationale for this procedure is that the essence of the theory – the resolution of individual attempts in a competitive economy to maximise utility subject to constraints – is most easily developed in the context of pure exchange, in which the constraints consist only of the stocks of commodities to be exchanged, and then the *same* principles may be extended to more complex scenarios.

Yet the transposition of the analysis from pure exchange to production involves the incorporation of two *rather different* modes of price formation within the model. The rentals paid for the factor services will be determined, as in pure exchange, by the balance of offer and demand. The prices of produced commodities will be equal to their costs of production.

Since, in equilibrium, the prices of produced commodities are equal to the sum of the rentals paid for the factor services used in their production, and the demand for those factor services is derived from the demand for products, the essential relation between utility maximization and the constraint of endowment, so evident in the case of pure exchange, is replicated in production as an indirect relation between utility maximization and fixed endowments of factors services. Or, to put it another way, as a relation between offers of factor services and the demands for them as mediated through demands for and supplies of products.

Production may therefore be considered a process of indirect exchange, in which *offers* of factor services are exchanged for one another, embodied in the form of produced commodities. This is the rationale behind Walras's argument (1874–77, p. 143) that:

> The exchange of two commodities for each other in a perfectly competitive market is an operation by which all holders of either one, or of both, of the two commodities can obtain the greatest possible satisfaction of their wants . . . . The main object of the theory of production of social wealth is to show how the principle of organization of agriculture, industry and commerce can be deduced as a logical consequence of [this] proposition.

The primacy of the balance of offer and demand within the logic of the theory also underpins Wicksteed's famous assertion (1914) that the supply curve does not exist. What are drawn as supply curves are simply the reverse of the individuals' demand to retain (not to offer) their original endowments. So the balance of offer and demand in exchange may be equally well represented as a balance of the sum of market demand and 'own-demand' with the fixed quantity of given endowment. And the balance of supply and demand in the market for products may be represented as a relation between the sum of market (derived) demands and own-demands for factor services and the fixed stocks of those services.

See Also

► Walras, Léon (1834–1910)

Bibliography

Walras, L. 1874–7. *Eléments d'économie politique pure.* Trans. as *Elements of Pure Economics,* ed. W. Jaffé. Homewood: Irwin, 1954.

Wicksell, K. 1901. *Lectures on political economy,* vol. 1, ed. L. Robbins. London: Routledge & Kegan Paul, 1931.

Wicksteed, P.H. 1914. The scope and method of political economy. *Economic Journal* 24: 1–23.

Offer Curve or Reciprocal Demand Curve

Harvey Gram

Keywords

Adjustment process; Community indifference curves; Continuity; Convexity; Duality; Excess demand; Excess supply; External economies; Homogeneous programming; Income effects; Increasing returns; Marshall, A.; Mill. J. S.; Monotonicity; Offer curve; Reciprocal demand curve; Recontracting; Stability of equilibrium; Substitution effects; Tatonnement; Trading curve; Uniqueness of equilibrium; Walras's Law

JEL Classifications
D0

The offer curve made its first appearance in Alfred Marshall's *Pure Theory of Foreign Trade* (1879), a privately printed paper consisting of the second and third chapters (chosen by Henry Sidgwick) of a four chapter manuscript. Almost 50 years passed before Marshall's analysis became generally available under his own name as Appendix J to *Money, Credit and Commerce* (1923). Thus, it was mainly through the writings of Edgeworth (1894) and others who had read Marshall's original contribution (see Whitaker 1975, p. 114n), that the offer curve came to be known.

Newman (1965, p. 104) notes the objections raised by Edgeworth (1924) and Wicksell (1925) to the name *offer curve* which was coined by W.E. Johnson (1913) and used by Bowley (1924). They were concerned that *offer curve* might suggest an asymmetry between supply and demand where, in fact, there was none. The alternative name, *reciprocal demand curve,* or *trading curve* (Newman 1965, pp. 89 ff.), avoids any such suggestion.

Marshall commented that his 'International Trade curves . . . were set to a definite tune, that called by [John Stuart] Mill' (Pigou 1925, p. 451). It was Mill who had written that

> supply and demand are but another expression for reciprocal demand; and to say that value will adjust itself so as to equalize demand with supply, is in fact to say that it will adjust itself so as to equalize demand on one side with the demand of the other. (Mill 1852, p. 604)

Mill's purpose was to close Ricardo's trade model by finding prices such that 'demand will be exactly sufficient to carry off the supply' (Mill 1844, p. 238). Edgeworth, though giving high praise to Mill's mature statement of his equation of international demand, thought little of Mill's exact solution; and Marshall commented only 'that the special example which [Mill] has chosen does not illustrate the general problem in question' (Whitaker 1975, p. 148). Chipman has argued, however, that Mill, in effect, solved a problem in what would now be called *homogeneous programming.* In claiming Mill's result to be a 'genuine and correct proof of the existence of equilibrium . . . [pre-dating the next such proof] by eighty years', Chipman remarks of Mill's law of international value, or *reciprocal demand,* that in 'its astonishing simplicity, it must stand as one of the great achievements of the human intellect' (Chipman 1965, Part 1, pp. 491 and 486, respectively).

Modern uses of the offer curve in trade theory and other areas (see Cass et al. 1980; Cass 1980; Grandmont 1985) have a greater affinity with Mill's analysis of a general equilibrium of supply and demand than with Marshall's original argument. That argument had three parts. The first is directly relevant to modern theory and concerns what would now be called the *income and substitution effects* of relative price changes. The second part deals with *increasing returns in production,* a phenomenon whose formulation and implications for traditional theory remain controversial. And finally, there is the problem of the *adjustment mechanism,* a part of Marshall's theory which, though highly regarded (Whitaker 1975, p. 115; Kemp 1964, p. 60), has been almost completely eclipsed by a Walrasian inspired 'stability' analysis. What follows is accordingly divided into three sections, following a brief discussion of the formal basis for the offer curve as it exists in modern theory.

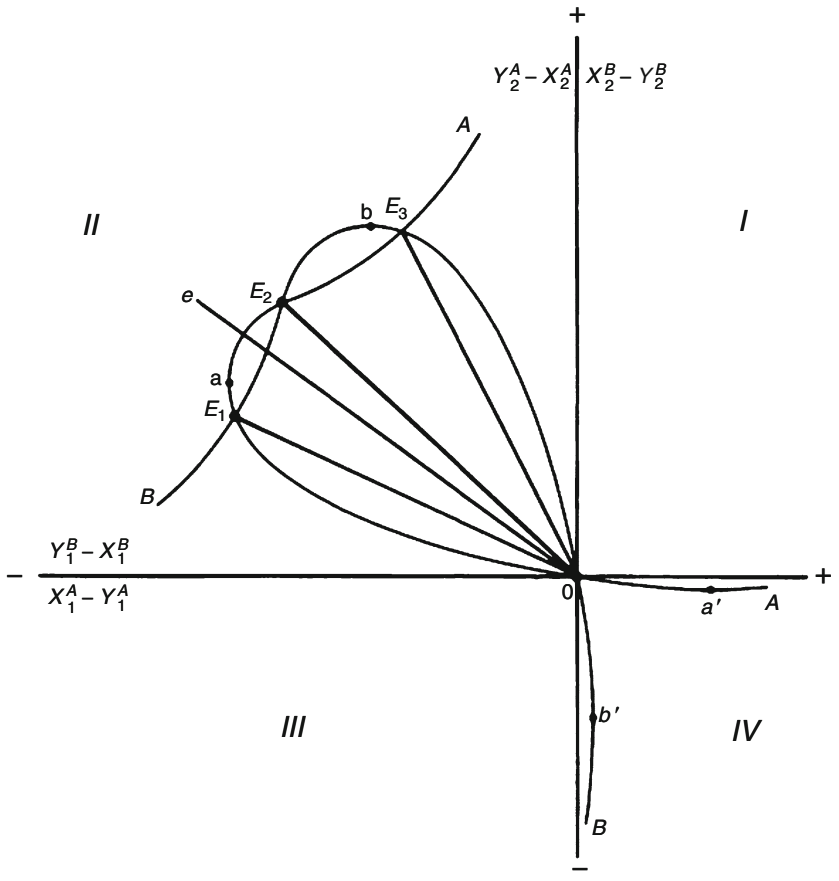
General Equilibrium

The traditional offer curve arises in the context of a two-country general equilibrium model. Each country has an endowment of resources and a technology for transforming the associated factor service flows into flows of output of two tradable commodities. Resources are owned by the country's consumers, each of whom has a preference ordering which is continuous, convex, and monotonic. Under constant or decreasing returns to scale, resources and technology generate a *convex* production possibilities set (although some degree of increasing returns to scale is not inconsistent with convexity). The assumptions on preferences guarantee that the set of points ranked 'at least as good as' any given point is also *convex,* its boundary defining an indifference curve. If

consumers have identical preferences and factor endowments, community indifference curves are simply radial expansions of individual indifference curves (cf. Chipman 1965, part 2, pp. 690–8).

Geometrical derivations of the offer curve utilize techniques introduced by Leontief (1933), Lerner (1932, 1934), and Meade (1952). Implicit in the derivation is the solution to a pair of problems in constrained optimization. At given commodity prices, P_j, $j = 1,2$, outputs, Y_j^k, in each country k, $k = A, B$, are such that the value of production, $\Sigma_j P_j Y_j^k$ is a maximum, subject to resource constraints which define the production possibilities sets. Simultaneously, for given factor supplies, F_i^k, $i = 1, 2$ (assuming two factors for simplicity), rental rates or factor prices, W_i^k, are such that cost of production, $\Sigma_i W_i^k K_i^k$, is a *minimum,* subject to price constraints which state that equilibrium profits are nowhere positive. Duality theory establishes an equality between maximum value and minimum cost. Because consumers own all resources, total cost is equal to total income, and so consumption choices, X_j^k satisfy Walras's Law: $\Sigma_j P_j X_j^k = \Sigma_j P_j Y_j^k$.

Given P_1 and P_2, there may or may not exist a solution or set of solutions, $\left(Y_1^k - X_1^k, Y_2^k - X_2^k\right)$, $k = A, B$. If $Y_1^k - X_1^k$ is positive (negative), Walras's Law ensures that $Y_2^k - X_2^k$ is negative (positive): country k offers an excess supply of good 1 (good 2) in order to satisfy its excess demand for good 2 (good 1). If both prices are positive, $Y_1^k - X_1^k = 0$ implies $Y_2^k - X_2^k = 0$; while if one price is zero and satiation is ruled out, the corresponding excess demand will be unbounded. In Fig. 1, the offer curves therefore



Offer Curve or Reciprocal Demand Curve, Fig. 1

O

occupy quadrants II and IV, passing through the origin and approaching the axes asymptotically.

At a given price ratio, measured by the (absolute) slope of a straight line through the origin, the solution for excess supply and excess demand in each country is unique in Fig. 1 and therefore (trivially) convex-valued. The solutions are also upper semicontinuous (Chipman 1965, part 2, p. 717). These two conditions are the basis for the idea of 'connectedness' or 'continuity' of the offer curve. They follow from the postulates on preferences: continuity, convexity, and monotonicity (where the last can be replaced by the assumption that outputs are strictly positive). The importance of the postulates turns on the fact that when the set of offers by each country, at a given price vector, is closed, convex, and upper semicontinuous, it can be shown that an equilibrium price vector exists. Mill found a *unique* equilibrium for Ricardo's trade model by assuming (implicitly) unitary price and income elasticities of demand for the two commodities in each country (Chipman 1965, part 1, pp. 483–91). The offer curves in Fig. 1 intersect three times, indicating *three isolated* equilibrium price vectors, OE_1, OE_2, and OE_3. Perpendiculars from E_1, E_2, and E_3 to the axes mark off the matching reciprocal demands of each country.

Income and Substitution Effects

The shape of each curve in Fig. 1 reflects the income and substitution effects of relative price changes. Consider country *A*. When P_1/P_2 is zero output of good 1 is zero and demand is unbounded so that the offer curve shoots off to the right in quadrant IV. As P_1/P_2 increases, convexity of the production possibilities set ensures that Y_1^A increases and Y_2^A decreases (unless both remain constant at a vertex). Assuming hypothetically that country *A* is confined to a given, convex-to-the-origin community indifference curve, X_1^A decreases and X_2^A increases (unless both remain constant at a 'corner' of the curve). These two substitution effects, one in production and one in consumption, reduce excess demand for good 1 (and excess supply of good 2) in quadrant IV, while raising excess supply of good 1 (and excess demand for good 2) in quadrant II. Note, however, that excess supply of good 1 reaches a maximum at *a* in quadrant II, while the same is true for good 2 at a' in quadrant IV. The reason for this is the income effect of the relative price change. In quadrant IV, a higher P_1/P_2 reduces the real purchasing power of country *A* which is an importer of good 1. If both goods are normal, the reduction in X_1^A associated with substitution is reinforced while the increase in X_2^A is offset. In quadrant II, the reverse is true. Country *A*, as an exporter of good 1, gains from an increase in the relative price of good 1. Now the income effect is pushing against the substitution effect in determining X_1^A and reinforcing it in determining X_2^A. Along the offer curve between *a* and a' substitution effects in production and consumption dominate the income effect. Beyond those critical points, the income effect is dominant in the sense that the excess supply of a commodity is lower when its relative price is higher.

Marshall's explanation of the critical point *a* is somewhat different. The independent variable in his analysis is the quantity of imports rather than the relative price ratio. In Marshall's normal class, an increase in imports in the neighbourhood of the origin results in an increase in receipts and, for this reason, the volume of exports which can be produced at normal profits increases. Receipts from imports pay the cost of exports. The slope of the offer curve increases (in absolute value) from the origin to point *a* because demand for imports is elastic. Beyond point *a* import demand turns inelastic, receipts fall off, and so the volume of exports which can be produced at normal profits declines. Marshall referred to this situation as Class I.

A final aspect of the income effect of relative price changes concerns changes in the distribution of purchasing power among consumers within each country, a problem which can only be addressed if consumers have different resource endowments. Assume therefore that country *A* and country *B* in Fig. 1 are, in fact, two groups of consumers within a single country. Aggregate excess demand for good 1 and excess supply of good 2 are *positive* for price vectors flatter than

OE_1 and for vectors intermediate between OE_2 and OE_3, and *negative* for vectors intermediate between OE_1 and OE_2 and steeper than OE_3. An offer curve defined for the two groups of consumers would therefore pass through the origin three times, tying itself in a bow. Its slope at the origin has three isolated values given by the slopes of OE_1, OE_2, and OE_3 (cf. Johnson 1959, 1960). A pair of such curves, constructed for two countries, each composed of two differentiated groups of consumers, would intersect at various points in quadrants II and IV. This indicates the possibility of trade pattern reversals as the relative price ratio takes on different equilibrium values.

The one proposition that Marshall insisted upon as 'the only law to which the curves must conform under all circumstances' is violated by offer curves which form loops through the origin. This was his Proposition VI to the effect that country *A*'s offer curve 'cannot in any case be cut more than once by a horizontal line. Similarly [country B's curve] cannot in any case be cut more than once by a vertical line' (Whitaker 1975, p. 140). Marshall's argument, however, had nothing to do with the income redistribution effects which, upon aggregation of consumer groups (as above), countries (see Chipman 1965, part 2, p. 217), or generations (see Cass et al. (1980) pp. 25–6), can result in offer curves exhibiting the floral patterns first noted by Johnson (1959, 1960). Rather he was concerned with the problem of increasing returns.

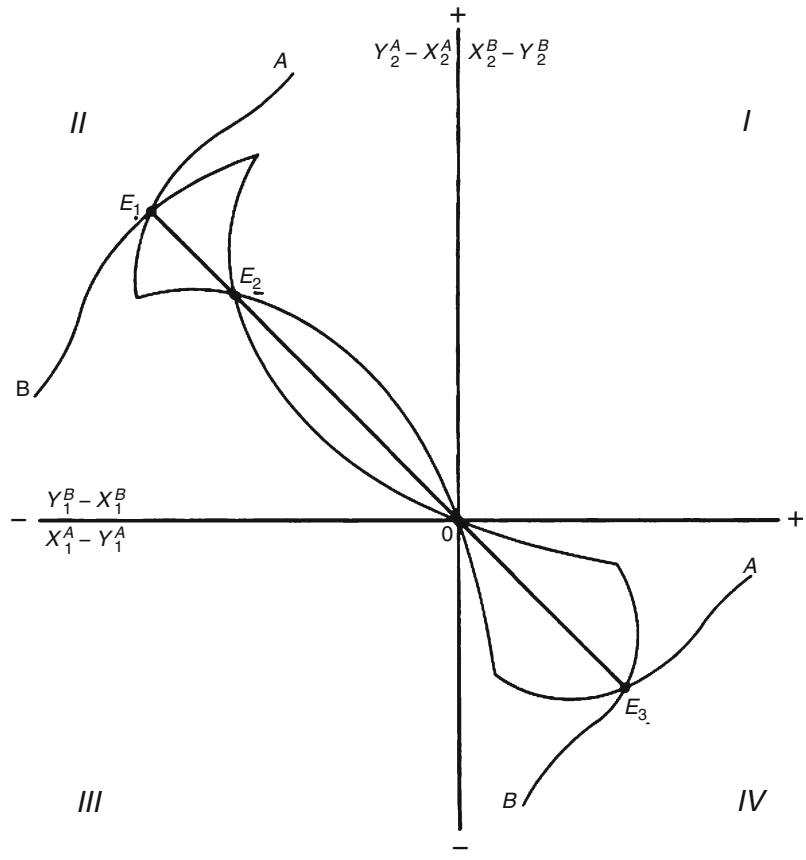
Increasing Returns

Marshall put increasing returns under the heading of 'problems of Exceptional Class II' (Whitaker 1975, p. 144). Where an increase in the production of exports leads to the introduction of extensive economies, a reduction in the volume of exports to a level previously experienced would not require as large a volume of imports to cover their costs of production as had previously been the case (assuming implicitly an elastic demand for imports). Thus, a movement along the offer curve would simultaneously shift the curve *towards* the export axis. Moreover, 'if time was allowed for the development of economies of production on a large scale, time ought to be allowed for the general increase of demand' (Pigou 1925, p. 49). In that event, a given volume of imports would yield higher receipts thereby shifting the offer curve *away* from the import axis.

Marshall's long period offer curve does not show any maximum level of exports, as does every static curve constructed on the basis of given resources and technology. Moreover, the slope of the long period curve can *decrease* (in absolute value) as a consequence of technological change. It was in this context that Marshall denied that any given volume of imports would cover the costs of more than one volume of exports. His Proposition VI claimed that economies of scale would never be sufficient to lower the total cost of a larger volume of exports below that of a smaller volume previously produced. Marshall had made the same assumption in the first edition of his *Principles,* but dropped it subsequently (Whitaker 1975, p. 116).

Modern discussions of the offer curve in the presence of increasing returns are concerned with technological externalities rather than with irreversible economies of large-scale production. A firm's output may depend on the output of the industry to which it belongs. Output in one industry, or the level of employment of particular factors in that industry, may have external effects on the output of another industry. Theoretical questions then arise concerning the convexity of the production possibilities set, the relationship between opportunity cost and relative price, and whether or not production occurs at a limit point of the feasible set. What the models have is common with Marshall's discussion is that the associated offer curves are no longer convex-valued functions of the relative price ratio. Marshall indicated this by drawing offer curves with several inflexion points. In modern treatments of external economies in production, offer curves typically have the shape indicated in Fig. 2. Curvature at the origin is opposite to that indicated in Fig. 1, changing abruptly at points of complete specialization. (The latter may or may not correspond to the critical points in Fig. 1 where excess supply reaches a maximum.)

Offer Curve or Reciprocal Demand Curve, Fig. 2



The curves in Fig. 2 have three intersections indicating three equilibrium trades. (These may be reduced to two by drawing curves which are mutually tangent at the origin indicating equal pre-trade price ratios which, in Fig. 1, would be sufficient to rule out an equilibrium with positive trade.) There is nothing in principle to prevent all three points from falling along a single ray. The resulting indeterminacy in the volume and direction of trade is the main distinguishing feature of trade models with external economies in production. Chacholiades (1978, pp. 197–9) has considered the problem in some detail, arguing that, in general, a country benefits from specialization in the production of the commodity subject to external economies. If this is the same commodity in each country, then, depending on the pattern of demand, one country may lose from trade. This suggests an even sharper conflict of interest than is evident in Fig. 1 where a country is clearly better off in that equilibrium in which its exports are smallest and its imports are largest.

Stability

Stability of equilibrium is defined in relationship to a process of adjustment which determines the movement of prices and/or quantities when the system is out of equilibrium. A distinction has been drawn between processes which focus on price changes and those which focus on quantity changes. A frequently considered case is that in which prices respond to differences between *hypothetical* supply and demand (those quantities which would prevail on each side of the market if the current price were an equilibrium price). Transactions, however, only take place in

equilibrium. This is a *recontracting process* and its convergence to an equilibrium of supply and demand is often referred to as the Walrasian *tâtonnement* or 'groping' process. Walras, in fact, referred to *tâtonnement* in connection with the problem of bringing a set of interrelated markets into equilibrium *sequentially,* and, as such, it was problematical since 'few prices will lie quiet at equilibrium while others are brought to heel, and the whole thing may turn out to be like the labour of Sisyphus' (Newman 1965, p. 103).

The offer curves in Fig. 1 can be used to illustrate the stability of a recontracting process. Consider a price ratio, P_1/P_2, slightly greater than the (absolute) slope of OE_1. Hypothetical exports of good 1 by country A exceed hypothetical imports by country 2, while the opposite is true for good 2. If P_1 falls and P_2 rises in this situation, P_1/P_2 falls back towards OE_1. The opposite is true for price ratios slightly lower than the (absolute) slope of OE_1. Thus, E_1 is a stable point. Note that, as prices move, the four substitution effects (in production and consumption in both countries) contribute towards reducing the initial divergence between supply and demand for each good. The same is true of part of the income effect. As the price ratio falls towards OE_1, for example, country A is made worse off and country B is made better off. As *importers,* country A demands less of good 2 and country B demands more of good 1 (assuming that the goods are normal), and this reinforces the substitution effects. But, as *exporters,* country A demands less of good 1 while country B demands more of good 2, thereby exacerbating the initial excess supply (of good 1) and excess demand (for good 2). At stable points, such as E_1 and E_3, exporters' income effects are not strong enough to swamp importer's income effects plus all substitution effects. At E_2, however, a slight increase in the relative price of good 1 would be associated with a hypothetical excess demand for good 1 and excess supply of good 2. The recontracting hypothesis would therefore result in a further increase in P_1/P_2, reflecting the fact that the initial increase has generated exporters' income effects which swamp all other income and substitution effects (cf. Caves and Jones 1985, pp. 492–4).

A variation on the above analysis allows trade to take place out of equilibrium but assumes that demand for imports is always satisfied. A disequilibrium exchange ray, such as Oe in Fig. 1, cuts the two offer curves in distinct points. Perpendiculars to the axes from these points indicate excess supply of good 1, which causes inventories to rise, and excess demand for good 2, which causes inventories to fall. If P_1 then falls and P_2 rises, P_1/P_2 once again falls back toward OE_1. During the process, however, country A must be selling assets to country B in order for the trade flow to be financed. If the consequence of this is to alter the position and shape of the offer curves, a more complete and undoubtedly more complex analysis of the convergence to equilibrium would be required (cf. Jones 1961, p. 203). Marshall's discussion of the adjustment mechanism is concerned neither with a recontracting process nor with inconsistent trades 'financed' by changes in inventories. In this theory, profits in export industries are abnormally high at points between a country's offer curve and the axis measuring its imports. On the other side of the curve, profits in exports are abnormally low. Marshall's adjustment mechanism is summed up as follows:

> when the terms on which a country's foreign trade is conducted are such as to afford a rate of profits higher than the rate current in other industries, the competition of traders to obtain these higher profits will lead to an increase in the exportation of her wares: and *vice versa* when the rate of profits in the foreign trade [is] exceptionally low. (Whitaker 1975, p. 151)

This adjustment in the production of exports (imports and the domestic consumption of exports held constant) appears to have been meant by Marshall to reflect a concomitant change in the production of non-traded goods (Marshall 1923, pp. 354–5n). Thus, at points off the offer curves

> production is changing in both countries, [and so] the dimensions of the Edgeworth box must be changing, as are also the shapes of the offer curves. The extreme subtlety of the Marshallian conception becomes more apparent the further one probes into it. (Chipman 1965, Part 2, p. 723)

One can only conclude that efforts to formalize Marshall's 'dynamics' (Samuelson 1947,

pp. 266–8; Kemp 1964, pp. 66–9; Amano 1968, pp. 326–39) are but valiant attempts to come to terms with an approach to equilibrium which itself moves in an unspecified manner as a consequence of not being attained initially.

Conclusion

Not surprisingly, that part of Marshall's *Pure Theory of Foreign Trade* which is most evident in modern discussions of the offer curve concerns the income and substitution effects which are central to the theory of supply and demand equilibrium. His treatment of increasing returns and his discussion of the adjustment process raise dynamic considerations associated with changes in technology and with changes in the structure of productive capacity. It is just such changes which present the equilibrium theory of supply and demand with some of its greatest difficulties.

See Also

- ▶ Marshall, Alfred (1842–1924)
- ▶ Terms of Trade

Bibliography

Amano, A. 1968. Stability conditions in the pure theory of international trade – Rehabilitation of the Marshallian approach. *Quarterly Journal of Economics* 82: 326–339.

Bowley, A. 1924. *The mathematical groundwork of economics*. New York: Oxford University Press.

Cass, D. 1980. Money in consumption loan type models: An addendum. In *Models of monetary economics*, ed. J.H. Kareken and N. Wallace. Minneapolis: Federal Reserve Bank of Minneapolis.

Cass, D., M. Okuno, and I. Zilcha. 1980. The role of money in supporting the Pareto optimality of competitive equilibrium in consumption loan type models. In *Models of monetary economies*, ed. J.H. Kareken and N. Wallace, 13–48. Minneapolis: Federal Reserve Bank of Minneapolis.

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

Chacholiades, M. 1978. *International trade theory and policy*. New York: McGraw-Hill.

Chipman, J.S. 1965. A survey of the theory of international trade. *Econometrica* 33, Pt. I, 477–519; Pt. II, 685–760.

Edgeworth, F.Y. 1894. The theory of international values. *Economic Journal* 4: 35–50, 424–443, 606–638. Reprinted in F.Y. Edgeworth, *Papers relating to political economy* [1925], vol. 2. New York: Burt Franklin, 1970.

Edgeworth, F.Y. 1924. Review. *Economic Journal* 34: 430.

Grandmont, J.-M. 1985. On endogenous competitive business cycles. *Econometrica* 53: 995–1045.

Johnson, H.G. 1959. International trade, income distribution, and the offer curve. *Manchester School of Economic and Social Studies* 27: 241–260.

Johnson, H.G. 1960. Income distribution, the offer curve, and the effects of tariffs. *Manchester School of Economic and Social Studies* 8: 215–242.

Johnson, W.E. 1913. The pure theory of utility curves. *Economic Journal* 23: 483–513.

Jones, R.W. 1961. Stability conditions in international trade: A general equilibrium analysis. *International Economic Review* 2: 199–209.

Kemp, M.C. 1964. *The pure theory of international trade*. Englewood Cliffs: Prentice-Hall.

Leontief, W.W. 1933. The use of indifference curves in the analysis of foreign trade. *Quarterly Journal of Economics* 47: 493–503.

Lerner, A.P. 1932. The diagrammatical representation of cost conditions in international trade. *Economica* 12: 346–356.

Lerner, A.P. 1934. The diagrammatical representation of demand conditions in international trade. *Economica,* NS 1: 319–334.

Marshall, A. 1879. *The pure theory of foreign trade.* Reprinted in Whitaker (1975).

Marshall, A. 1923. *Money, credit and commerce*. New York: Augustus Kelley, 1965.

Meade, J.E. 1952. *A geometry of international trade.* Reprinted. New York: Augustus Kelley, 1971.

Mill, J.S. 1844. On the laws of interchange between nations; and the distribution of the gains of commerce among countries of the commercial world. In *Essays on some unsettled questions of political economy.* Reprinted in *Collected works of John Stuart Mill,* vol. 4. Toronto: University of Toronto Press, 1967.

Mill, J.S. 1852. *Principles of political economy.* 3rd ed. Reprinted in *Collected works of John Stuart Mill,* vol. 3. Toronto: University of Toronto Press, 1965.

Newman, P. 1965. *The theory of exchange*. Englewood Cliffs: Prentice-Hall.

Pigou, A.C. (ed.). 1925. *Memorials of Alfred Marshall*. London: Macmillan.

Samuelson, P.A. 1947. *Foundations of economic analysis*. Cambridge, MA: Harvard University Press.

Whitaker, J.K. (ed.). 1975. *The early writings of Alfred Marshall, 1867–1890*, vol. 2. New York: Free Press.

Wicksell, K. 1925. Matematisk nationalekonomi. *Ekonomisk Tidskrift* 27: 103–125. Trans. in *Knut Wicksell, Selected Papers on Economic Theory,* ed. E. Lindahl. Reprinted, New York: Augustus Kelley, 1969.

Ohlin, Bertil Gotthard (1899–1979)

Hans Brems

Keywords

Accelerator; Adaptive expectations; Bonds; Budget deficits; Cassel, G.; Factor endowments; Factor price equalization; Factor substitution; Heckscher–Ohlin trade theorem; Hecksher, F. H.; Input–output coefficients; International capital flows; International trade theory; Interregional trade; Keynesianism; Lindahl, E. R.; Liquidity preference; Money supply; Multiplier; Natural rate and market rate of interest; Ohlin, B. G.; Preferences; Propensity to consume; Public works; Stockholm School; Stolper–Samuelson theorem; Technology; Wicksell, J. G. K.

JEL Classifications

B31

Ohlin was born on 23 April 1899 in Klippan, Sweden. He took a degree in mathematics, statistics and economics at the University of Lund in 1917, a degree in economics under Heckscher at the Stockholm School of Business Administration in 1919, an AM degree under Taussig and Williams at Harvard in 1923, and a Ph.D. degree under Cassel at the University of Stockholm in 1924. Ohlin taught at the University of Copenhagen (1925–30) and, as Heckscher's successor, at the Stockholm School of Business Administration (1930–65). He was a visiting professor at the University of California at Berkeley in 1937 and at Columbia and Oxford in 1947.

For the League of Nations Ohlin prepared a report on the world depression in 1931 and for the Swedish government a report on unemployment in 1934. He was a member of the Swedish parliament (1938–70), a member of the Cabinet (1944–45), the leader of the Liberal Party (1944–67); he died on 3 August 1979 in Stockholm.

Trade Theory

Ohlin is best known for, and received the 1977 Nobel Prize for, his modernization of the theory of international trade. The modernization was long overdue: discredited in general economic theory after 1870, the labour theory of value was still surviving in the province of international-trade theory half a century later.

Ohlin's teacher at Stockholm was Gustav Cassel, and his point of departure was Cassel's (1918) version of a Walrasian general equilibrium of a closed economy with perfect mobility of goods and factors. Unlike Walras, Cassel assumed the factor endowments of all households to be fixed. Household income would then be the sum of the products of factor price and all factor endowments of that household. Like Walras, Cassel assumed the input–output coefficients of all goods to be fixed. The competitive price of a good would then be the sum of the products of factor price and all input–output coefficients of that good. Facing such household income and such competitive goods prices, every household would reveal its preference. Goods–market equilibrium would require industry supply and such household demand to be equal for every good. Industry demand for a factor would be the sum of the products of such industry goods supplies and all input–output coefficients of that factor. Factor-market equilibrium would require household supply and such industry demand to be equal for every factor.

The ultimate determinants of all quantities and relative prices in such a general equilibrium were, first, factor endowments; second, technology in the form of the input–output coefficients; and, third, preferences. Inspired by his other teacher at Stockholm, Eli Filip Heckscher (1919), Ohlin (1924, 1933) set out to modify the Cassel model to fit interregional and international trade.

As his first modification Ohlin visualized an economy composed of regions within which factor mobility was perfect but between which it was imperfect or, as a first approximation, nonexistent. In the absence of goods trade, isolation would be complete, and such regions would simply constitute a system of miniature Casselian

closed economies. Between them relative prices could differ because factor endowments, technology, or preferences differed. As another first approximation, Ohlin assumed regions to differ solely in their factor endowments, not in their technology or preferences. Finally, Ohlin unfroze Cassel's fixed input–output coefficients, thus making room for factor substitution. With such assumptions he had the ingredients to what later became known as the 'strong' Heckscher–Ohlin theorem. In the simple case of two factors, two goods and two regions the theorem becomes very tractable. In isolation each region would have a relatively low-priced and a relatively high-priced good. Since nothing else than factor endowments differed between regions, the low-priced good would be low-priced because it required relatively much of that region's relatively abundant, hence low-priced, factor. That good will be a candidate for export once we remove isolation. The high-priced good would be high-priced because it required relatively much of that region's relatively scarce, hence high-priced, factor. That good will be a candidate for import once we remove isolation; but we are not removing it yet. As we know, under profit maximization, pure competition, and factor substitution the physical marginal productivity of either factor in terms of either good will equal the real price of that factor in terms of that good.

Now remove isolation and let goods be traded. Export would expand a region's demand for its abundant factor and import reduce the demand for its scarce factor. Thus trade would raise the price of the abundant factor, reduce the price of the scarce one, and encourage substitution between them: either good would use less abundant factor per unit of scarce factor than in isolation. The abundant factor would then have a higher physical marginal productivity and a higher real price in terms of either good than in isolation. Vice versa for the scarce factor. Does all this mean that trade would eventually equalize real factor prices in terms of either good between regions – although no factor ever crossed the border? Yes, in the absence of transportation costs and in the absence of specialization. One reason for specialization would be increasing returns to scale. Specialization would leave a region with an unproduced good. Where nothing is produced, no factor can have a marginal productivity. In terms of the unproduced good, then, physical marginal productivity could no longer equal real factor price, and the theorem would fail. So it would in case of transportation costs or in case regions differed, not in factor endowments but in technology or preferences. And so it might if there were more than two factors, goods, or regions.

Few theorems have been as fruitful, that is, few inspired as much later work, theoretical and empirical, as the Heckscher–Ohlin theorem. Neither Heckscher nor Ohlin applied present-day rigour. To Heckscher factor-price equalization would be complete; to Ohlin – more aware of the many qualifications – incomplete. The theorem was first taken up, baptized, and rigourized by Stolper and Samuelson (1941), who examined a scarce factor's case for protectionism but found 'the definiteness of the Heckscher–Ohlin theorem [beginning] to fade' with more than two factors. More groundwork was done by Samuelson (1948, 1949). Using his domestic US input–output table with many goods but only two factors, Leontief (1953, 1956) found the capital–labour ratio to be lower in US exports than in US import-competing goods. If the Heckscher–Ohlin theorem were true, then, capital would have to be the scarce and labour the abundant US factor. This Leontief paradox did not make the theorem go away but stimulated new contributions. A good guide to them is the third part of Chipman's (1966) survey of the theory of international trade.

Ohlin's second modification of Cassel saw international trade as a special case of interregional trade. What was special about nations?

First, national differences in factor endowments, technology, and preferences might be rooted in differences in climate, language, cultural, and legal institutions. Of international movements of factors, labour as well as capital, and such obstacles to them Ohlin gave a full account. His account of international capital

movements found an early and specific expression (1929) in his discussion with Keynes of the mechanism of the reparation payments imposed upon Germany by the Versailles Treaty. Still influenced by Marshallian tradition, Keynes saw a drastic worsening of Germany's terms of trade as a necessary condition for such payments. To Ohlin reparations were nothing but huge international transfers of 'buying power'. Against an uncomprehending 1929 Keynes, Ohlin advocated the view of a 1936 Keynes, that is, the income mechanism would do; no price mechanism was needed.

Second, nations were special in having their own currency and monetary authorities. In a two-country world such separate currencies would add a new unknown, that is, the price of one currency in terms of the other – the exchange rate. Fortunately there would also be a new equation, that is, the equilibrium condition that in a pure-trade model the balance of trade would be zero or that in a trade *cum* lending and borrowing model the balance of payments would be zero.

Macroeconomic Theory

Less well known to the English-speaking world is Ohlin's macroeconomic theory: its most important work (1934) was never fully translated. Here, Ohlin was inspired by Wicksell and Lindahl.

Wicksell (1893) had restated Böhm-Bawerk mathematically and (1898) wondered how a Böhm-Bawerk 'natural' rate of interest was related to the rate of interest observed in markets where the supply of money met the demand for it. If such a 'money' rate of interest were lower than the natural rate of interest, entrepreneurs would be induced – and the money supply correspondingly expanded – to pay a higher money wage rate. Physically speaking, nothing would come of this, for when labour spent the higher money wage rate, prices would rise correspondingly and unexpectedly leave the real wage rate unchanged. There would be a cumulative process of inflation expected by nobody.

Wicksell's answer was made possible by a method fundamentally new in three respects. Wicksell's method was a macroeconomic, dynamic disequilibrium method based upon adaptive expectations whose disappointment constituted the motive force of the system. But Wicksell had applied his method to a model with price as the only variable. Using Wicksell's method and inspired by Lindahl's (1930) refinement of it, Ohlin (1933, 1934) added physical output as an additional variable. Two years ahead of Keynes, Ohlin used three Keynesian tools, that is, the propensity to consume, liquidity preference and the multiplier, and one non-Keynesian tool, that is, the accelerator. The four tools would interact as follows in Ohlin's feedback mechanism. Let consumption demand be stimulated. As a result physical output would rise, generating new income. The propensity to consume would link physical consumption to the *level* of physical output and thus establish a consumption feedback. The accelerator would link physical investment to the *growth* of physical output and thus establish an investment feedback. As did the Wicksellian one, Ohlin's two feedbacks unfolded in a cumulative process along a time axis as a succession of disequilibria: expectations and plans were for ever being revised in the light of new experience. By contrast, Keynes used only the consumption feedback and telescoped it into an instant static equilibrium along an output axis.

Ohlin's relation to Keynesian economics was discussed by Steiger (1976), Patinkin (1978), and Brems (1978). Forty-one years apart Ohlin expressed his own view on the matter in (1937) and (1978).

Ohlin's (1934) analysis appeared in a report on unemployment requested by the Swedish government, and his policy conclusions were quite specific. In times of excess capacity the government should undertake investment projects – say highway construction or the electrification of state railroads – which would not compete with private investment and which should be allowed to generate fiscal deficits. Tax financing would reduce consumption and thus defeat the purpose of public works. Ohlin wrote the government budget constraint: deficits might be financed by expanding

either the bond or the money supply. Sale of government bonds would depress bond prices and thus discourage private investment, again defeating the purpose of public works. That left central-bank discounting of treasury bills as the only way which would not deprive private investment of finance. Thus financed, public works would generate income. Such income generation would be magnified by the multiplier and the accelerator.

Except for a nine-page algebraic two-country Cassel general equilibrium, banished to an appendix, Ohlin used neither algebra nor diagrams. But in all his work his style was accurate, cautious and lucid, often enlivened by relevant statistical and historical illustrations.

See Also

- ▶ Heckscher–Ohlin Trade Theory
- ▶ Stockholm School

Selected Works

1924. *Handelns teori*. Stockholm: Centraltryckeriet.

1929. Transfer difficulties, real and imagined. *Economic Journal* 37: 172–178. Reprinted in *Readings in the theory of international trade*, ed. H.S. Ellis and L. Metzler. Philadelphia/Toronto: Blakiston, 1949.

1933a. *Interregional and international trade*. Cambridge, MA: Harvard University Press.

1933b. Till frågan om penningteoriens uppläggning. *Ekonomisk Tidskrift*. Trans. as 'On the formulation of monetary theory', *History of Political Economy* 10 (1978): 353–388.

1934. *Penningpolitik, offentliga arbeten, subventioner och tullar som model mot arbetslöshetbidrag till expansionens teori*. Unemployment Report II, 4. Stockholm: P.A. Norstedt & Söner. Summarized by Brems (1978).

1937. Some notes on the Stockholm theory of saving and investment. I–II. *Economic Journal* 47: 53–69; June, 53–69, 221–240. Reprinted in *Readings in business cycle theory*, ed. G. Haberler. Philadelphia/Toronto: Blakiston, 1951.

1978. Keynesian economics and the Stockholm School: A comment on Don Patinkin's paper. *Scandinavian Journal of Economics* 80(2): 144–147.

Bibliography

Brems, H. 1978. What was new in Ohlin's 1933–34 macroeconomics? *History of Political Economy* 10: 398–412.

Cassel, G. 1918. *Theoretische Sozialökonomie*. Leipzig: Winter. Trans. from the 5th ed as *The theory of social economy*, New York: Harcourt, Brace, 1932.

Chipman, J.S. 1966. A survey of the theory of international trade: Part 3, the modern theory. *Econometrica* 34: 18–76.

Heckscher, E.F. 1919. Utrikeshandelns verkan på inkomstfordelningen. *Ekonomisk Tidskrift* 497–512. Trans. as 'The effect of foreign trade on the distribution of income' in Readings in the *Theory of international trade*, ed. H.S. Ellis and L. Metzler. Philadelphia/Toronto: Blakiston, 1949.

Leontief, W.W. 1953. Domestic production and foreign trade; the American capital position re-examined. *Proceedings of the American Philosophical Society* 97: 332–349.

Leontief, W.W. 1956. Factor proportions and the structure of American trade: Further theoretical and empirical analysis. *The Review of Economics and Statistics* 38: 386–407.

Lindahl, E. 1930. *Penningpolitikens medel*. Lund: Gleerup. Partially trans. in E. Lindahl, *Studies in the theory of money and capital*. London: Allen & Unwin, 1939.

Patinkin, D. 1978. On the relation between Keynesian economics and the 'Stockholm School'. *Scandinavian Journal of Economics* 80(2): 135–143.

Samuelson, P.A. 1948. International trade and the equalization of factor prices. *Economic Journal* 58: 163–184.

Samuelson, P.A. 1949. International factor-price equalization once again. *Economic Journal* 59: 181–197.

Steiger, O. 1976. Bertil Ohlin and the origins of the Keynesian revolution. *History of Political Economy* 8: 341–366.

Stolper, W.F., and P.A. Samuelson. 1941. Protection and real wages. *Review of Economic Studies* 9: 58–73. Reprinted in *Readings in the theory of international trade*, ed. H.S. Ellis and L. Metzler. Philadelphia/Toronto: Blakiston, 1949.

Wicksell, K. 1893. *Über Wert, Kapital und Rente*. Jena: G. Fischer. Trans. as *Value, capital and rent*. London: Allen & Unwin, 1954.

Wicksell, K. 1898. *Geldzins und Güterpreise*. Jena: G. Fischer. Trans. R.F. Kahn with an introduction by Bertil Ohlin as *Interest and prices*. London: Macmillan, 1936.

Oil and Politics in the Gulf: Kuwait and Qatar

Jill Crystal

Abstract

The discovery of oil in the early 20th century had a dramatic effect on the formation and destruction of political coalitions and state institutions in the Arab states of the Persian Gulf. In particular, it fundamentally restructured the relationship between the rulers of the Gulf's sheikhdoms and the merchants, the business elite of that era, shifting political power away from the merchants and into the hands of the rulers and ruling families. In the process, oil dramatically restructured politics and economics, creating new alliances and institutions that would continue to shape politics into the 21st century. The details of this initial arrangement in Kuwait and Qatar are developed in Crystal (1995).

Keywords

Kuwait; Oil; Persian Gulf; Politics; Qatar

JEL Classifications

L710; O130; Q350

Introduction

The discovery of oil in the early 20th century had a dramatic effect on the formation and destruction of political coalitions and state institutions in the Arab states of the Persian Gulf. In particular, it fundamentally restructured the relationship between the rulers of the Gulf's sheikhdoms and the merchants, the business elite of that era, shifting political power away from the merchants and into the hands of the rulers and ruling families. In the process, oil dramatically restructured politics and economics, creating new alliances and institutions that would continue to shape politics into the 21st century. The details of this initial arrangement in Kuwait and Qatar are developed in Crystal (1995).

In the pre-oil era, the relationship between rulers and merchants in Kuwait and Qatar was similar. In both countries, politics was characterised by a coalition between the ruler and those merchant families who controlled the lucrative pearl diving industry and long-distance commodity trade and who provided rulers with the revenues (and sometimes the manpower) they needed to rule the country. Until oil was discovered in the 1930s, rulers in both countries (neither of which possessed significant non-oil resources), shared a dependence on the local merchants for revenues. In Kuwait, the merchants were also the country's main employers into the 1950s (Tetreault 2000, p. 39). This was almost certainly the case in Qatar as well, where pearling and trade dominated the economy, accompanied by seasonal nomadic pastoralism. This dependence of the rulers gave merchants economic influence and with it a degree of political influence, exercised informally through social institutions, among them *majlis* (the regular weekly meetings that allowed merchants to air their opinions and grievances with the rulers), inter-marriages between dominant merchant families and the ruler's family, and, in Kuwait's case, proximity (the families of the rulers and merchants all lived within the walls of the old city). The ruler's family, while enjoying a high social rank, did not function as a ruling family – that is, as an institution through which the ruler designed and implemented policy.

In the interwar period, however, the political economy of the Gulf changed dramatically. The merchants' position was weakened by the crash of the pearl market following the invention of Japanese cultured pearls and then by the Great Depression. At about the same time, however, oil was discovered in the Gulf. In Kuwait the Amir signed a concession agreement in 1934 with the Kuwait Oil Company (an Anglo-American consortium) to search for oil, which was discovered later in 1938. In Qatar the ruler signed a concession agreement

with the Anglo- Persian Oil Company (a predecessor of British Petroleum) in 1935; oil was also discovered there in 1938. However, in both countries the oil wells were capped during the Second World War, and it was not until the late 1940s that oil came to dominate the economy.

Oil and Politics

Oil forged a new relationship between rulers and merchants. Since they came from outside the country and went directly from the transnational oil companies to the ruler, and later to the state, oil revenues freed rulers from their historical economic (and hence political) dependence on established economic elites. Unlike rulers in other states, forced by the need for revenues to either crush economic elites or absorb them into the political process, Gulf rulers could simply buy out those elites using the newly found oil resources.

To do so they worked out a new arrangement with the merchants: a trade of wealth for access to formal power. The sheer volume of revenues prompted the merchants to re-examine their core interests and coalesce around them. With oil, the merchants – the group that had historically pressed its claims most effectively on the state – now renounced their claim to participate in formal decision-making. In exchange, the rulers guaranteed them a continuing share of oil revenues. While the energy sector remained government-owned and operated, steps were taken to ensure that a thriving private sector, dominated by the older merchants, would handle most of the rest of the economy. This was accomplished through a variety of measures. In Kuwait, one important mechanism was a government land acquisition programme which purchased land from merchants at above-market prices, then rented or sold the land back to the merchants at below-market prices (Moore 2004, p. 43). From the 1940s to the early 1970s roughly a quarter of Kuwait's oil revenues went to this programme (Khouja and Sadler 1979, pp. 44–5).

In Qatar, another government programme was the public purchase (and then manumission) of slaves, owned by wealthy families (Crystal 1995, p. 143), creating support from both the former slave owners and former slaves. In both countries, other measures included an array of pro-business policies: no taxes, free movement of capital and preferential government contracts to merchant families, as well as direct loans (initially in a period when credit was scarce) and subsidies. In the early years, the agency system was particularly beneficial to merchants in both countries. This system required foreign investors to take local partners. No product or service could be sold locally without a local agent, who received a percentage of the profit. Typically these local agents were members of the established merchant families. At first an informal requirement, agency agreements were later written into commercial laws. In the early boom years, when the infrastructures of these countries were built, great fortunes were made by men who sometimes did little more than serve as a silent local partner. The ruling family agreed (for the most part) to stay, at least visibly, out of the private sector (although this was honoured more in Kuwait than in Qatar).

The merchants, in turn, agreed to stay out of formal politics. Where, before oil, economic elites entered politics to protect their economic interests, after oil, merchants left formal politics to preserve their economic interests. Oil revenues thus preserved the apparent continuity at the top of the political system, retaining a monarchical form of government, but actually altered politics considerably by forcing the breakdown of the old ruling coalition between the rulers and merchants. Economic elites withdrew from politics, but did not disappear as a social force. The differences in the relative strengths of both the rulers and merchants in Kuwait and Qatar before oil affected the nature of the transformation. In Kuwait, the merchants retained a particularly strong sense of corporate identity, one which was reinforced by institutions of marriage and *majlis*, and hence an ability to re-enter politics should the Amir renege on his initial arrangement. Their economic autonomy was institutionalised in such bodies as the Kuwait Chamber of Commerce and Industry (KCCI), established in 1961 (Moore 2004).

In Qatar, the merchants formed a smaller, yet less cohesive group, one that was also more divided along sectarian lines. While in Kuwait,

most of the dominant merchant families were Sunni, in Qatar, some of the wealthiest families, such as the al-Fardan, Jaidah and Darwish were Shia families of Iranian origin (Kamrava 2012, p. 64). The ruling family, on the other hand, was far larger relative to the national population and far more quarrelsome than its Kuwaiti counterparts. Its differences frequently spilled into the public view and were exacerbated then (as today) by Saudi meddling. While the rulers bought the merchants out of politics, as they did in Kuwait, in the years following the discovery of oil, the Qatari merchants' strength was further diminished by the emergence of new economic elites created by and more dependent on the rulers. In Kuwait, the dominant merchant families were largely Sunni, and a similar process occurred among the Shia merchant families, who were bypassed by the Amir in favour of newer Shia business elites, more dependent on the state. These families were also used to secure the middle class Shia support (Azoulay 2013). The Qatari merchants' strength was also undercut by the more frequent direct intrusion of the Qatari ruling family itself into the ownership of businesses and real estate, both commercial and residential. A far smaller number of old business families, such as al-Mana and the Darwish, rebuilt themselves as wealthy families in the modern economy, but they did so by remaining close to the ruling family. The existence of a much larger and unruly ruling family, which each Amir struggled to control, also left Qatar less stable domestically than Kuwait.

By the end of the 20th century, some important changes in the structure of the local market had occurred. Some privately owned family businesses became publicly held corporations. Agency monopolies were eroded when Kuwait (1995) and Qatar (1996) joined the World Trade Organization (WTO). But by then the merchants had already established their domination of the local market and the business community did not voice objections to joining the WTO (Seznek 2007, p. 75). Both governments were now also anxious to attract foreign direct investment in order to create jobs for their growing and increasingly youthful national population. Merchants in both Kuwait and Qatar also became active participants in a larger regional market, expanding into other Gulf Cooperation Council states' markets and facing competition at home from businesses based in other GCC states, largely from what Hanieh (2011) calls *khaleeji* [Gulf] capital, dominated by Saudi Arabia and the United Arab Emirates. In Kuwait and Qatar, however, local businesses maintained the advantage they had always enjoyed because these were never simply about money, but embedded in larger coalitional arrangements that underpinned the basic power structure of the state. If anyone suffered from the additional competition, it was the newer merchants. This historical local advantage insulated the Qatari and Kuwait private sector somewhat from the *khaleeji* capital that Hanieh explores. Consequently, Qatar and Kuwait's private sector did not suffer as greatly from the fiscal crisis of 2008 as did the other GCC states.

In Kuwait, the merchants were a more cohesive group of wealthy families, with eight principal families at their core, who retained their social connections even in the absence of the pearl and commodity trade industries that had given them their initial wealth. Even as the economy changed, the old merchant families continued to dominate the leading non-oil sectors and influential economic institutions such as the KCCI. While new entrepreneurs also emerged in Kuwait in the following decades, they were dwarfed by the size and strength of the established business community and were not as dependent on the state. The result in Kuwait was thus a precarious balance: economic elites withdrew from formal politics, but did not disappear as a social force. In Qatar, the old families retained their connections to the rulers along with the economic benefits that followed, but failed to cohere as thoroughly as a social force.

Creating New Alliances

In both countries the rulers also formed new ties with the national population, which began to receive economic benefits directly (in the form of grants, loans, subsidised housing and utilities, and free education and healthcare) and indirectly (through the creation of a vast welfare state and

through massive state employment). All nationals also benefited from a generally tax-free system (although some fees for services were later added). A once politically active labour force was removed from politics by the importation of a new and tightly restricted working class, brought in for temporary work from other countries, initially from elsewhere in the Arab world and later from the subcontinent. The indigenous working class, once politically quite active, was transformed into a more comfortable and docile middle class, largely loyal to the rulers or, at least, politically quiescent. Foreign labour also benefitted the merchants by providing them with an inexpensive and pliant labour force. While workers were sometimes protected by labour law, in practice, any labour action would result in the deportation of the activists. These policies explain in part why the ruling families of both states were able to withstand and survive the revolutionary wave of Arab nationalism in the 1950s and 1960s and, later, the Arab Spring.

The governments' distributive policies in turn inadvertently triggered the creation of large state bureaucracies: distributive states that emerged from the imperative to expend rather than extract revenues. In the 1960s and especially the 1970s, governments in both Kuwait and Qatar were able to create massive welfare states which provided both services and state employment to the majority of its nationals. As these bureaucracies grew, they became both bloated and less susceptible to control through ruling kinship networks, developing into independent power centres. In Kuwait especially, the merchants used these state institutions to re-enter politics through the back door of the bureaucracy, placing their now educated sons (and later daughters) in key posts where they could create administrative fiefdoms and use state resources to rebuild patron–client ties. In Kuwait, some pockets of efficiency nonetheless emerged in the government. In Qatar, state fiefdoms also arose, although these were typically dominated by ruling family members, who took over key state institutions as they were created. Even today a few key family members dominate the core state institutions. Only in more recent years, under Sheikh Hamad (r. 1995–2013), has the ruler in Qatar been able to circumvent the cumbersome bureaucracies to create more effective policy implementation by establishing separate laws and procedures for industrial zones, tourism zones and state-owned enterprises. The cost, however, has been fragmentation and redundancy (Hertog 2010a, p. 268). In keeping with historical patterns, the government has made a clear effort to incorporate established merchants, along with technocrats, into the top management positions of these SOEs (Kamrava 2013, p. 149).

To counter the merchants' power, the rulers in Kuwait extended not only economic benefits but also political benefits to the national population by creating an elected National Assembly following independence in 1961. The ruler's goal was to ensure the continued political marginalisation of the merchants by empowering groups outside the old economic elite (notably bedouins) and enticing others to shift their historical clientelistic ties from the merchants to the rulers. Tribes were settled in subsidised housing in the outer circles of the capital in the 1950s and given employment in the police and military. Tribal deputies were then enticed into the National Assembly with electoral laws that favoured their districts, further weakening the political power of the merchants, some of whom entered parliament as liberals in its early years. The decision to offset any potential power of the merchants in parliament was shaped, in part, by the memory of an earlier merchant-led uprising, the Majlis Movement, in 1938, which had resulted in the election by the leading merchant families of a legislative assembly that directly challenged the Amir's authority. This act followed the failure of the Amir to share his initial oil royalties with the population in any significant way. Qatar's Amir at the time, Sheikh Abdallah bin Jassim al-Thani, behaved in a similar way, treating oil revenues as personal income. There, however, because of the weakness of the merchant class and the size of the royal family, opposition came largely from his relatives, and the Amir was forced to relent to their demands by granting larger family allowances. (Fromherz 2012, p. 131). Well into the 20th century the government of Qatar was spending one-quarter of its income on royal family allowances to placate the various factions (Gray 2013).

Following independence, one goal of Kuwait's Amirs was to preempt any such merchant-led assembly by filling the new National Assembly with middle class supporters. Counter-intuitively, the expansion of popular input into decision-making appeared as a result of efforts by rulers to centralise power. In exchange for financial guarantees, merchants willingly opted out of the political process. Exclusionary and inclusionary policies, rather than being opposites, were in fact two sides of the same coin. Political liberalisation was a top-down strategy, a calculated and limited tactical move by the ruler to maintain power. That it was popular and occurred in the presence of some pressure from below should not obscure this fact.

Over time new groups were incorporated into this Assembly (and the bureaucracy): Shias, Sunni liberals, Sunni Islamists and finally women (granted suffrage in 2005), allowing the ruler to play groups off against each other to maintain power. But even as its social base became more diverse, the underlying structure and purpose of the National Assembly endured. And while the Assembly could sometimes block government initiatives, the government could counter by balancing supporters and opponents, and, when that failed, by calling for new elections. Unlike most parliaments, Kuwait's National Assembly also allowed unelected cabinet ministers to vote, giving the ruler an advantage even in that body.

The decision to create a parliament had costs. Some were political: the Assembly has been successful in blocking government initiatives and even forcing the resignation of some government ministers. Popular rentierism, as Yom (2011) describes it, has also had economic costs. Public expectations of government largesse have grown over time until the parliament has acquired what Hertog (2010b) describes as a 'fiscally reckless character', passing costly increases in state salaries and direct subsidies, maintaining bloated bureaucracies and inhibiting the formation of effective state-owned enterprises (SOEs), notably in Kuwait's case by blocking Project Kuwait, a government-favoured SOE designed to develop new oil fields in partnership with transnational oil companies. This cost, however, remains bearable for the government because the business community has stood behind the rulers through economic and political crises ranging from the collapse of the Suq al-Manakh stock market in 1982 to the 1990 Iraqi invasion and the Arab Spring.

Qatar's rulers did not create a National Assembly, because the weaker merchant class did not pose as great a potential political challenge, and so did not require the rulers to construct an alternate centre of power. However, Qatar did begin to hold regular elections for a Municipal Council with limited authority in 1999. The government of Qatar has promised to hold elections for a national body for years, but actual elections have been repeatedly postponed. The decision to hold municipal elections came about partly in response to external pressure and was part of a regional movement in the same period in all the Gulf Cooperation Council states to increase political participation, typically by introducing partially elected consultative councils. As in Kuwait, political liberalisation, although more limited, was a tactical move by the ruler to contain reform rather than a serious step towards a more democratic state. While Qatar, like Kuwait, has a generous welfare state, which its tremendous hydrocarbon revenues and small population allow, the absence of any countervailing institution comparable to Kuwait's National Assembly has meant that the Qatari government has been much more successful in developing state-owned enterprises, in turn endowing it with the financial ability to placate everyone: the contentious ruling family, the merchants (old and new) and the new national middle class.

The arrangement with the merchants was one of many policies the rulers implemented to maintain domestic peace in the new environment created by oil revenues. Rulers also formed new and independent ties with their own family members. Political kinship, usually considered a traditional vestige, was in fact a new response to the demands of the oil-induced bureaucratic state. Before oil, the rulers' families played a more modest role in governing, one they shared with the merchants. After oil, rulers built their families into ruling institutions, ones that came to control the most important cabinet posts, or sovereign ministries (e.g., defence, interior, foreign affairs) and other key institutions such as the

ruler's *diwan* (advisory council) and the position of prime minister, a system Michael Herb has characterised as dynastic monarchism (Herb 1999). Members of the ruling family, both those who held formal office and the many who did not, also came to participate in ruling family councils, held outside public view, which handle disputes within the ruling family and potentially embarrassing public misbehaviour of ruling family members, as well as making key decisions on the most important domestic and foreign policy issues. The ruling family as an institution is thus a relatively new phenomenon, emerging after and as a result of the explosion of oil revenues. It has been reinforced by shifting marriage patterns, notably an increase in ruling family endogamy (particularly for women), which has replaced historical intermarriage with merchant family members.

In Kuwait, where the al-Sabah family had achieved political ascendancy in the early 18th century, the ruling family emerged as a more cohesive and powerful force. In Qatar, where the relatively larger al-Thani family rose to power much later, in the mid-19th century (1868), and only with the help of the British, the ruling family remained less cohesive and the ruler less powerful. (The al-Sabah family relied on the British to maintain independence from the Ottoman Empire, but they came to power on their own.) In Qatar, members of the al-Thani family dominated the sovereign ministries and state-owned enterprises, as well as many private firms. There, however, some family members also carved out their own niches in the state bureaucracy, resistant to the ruler's oversight. In Qatar the process of reining in the ruling family following the emergence of an oil economy took decades and two bloodless coups (in 1972 and 1995). A degree of centralised authority over the ruling family was ultimately achieved by Sheikh Hamad, who, after seizing power in 1995, introduced a series of economic reforms that turned Qatar into a major gas producer, giving it one of the highest per capita GNPs in the world. His economic success was coupled with political decisions that helped centralise rule. He included 13 ruling family members in his initial cabinet (Gray 2013, p. 61). He moved very quickly to name a son as heir apparent (thus pre-empting family squabbling over a favourite issue). He forced out of office or otherwise marginalised family members he suspected of closer ties to his deposed father, replacing them with younger and more loyal members of the family (Kamrava 2013, p. 117). His insistence on making decisions previously taken by other family members, especially in the economic realm, gave him a larger degree of control over the family, especially after he survived an attempted family coup in 2011 (Gray 2013, p. 60). In 2003 he created a new constitution which institutionalised a decision he had announced soon after his accession, limiting succession to the Amir's son, rather than to any al-Thani family member. It took longer, but in the end Sheikh Hamad created a system of dynastic monarchism similar to Kuwait's. Ruling family members dominated the government, but their selection was now based more on loyalty and competence than on the influence of their family factions (Gray 2013, p. 62).

The system was strong enough in both countries to endure the crash of oil prices in the mid-1980s and, in the case of Kuwait, the Iraqi invasion of 1990. Paying for the war to liberate Kuwait depleted much of the state's assets built up in previous years. But with the return of independence, and higher oil prices, the old system resumed. In the 1990s the Amir deepened his relationship with the old merchant elite by bringing more merchant technocrats into cabinet and government positions (Yom 2011, p. 236). In Kuwait the business community has remained largely outside the National Assembly, yet supportive of the democratic process because of the checks that body offers on corruption and any autocratic tendencies in the ruling family.

The Arab Spring

The relationship between rulers and merchants established in the post-oil era explains in some key ways the evolution of the Arab Spring in Kuwait and Qatar. The Arab Spring swept through both countries, but did not openly challenge, let alone overthrow, the leaders. In Qatar the

protests were almost non-existent. In Kuwait, the protests, while larger, simply amplified ongoing complaints against the government. In both countries the wave of protest was significantly weaker than those seen in other countries in the region. This was not the result of political quiescence. Historically, both countries, especially Qatar, had seen large demonstrations, most notably during the era of Arab nationalism. The quieter response to the Arab Spring was, in part, the consequence of the presence of oil wealth, but also a result of the experience that the rulers had gained, initially through dealing with the merchants, in quickly and effectively deploying those revenues to stave off or contain opposition. Even in Qatar, where the Arab Spring was barely a whisper, the government preemptively increased public sector salaries by 120% in September 2011 (Abdulla 2014, p. 44). In Kuwait, the government responded to the nightly protests by handing out 1,000 Kuwait dinars ($3,000) to each Kuwaiti citizen and granting a year of food subsidies (Abdulla 2014, p. 47). The Arab Spring demonstrated that oil wealth, carefully and quickly spent, can protect governments from the kinds of protests that brought down leaders elsewhere in the region, without requiring resort to force (Gause 2013; Yom and Gause 2012). This last point is worth emphasising. While much of the rentier literature has focused on the ability of oil-producing states to stave off demands for democratisation, the Kuwaiti case demonstrated from the early days that some political participation might be useful even to rulers in a rentier state. What oil revenues also provide is the ability to deal with the opposition that does arise without resorting to force, by giving rulers the option of allocating oil revenues to co-opt potential and real opposition.

However, oil revenues, even when carefully distributed, do not eliminate political demands completely. While oil revenues have bought the state a degree of distance from society, as Gray (2013, p. 9), points out, no state is fully autonomous and thus needs to maintain a degree of legitimacy, or at least popular support. As the state's scope grew over the decades, and the system consolidated, the distance between the ruler and the population in both states also grew. As social services became the norm, citizens came to view them as legitimate claims on the government rather than evidence of the rulers' generosity. Historically, rulers in Kuwait and Qatar had responded to problems of maintaining popular loyalty in part with a stress on normative socialisation, directing state revenues to the development of a strong national identity, including socialisation through public education. National dress, by custom barred to non-nationals, also guaranteed privileged treatment. Even granting suffrage was a way of reinforcing national identity, rewarding citizens and heightening the distinction between nationals and expatriates. This was particularly the case in Kuwait (Tetreault 2000).

Both states, but especially Qatar, also devoted significant effort to creating a new civic myth linking the pre-oil past to a modern national identity that transcends lines of class, tribe and sect. Particularly in Qatar, with a weaker initial sense of national identity, the government devoted considerable resources to building museums and heritage sites and to creating an idealised Qatar of the imagination which privileged the narratives of the ruling family and the desert (where the rulers originated) over the maritime identity (in which the merchants played the key roles) that had actually dominated the economy historically. Qatar began with a small national museum, built around the original Amiri palace of Sheikh Abdallah bin Jassim al-Thani who ruled Qatar in the early 20th century (the museum is presently being expanded and reconstructed), then followed with several more museums and heritage sites, notably *Souq Waqif*, a market rebuilt on the site of an old market and designed to convey a vision of traditional Qatar. Kuwait's heritage museums and sites, while not as extensive, likewise celebrate a similar vision of Kuwait's history, as do holidays such as Kuwait's *Yawm al-Bahr*, a holiday linking symbols and activities of Kuwait's economic and social past with patriotic songs and a sense of modern nationalism (Abou-Samra 2014, p. 185). (On the development of government-sponsored cultural heritage in the Gulf, see Exell and Rico 2014.)

When the Arab Spring arrived, muted political demands arose, although notably largely not from the business community. The pacts of the early oil

era remained intact. In neither country did the opposition call for the fall of the regime; nor did it call for a change in the coalitional arrangements on which that regime rested. The ruling family's right to rule was not challenged; rather people used the Arab Spring to protest issues of pre-existing concern such as corruption, education and the role of Islam in society. In Qatar, political demands were relatively few and were deflected by distributing more oil and (especially) gas revenues, initially in the form of public sector raises, reaching nearly all Qataris (Gray 2013, p. 235). The ruler also preemptively removed any potential demand for a change in leadership from a largely youthful population by abdicating in favour of his son Sheikh Tamim (born in 1980) in 2013, an indication of the extent to which he had overcome much of the historical ruling family factionalism.

In Kuwait protest was greater, owing largely to the existence of a National Assembly as a natural focal point for expressing grievances. But the core political disagreements now expressed again predated the Arab Spring, emanating from political confrontations between the Assembly and Amir Sabah al-Ahmad al-Sabah, dating back to his somewhat complicated succession in 2006 (Tetreault 2006). The Arab Spring only amplified the power struggle between the parliament and the ruling family which had continued since that succession. The National Assembly site now became the focal point for opposition. Demonstrations, some with as many as 50,000 protesters, emerged, culminating in the storming of parliament in November 2011, which prompted the prime minister (facing corruption accusations) to resign. The opposition, however, remained largely loyal: although some called for gentle moves towards a constitutional monarchy, there were no public calls for the fall of the regime. As in Qatar, the government's response was also, by regional standards, restrained.

Conclusion

The sources of capital remain centrally important to understanding politics in Kuwait and Qatar, as well as the other Arab Gulf states. However, the mechanisms through which these revenues enter the economy are equally important. The creation of new alliances, in this case between rulers and merchants, in the early days of oil set in place patterns in politics that continue to shape political events today. The arrangements struck in the early days of oil have proven quite resilient. They help explain the continued existence of powerful ruling families, the extent of wealth and power of the business community, and the vehicles they choose to exercise that power. Just as these arrangements allowed superficially anachronistic monarchs to survive the regional upheavals that swept away other monarchs in the Arab nationalist era of the 1950s and 1960s, they have also enabled the rulers in Qatar and Kuwait to survive the Arab Spring.

See Also

► Oil and the Macroeconomy

Bibliography

Abdulla, A. 2014. The impact of the Arab Spring on the Arab Gulf states. In *The silent revolution: the Arab spring and the gulf states*, ed. M. Seikaly and K. Mattar. Berlin: Gerlach Press.

Abou-Samra, R. 2014. A spring of concentric circles: Overlapping identities in Kuwait and Bahrain and their effect on the Arab Spring. In *The silent revolution: The Arab spring and the gulf states*, ed. M. Seikaly and K. Mattar. Berlin: Gerlach Press.

Azoulay, R. 2013. The politics of Shi'i merchants in Kuwait. In *Business politics in the middle East*, ed. S. Hertog, G. Luciani, and M. Valeri. London: Hurst & Company.

Crystal, J. 1995. *Oil and politics in the gulf: Rulers and Merchants in Kuwait and Qatar*. Cambridge: Cambridge University Press.

Exell, K., and T. Rico, eds. 2014. *Cultural heritage in the Arabian Peninsula: Debates, discourses, and practices*. Burlington: Ashgate.

Fromherz, A. 2012. *Qatar: A modern history*. Washington: Georgetown University Press.

Gause, F. G., III. 2013. Kings for all seasons: How the Middle East monarchies survived the Arab Spring. *Brookings Doha Center Analysis Paper,* Number 8.

Gray, M. 2013. *Qatar: Politics and the challenges of development*. London: Lynne Rienner.

Hanieh, A. 2011. *Capitalism and class in the gulf Arab States*. New York: Palgrave Macmillan.

Herb, M. 1999. *All in the family: Absolutism revolution, and democracy in middle East Monarchies*. Albany: State University of New York Press.

Hertog, S. 2010a. *Princes, brokers, and bureaucrats: Oil and the State in Saudi Arabia*. Ithaca: Cornell University Press.
Hertog, S. 2010b. Defying the resource curse: Explaining successful state-owned enterprises in rentier states. *World Politics* 62(2): 261–301.
Kamrava, M. 2012. The political economy of rentierism. In *The political economy of the persian gulf*, ed. M. Kamrava. New York: Columbia University Press.
Kamrava, M. 2013. *Qatar: Small state, big politics*. Ithaca: Cornell University Press.
Khouja, M.W., and P.G. Sadler. 1979. *The economy of Kuwait: Development and role in international finance*. London: MacMillan.
Moore, P.W. 2004. *Doing business in the middle east: Politics and economic crisis in Jordan and Kuwait*. Cambridge: Cambridge University Press.
Seznek, J.-F. 2007. Changing circumstances: Gulf trading families in the light of free trade agreements, globalization and the WTO. In *The gulf family: Kingship policies and modernity*, ed. A. Alsharekh. London: Saqi.
Tetreault, M.A. 2000. *Stories of democracy: Politics and society in contemporary Kuwait*. New York: Columbia University Press.
Tetreault, M. A. 2006. *Kuwait's annus mirabilis. Middle East Research and Information Project,* 7 September.
Yom, S.L. 2011. Oil, coalitions, and regime durability: The origins and persistence of popular rentierism in Kuwait. *Studies in Comparative International Development* 46: 217–241.
Yom, S., and F.G. Gause III. 2012. Resilient royals: How arab monarchies hang on. *Journal of Democracy* 23: 4.

Oil and the Macroeconomy

James D. Hamilton

Keywords
Oil and the macroeconomy; Real business cycles; Inflation

JEL Classifications
Q43

Nine out of ten of the US recessions since the Second World War were preceded by an upward spike in oil prices. One way to inquire whether this might be just a coincidence is with a statistical regression of real GDP growth rates (quoted at a quarterly rate) on lagged changes in GDP growth rates and lagged logarithmic changes in nominal oil prices. The results from an ordinary least squares (OLS) estimation of this relation for $t = 1949{:}\text{II}$ to $1980{:}\text{IV}$ are as follows (standard errors in parentheses):

$$y_t = \underset{(0.18)}{1.14} + \underset{(0.09)}{0.20}\, y_{t-1} + \underset{(0.09)}{0.05}\, y_{t-2} - \underset{(0.09)}{0.10}\, y_{t-3} - \underset{(0.09)}{0.19}\, y_{t-4} - \underset{(0.026)}{0.004}\, o_{t-1} - \underset{(0.026)}{0.027}\, o_{t-2} - \underset{(0.026)}{0.034}\, o_{t-3} - \underset{(0.027)}{0.065}\, o_{t-4}.$$

The coefficient on the fourth lag of oil prices ($1o_t - 4$) is negative and highly statistically significant (t-statistic $= -2.4$), and an F-test leads to a rejection of the null hypothesis that the coefficients on lagged oil prices are all zero with a p-value of 0.005. Quite a few studies have tested and rejected the hypothesis that the relation between oil prices and output could just be a statistical coincidence, including Rasche and Tatom (1977, 1981), Hamilton (1983), Burbidge and Harrison (1984), Santini (1985, 1992), Gisser and Goodwin (1986), Rotemberg and Woodford (1996), Daniel (1997), Raymond and Rich (1997), Carruth et al. (1998), and Hamilton (2003).

Another possibility is that the correlation between oil prices and output results from common dependence on some third factor or factors that are the true cause of both the increase in oil prices and the subsequent recession. For example, something about the last stages of an economic expansion may often produce a surge in oil prices just before output is about to turn down, so that both the oil price increase and the subsequent recession result from the same business cycle dynamics. This is difficult to reconcile with the fact that, at least for the early post-war period, oil price changes could not be predicted from earlier movements in other macro variables (Hamilton 1983), and that most of the oil spikes can be attributed to exogenous events such as military conflicts (Hamilton 1985). However, Barsky and Kilian (2002, 2004) have recently developed challenges to the latter claim.

Predicted Size of Effects

Economic theory suggests that it is the real oil price rather than the nominal price that should matter for economic decisions. It does not make much difference in summarizing the size of any given shock whether one uses the nominal price o_t or the real price of oil, since in most of the shocks discussed here the move in nominal prices is an order of magnitude larger than the change in overall prices during that quarter. However, particularly in the early part of the sample, the nominal oil price would stay frozen for years and then adjust suddenly. To the extent that there is a difference between using nominal and real prices as the explanatory variable in such regressions, the real price results from the confluence of two forces: events such as the Suez crisis, which accounts for almost all of the movement in the nominal price between 1955 and 1965, and the quarter-to-quarter change in inflation, which is completely endogenous with respect to the economy and whose consequences for future output are likely to be quite different from those of an oil shock. In so far as the statistical exogeneity of the right-hand variables is important for interpreting the regression, many researchers have for this reason used the nominal oil price change rather than the real oil price change as the explanatory variable.

One simple framework for thinking about what the effects of energy supply disruptions should be comes from examining a production function relating the output Y produced by a particular firm to its inputs of labor N, capital K, and energy E:

$$Y = F(N, K, E).$$

Suppose that output is sold for a nominal price P dollars per unit, labour is paid nominal wage W, energy's nominal price is Q, and capital is rented at nominal rate r. The profits of the firm are given by

$$PY - WN - rK - QE.$$

A price-taking profit-maximizing firm would purchase energy up to the point where the marginal product of energy is equal to its relative price,

$$F_E(N, K, E) = Q/P,$$

where $F_E(N, K, E)$ denotes the partial derivative of $F(\,)$ with respect to E. If we multiply both sides of the above equation by E and divide by Y, we find

$$\frac{\partial \ln F}{\partial \ln E} = \frac{QE}{PY}.$$

In other words, the elasticity of output with respect to a given change in energy use can be inferred from the dollar share of energy expenditures in total output.

This dollar share for the economy as a whole is fairly small. For example, in 2000 the United States consumed about 7.2 billion barrels of oil. At a price of $30 a barrel, that represents only 2.2 per cent of a $9.8 trillion nominal GDP. With the rapid price increases of 2003–5, that share has risen to 3.8 per cent of GDP. Table 1 reports Hamilton's (2003) values for the size of the supply disruptions associated with the five most important oil shocks, calculated from the magnitude of the drop in production in the affected countries. Kilian (2005) has more modest estimates based on his inference that production might have fallen even in the absence of the indicated events, and neither Hamilton's nor Kilian's figures take into account the fact that typically production

Oil and the Macroeconomy, Table 1 Exogenous disruptions in world petroleum supply, 1956–90

Date	Event	Drop as % of world production	Change in US real GDP (%)
Nov. 1956	Suez crisis	10.1	−2.5
Nov. 1973	Arab–Israel war	7.8	−3.2
Nov. 1978	Iranian revolution	8.9	−0.6
Oct. 1980	Iran–Iraq war	7.2	−0.5
Aug. 1990	Persian Gulf war	8.8	−0.1

Source: Hamilton (2003)

increased in other parts of the world to make up part of the gap. Even using the ten per cent figure and a four per cent crude oil share, however, such shocks would by the above calculation be predicted to reduce GDP by only 0.4 per cent. Table 1 also reports the amount by which US real GDP declined between the date of the oil shock and the trough of the subsequent recession, which trough usually was reached a little over a year after the oil shock. Since the US economy would grow 3.4 per cent during a typical year, these numbers imply declines of real GDP relative to trend in excess of four per cent, an order of magnitude greater than predicted by the factor share argument. Furthermore, Bohi (1991) failed to find statistically significant evidence that industries with greater energy factor shares suffered more than others in response to the oil shocks of the 1970s.

One would arrive at a similar prediction if one thought of the oil shock as an exogenous change in the price of oil rather than a decrease in the quantity supplied. Faced with an increase in fuel costs, one option a given consumer would always have would be to keep on buying as much gas as before and just pay the higher price, decreasing other expenditures as needed. The value of what is lost by such behaviour is given by $E \cdot \Delta Q$; or, to express this relative to total income PY,

$$\frac{E \cdot \Delta Q}{PY} = \frac{QE}{PY} \cdot \frac{\Delta Q}{Q},$$

in other words, the percentage change in oil prices $\Delta Q/Q$ is again multiplied by energy's value share QE/PY. This actually places an upper bound on the value of what the consumer loses, because, in so far as the consumer opts to reduce E rather than hold E fixed, it must be because the latter strategy is in fact an inferior option.

If these oil shocks did contribute to economic downturns, this would have to be attributed to the movements they induced in other factors of production rather than to the value of the lost energy input per se. Some modest adjustments of other factors would be anticipated in a frictionless neoclassical model, but these appear to be small. Kim and Loungani's (1992) real business cycle analysis suggested that oil price shocks could explain only a modest component of the variance of US output growth.

One modification that can make a difference is to replace the assumption of perfect competition with mark-up pricing. Rotemberg and Woodford (1996) showed that this can induce a response of labour utilization to an oil price shock that greatly amplifies the effects, with simulations in which a ten per cent increase in energy prices could lead to a 2.5 per cent drop in output six quarters later.

Another important margin is the capital utilization rate, as emphasized by Finn (2000), who was able to arrive at similar quantitative effects as Rotemberg and Woodford even under the assumption of perfect competition.

Other Mechanisms

Another explanation offered for the correlation between energy prices and output has to do with the role of monetary policy. Barsky and Kilian (2002, 2004) argued that a monetary expansion was the cause of much of the 1973–4 oil price increase, and that this monetary expansion also set the stage for a subsequent decline in output. Bernanke et al. (1997) took the view that the oil shocks were exogenous, but the Federal Reserve responded to them by raising interest rates in order to control inflation, with this monetary contraction itself the principal cause of the downturns. Hamilton and Herrera (2004) argued that the Bernanke, Gertler and Watson conclusion was due primarily to the fact that these authors omitted the biggest effects of oil shocks corresponding to the coefficients on o_{t-3} and o_{t-4} in the regression above. Leduc and Sill (2004) added sticky prices to a theoretical model generalizing the approach considered by Finn (2000), and concluded that monetary policy makes only a modest contribution. More empirically oriented studies also concluding that the oil shocks were more important than any monetary contraction include Dotsey and Reid (1992), Hoover and Perez (1994), Ferderer (1996), Brown and Yücel (1999), and Davis and Haltiwanger (2001).

A different class of explanations emphasizes the frictions in reallocating labour or capital across different sectors that may be differentially affected by an oil shock. For example, one common consequence of an oil price shock is a sudden drop in demand for certain kinds of cars, which leads to lower capacity utilization at affected plants (Bresnahan and Ramey 1993). Because labour and capital cannot move costlessly to alternative productive activities, the result is idle resources that can significantly multiply the effects described above. Manufacturing of transportation equipment is one of the industries most affected by oil shocks in the United States but has one of the lowest energy intensities, and thus is part of the reason that Bohi (1991) found no connection between energy intensity and output decline. Lee and Ni (2002) found that oil price shocks tend to reduce supply in oil-intensive industries but reduce demand in other industries such as autos. Davis and Haltiwanger (2001) found oil shocks reduce employment the most in industries that are more capital intensive, more energy intensive, and have greater product durability. Keane and Prasad (1996) documented significant differences across industries in the effects of oil shocks on workers' wages.

Hamilton (1988) and Atkeson and Kehoe (1999) provided theoretical analyses of the way in which technological costs of adjusting capital or labour can result in magnification of the disruptive effects of oil shocks. One of the key predictions of such models is that, unlike the factor share stories, the response of output to oil prices would not be log-linear. When oil prices go up, consumers may postpone their car purchases, but when oil prices go down, they do not go out and buy a second car. In fact, it is a theoretical possibility that, as a result of the output that is lost from trying to reallocate capital and labour, the short-run effect of an oil price decrease would actually be a decline rather than an increase in output.

Linearity

If one estimates a log-linear relation between GDP growth and lagged oil prices, the statistical significance of the relation falls as one adds more data (Hooker 1996), suggesting at a minimum that a linear relation is either mis-specified or unstable. For example, when the regression described above is re-estimated with data through 2005:II, the result is

$$y_t = \underset{(0.11)}{0.69} + \underset{(0.07)}{0.28}\, y_{t-1} + \underset{(0.07)}{0.13}\, y_{t-2} - \underset{(0.07)}{0.07}\, y_{t-3} - \underset{(0.07)}{0.12}\, y_{t-4} - \underset{(0.006)}{0.003}\, o_{t-1} - \underset{(0.006)}{0.006}\, o_{t-2} - \underset{(0.006)}{0.002}\, o_{t-3} - \underset{(0.006)}{0.015}\, o_{t-4}.$$

Although the t-statistic on o_{t-4} remains statistically significant with a p-value of 0.02, an F-test of the null hypothesis that all four coefficients on lagged oil prices are zero would be accepted with a p-value of 0.11. The size of the effect is substantially smaller as well – whereas the 1949–80 regression would predict that GDP growth would be 2.9 per cent slower (at an annual rate) four quarters after a ten per cent oil price hike, the 1949–2005 regression would predict only 0.7 per cent slower growth.

A number of authors have concluded that this instability is due to the nonlinearity of the relationship, with a linear relationship breaking down empirically when the huge oil price drops of 1985 failed to produce an economic boom. Loungani (1986) and Davis (1987a, b) were the first to report evidence of nonlinearity of these relations, which they interpreted as implying that the effects of oil shocks resulted from sectoral shifts with costly reallocation of resources. Mork (1989) estimated separate coefficients on oil price increases and decreases, and found that the latter were statistically insignificantly different from zero.

To the extent that the oil shocks are operating through an effect on demand for items such as less fuel-efficient cars, the influence would depend not just on the size of the oil price increase but also the context in which it occurred. Lee et al. (1995) found that much better forecasts of GDP growth were obtained if one divided the oil price increase by the standard deviation of recent price volatility. Hamilton (2003) used a flexible parametric model to investigate the nature of this nonlinearity, and

found support for the Lee, Ni and Ratti formulation as well as an alternative that looks at how much the oil price might exceed its previous three-year peak; if it does not exceed the previous three-year peak, no oil shock is said to have occurred. An OLS regression of quarterly GDP growth (quoted at a quarterly rate) on lags of this net oil price measure for 1949:II to 2005:II results in the following estimates:

$$y_t = \underset{(0.12)}{0.87} + \underset{(0.07)}{0.24}\, y_{t-1} + \underset{(0.07)}{0.11}\, y_{t-2} - \underset{(0.07)}{0.08}\, y_{t-3} - \underset{(0.07)}{0.13}\, y_{t-4} - \underset{(0.012)}{0.009}\, o^{\#}_{t-1} - \underset{(0.012)}{0.014}\, o^{\#}_{t-2} - \underset{(0.012)}{0.009}\, o^{\#}_{t-3} - \underset{(0.012)}{0.031}\, o^{\#}_{t-4}.$$

Here an F-test of the null hypothesis that all coefficients are zero is rejected with a p-value of 0.006, and a ten per cent increase in oil prices above their previous three-year high is predicted to reduce quarterly GDP growth (quoted at an annual rate) by 1.4 per cent.

Similar evidence of nonlinearity, with oil price increases reducing real output growth, has also been reported for a number of other countries by Mork et al. (1994), Cuñado and Pérez de Gracia (2003), and Jimenez-Rodriguez and Sanchez (2005).

Other Factors and Consequences

As noted by Kilian (2005), civil unrest in Venezuela in December 2002 led to a drop in production of 2.3 million barrels a day, representing 3.4 per cent of world production at the time. The net oil price series $o^{\#}_t$ reflected a surge in crude oil prices 20 per cent above their previous three-year high. Nevertheless, there was no discernible drop in GDP. Another surge in $o^{\#}_t$ of 18 per cent occurred in 2004:III, accompanied by a 1.3 per cent increase in world production, and a third surge of 21 per cent in 2005:I, accompanied by a 0.2 per cent increase in production, with no recession as of the time of this writing (August 2005). It is clear from the last two examples in particular that demand increases rather than supply reductions have been the primary factor driving oil prices over recent years. In so far as these demand increases resulted from global income growth, one wouldn't expect to see the sharp drop in consumer spending on other key items that accompanied the episodes in Table 1. At a minimum, the failure of a recession to result as of the time of this writing from the oil price increases of 2003–5 suggests that there is not simply a mechanical relation, even a nonlinear one, between oil prices and output. The experience is consistent with the claim that the key mechanism whereby oil shocks affect the economy is through a disruption in spending by consumers and firms on other goods and that, if this disruption fails to occur, the effects on the economy are indeed governed by the factor share argument.

Another potential macroeconomic effect of oil price shocks is on the inflation rate. The long-run inflation rate is governed by monetary policy, so ultimately this is a question about how the central bank responds to the oil shock. Hooker (2002) found evidence that oil shocks made a substantial contribution to US core inflation before 1981 but have made little contribution since, consistent with the conclusion of Clarida et al. (2000) that US monetary policy has become significantly more devoted to curtailing inflation.

See Also

- ▶ Cost-Push Inflation
- ▶ Inflation
- ▶ Real Business Cycles

Bibliography

Atkeson, A., and P. Kehoe. 1999. Models of energy use: Putty-putty versus puttyclay. *American Economic Review* 89: 1028–1043.

Barsky, R., and L. Kilian. 2002. Do we really know that oil caused the great stagflation? A monetary alternative. In *NBER macroeconomics annual 2001*, ed. B. Bernanke and K. Rogoff. Cambridge: MIT Press.

Barsky, R., and L. Kilian. 2004. Oil and the macroeconomy since the 1970s. *Journal of Economic Perspectives* 18: 115–134.

Bernanke, B., M. Gertler, and M. Watson. 1997. Systematic monetary policy and the effects of oil price shocks. *Brookings Papers on Economic Activity* 1997(1): 91–124.

Bohi, D. 1991. On the macroeconomic effects of energy price shocks. *Resources and Energy* 13: 145–162.

Bresnahan, T., and V. Ramey. 1993. Segment shifts and capacity utilization in the U.S. automobile industry. *American Economic Review: Papers and Proceedings* 83: 213–218.

Brown, S., and M. Yücel. 1999. Oil prices and US aggregate economic activity: A question of neutrality. *Federal Reserve Bank of Dallas Economic and Financial Review (Second Quarter)*: 16–23.

Burbidge, J., and A. Harrison. 1984. Testing for the effects of oil-price rises using vector autoregressions. *International Economic Review* 25: 459–484.

Carruth, A., M. Hooker, and A. Oswald. 1998. Unemployment equilibria and input prices: Theory and evidence from the United States. *The Review of Economics and Statistics* 80: 621–628.

Clarida, R., J. Galí, and M. Gertler. 2000. Monetary policy rules and macroeconomic stability: Evidence and some theory. *Quarterly Journal of Economics* 115: 147–180.

Cuñado, J., and F. Pérez de Gracia. 2003. Do oil price shocks matter? Evidence from some European countries. *Energy Economics* 25: 137–154.

Daniel, B. 1997. International interdependence of national growth rates: A structural trends analysis. *Journal of Monetary Economics* 40: 73–96.

Davis, S. 1987a. Fluctuations in the pace of labor reallocation. In *Empirical studies of velocity, real exchange rates, unemployment and productivity*, Carnegie-Rochester Conference Series on Public Policy, 24, ed. K. Brunner and A. Meltzer. Amsterdam: North-Holland.

Davis, S. 1987b. Allocative disturbances and specific capital in real business cycle theories. *American Economic Review: Papers and Proceedings* 77: 326–332.

Davis, S., and J. Haltiwanger. 2001. Sectoral job creation and destruction responses to oil price changes. *Journal of Monetary Economics* 48: 465–512.

Dotsey, M., and M. Reid. 1992. Oil shocks, monetary policy, and economic activity. *Federal Reserve Bank of Richmond Economic Review* 78(4): 14–27.

Ferderer, J. 1996. Oil price volatility and the macroeconomy: A solution to the asymmetry puzzle. *Journal of Macroeconomics* 18: 1–16.

Finn, M. 2000. Perfect competition and the effects of energy price increases on economic activity. *Journal of Money, Credit, and Banking* 32: 400–416.

Gisser, M., and T. Goodwin. 1986. Crude oil and the macroeconomy: Tests of some popular notions. *Journal of Money, Credit, and Banking* 18: 95–103.

Hamilton, J. 1983. Oil and the macroeconomy since World War II. *Journal of Political Economy* 91: 228–248.

Hamilton, J. 1985. Historical causes of postwar oil shocks and recessions. *Energy Journal* 6(1): 97–116.

Hamilton, J. 1988. A neoclassical model of unemployment and the business cycle. *Journal of Political Economy* 96: 593–617.

Hamilton, J. 2003. What is an oil shock? *Journal of Econometrics* 113: 363–398.

Hamilton, J., and A. Herrera. 2004. Oil shocks and aggregate macroeconomic behavior: The role of monetary policy. *Journal of Money, Credit, and Banking* 36: 265–286.

Hooker, M. 1996. What happened to the oil price-macroeconomy relationship? *Journal of Monetary Economics* 38: 195–213.

Hooker, M. 2002. Are oil shocks inflationary? Asymmetric and nonlinear specifications versus changes in regime. *Journal of Money, Credit and Banking* 34: 540–561.

Hoover, K., and S. Perez. 1994. Post hoc ergo propter hoc once more: An evaluation of 'does monetary policy matter?' in the spirit of James Tobin. *Journal of Monetary Economics* 34: 89–99.

Jimenez-Rodriguez, R., and M. Sanchez. 2005. Oil price shocks and real GDP growth: Empirical evidence for some OECD countries. *Applied Economics* 37: 201–228.

Keane, M., and E. Prasad. 1996. The employment and wage effects of oil price changes: A sectoral analysis. *The Review of Economics and Statistics* 78: 389–400.

Kilian, L. 2005. Exogenous oil supply shocks: How big are they and how much do they matter for the US economy? Discussion paper no. 5131. London: CEPR.

Kim, I., and P. Loungani. 1992. The role of energy in real business cycle models. *Journal of Monetary Economics* 29: 173–189.

Leduc, S., and K. Sill. 2004. A quantitative analysis of oil-price shocks, systematic monetary policy, and economic downturns. *Journal of Monetary Economics* 51: 781–808.

Lee, K., and S. Ni. 2002. On the dynamic effects of oil price shocks: A study using industry level data. *Journal of Monetary Economics* 49: 823–852.

Lee, K., S. Ni, and R. Ratti. 1995. Oil shocks and the macroeconomy: The role of price variability. *Energy Journal* 16(4): 39–56.

Loungani, P. 1986. Oil price shocks and the dispersion hypothesis. *The Review of Economics and Statistics* 58: 536–539.

Mork, K. 1989. Oil and the macroeconomy when prices go up and down: An extension of Hamilton's results. *Journal of Political Economy* 91: 740–744.

Mork, K., Ø. Olsen, and H. Mysen. 1994. Macroeconomic responses to oil price increases and decreases in seven OECD countries. *Energy Journal* 15(4): 19–35.

Rasche, R., and J. Tatom. 1977. Energy resources and potential GNP. *Federal Reserve Bank of St. Louis Review* 59(June): 10–24.

Rasche, R., and J. Tatom. 1981. Energy price shocks, aggregate supply, and monetary policy: The theory and international evidence. In *Supply shocks, incentives, and national wealth*, Carnegie-Rochester Conference Series on Public Policy, ed. K. Brunner and A. Meltzer, Vol. 14. Amsterdam: North-Holland.

Raymond, J., and R. Rich. 1997. Oil and the macroeconomy: A Markov state switching approach. *Journal*

of Money, Credit and Banking 29: 193–213 . Erratum, 29 (November, Part 1), p. 555.
Rotemberg, J., and M. Woodford. 1996. Imperfect competition and the effects of energy price increases. *Journal of Money, Credit, and Banking* 28: 549–577.
Santini, D. 1985. The energy-squeeze model: Energy price dynamics in U.S. business cycles. *International Journal of Energy Systems* 5: 18–25.
Santini, D. 1992. Energy and the macroeconomy: Capital spending after an energy cost shock. In *Advances in the economics of energy and resources*, ed. J. Moroney, Vol. 7. Greenwich: JAI Press.

Okun, Arthur M. (1928–1980)

James Tobin

Keywords

Budget deficits; Equality of opportunity; Implicit contracts; Inequality; Leaky bucket; New classical macroeconomics; Okun, A. M.; Okun's Law; Output gap; Potential GNP; Redistribution of income and wealth; Unemployment–inflation trade-off

JEL Classifications

B31

Okun was born in Jersey City, New Jersey, on 28 November 1928. He died suddenly in Washington, DC, on 23 March 1980.

Okun received his BA, ranked first in his college class, in 1949 and his Ph.D. in economics in 1956, both from Columbia University. He started teaching at Yale as Instructor in 1952, and advanced up the ladder to the rank of Professor in 1963. From September 1961 to January 1969 Okun was, except for two academic years 1962–4, on leave from Yale at the President's Council of Economic Advisers (CEA) in Washington, first as a staff member, then as a Council Member 1964–8, and finally as Chairman 1968–9. When Administrations changed in 1969, Okun joined the Brookings Institution as a Senior Fellow, an appointment he held the rest of his life.

Prior to his public service in the 1960s Okun was not well known outside Yale. Those who knew him personally appreciated his extraordinary talents and virtues. He was a great and generous teacher, both in the classroom and out. His open-door office was the place for students and colleagues to get things straight, confusions dispelled, errors corrected, models repaired. A thinker of natural integrity and inexhaustible curiosity, he pursued matters in depth, unsatisfied until logic was tight and facts fell into place. His teachings of policy-oriented macroeconomics created an oral tradition that many beneficiaries remember with deep gratitude. But little of it was published, because Art Okun was unduly modest and perfectionist about putting his wisdom into print.

At the CEA Okun found another metier, macroeconomic analysis directly related to the policy issues of the day. It began when President Kennedy's Council, of which one member was from Yale, enlisted Okun as a consultant. The Council wanted to convince the President, his White House staff, the Congress and the public that reduction of unemployment from seven per cent to four per cent would yield economy-wide benefits much greater than moving from 93 to 96 per cent employment superficially suggested. Okun was asked to estimate the gains of real Gross National Product associated with unemployment reduction. The answer became famous as Okun's Law, one of the most reliable empirical regularities of macroeconomics. Okun found that a reduction of one percentage point of unemployment was associated with a gain of three per cent in real GNP. His research, later published (1962), also provided a methodology for estimating potential GNP, the real output the economy can produce at a full-employment or 'natural' rate of unemployment, and the 'Gap' between actual and potential output.

These concepts are central to estimates of the 'high-employment' or 'structural' federal budget deficits implied by tax and spending policies, as distinguished from actual deficits, which depend also on the performance of the economy as indicated by the Gap. The entire apparatus was displayed in the 1962 *Economic Report of the*

President, and was a mainstay of subsequent *Reports* for 20 years. Okun himself was a major contributor to all the *Reports* 1962–70.

Okun was the Council's principal forecaster and estimator of the consequences of alternative policies. As he won the confidence of Council chairmen, White House staff, presidents, and even Treasury secretaries, he became the obvious choice for President Johnson to appoint Council Member and then Chairman. The period 1966–9 was difficult for the CEA and for Okun personally. The four per cent unemployment target had been achieved in 1965, with negligible cost in higher inflation. Then came the acceleration of Vietnam spending, overheating the economy and lifting the inflation rate three percentage points by 1969. At the beginning of 1966 Gardner Ackley, CEA Chairman, and Okun urged President Johnson to ask Congress to raise taxes. He would not do so until too late, and even then the temporary income surtax of 1968 had disappointingly small effects. When Okun left the government in 1969, the unemployment–inflation nexus became the foremost problem on his research agenda for the rest of his career.

From 1969 much of his energy and leadership went into his brainchild, the Brookings Panel on Economic Activity, which enlisted able economists from Brookings and elsewhere for research on the major macroeconomic developments and policies of the times. The papers are published in *Brookings Papers on Economic Activity*, which under the painstaking editorship of Okun and George Perry quickly became one of the most admired professional journals in economics. The editors put the contents of every issue in perspective with their analytical summaries of the papers and discussions.

Okun had nearly completed a major treatise on macroeconomics (1981) when he died; it was edited and finished by his colleagues at Brookings. The book is a culmination of his thinking and writing over many years, his search for a coherent model of an advanced capitalist economy in a democratic society, based on his understanding of how businesses, workers and consumers behave and relate to one another. Okun did not believe that the economists' favourite paradigm, purely competitive markets cleared by flexible prices – Adam Smith's 'invisible hand' – provided realistic foundations for macroeconomics. He was impressed by the informal reciprocal expectations and obligations that characterize repeated dealings between sellers and customers or employers and workers. A creative phrase-maker, Okun called this web of implicit contracts the 'invisible handshake'. His 'customer markets' are in many ways efficient substitutes for price-cleared auction markets, but they are also the source of endemic macroeconomic difficulties.

Okun saw no easy resolution of the cruel dilemma policymakers face in the trade-off between unemployment and inflation. All too often, and especially in the 1970s, fiscal and monetary demand management could achieve acceptable outcomes in one of these two dimensions only at the cost of unacceptable results in the other. Okun had no use for the monetarist view that inflation could be easily prevented or conquered if only the central bank mustered sufficient will and wisdom. Nor did he share the simplistic view of some theorists of various schools that inflations are neutral and innocuous, devoid of real consequences. He advocated structural anti-inflation policies, including wage and price guideposts strengthened by tax-based incentives for compliance, to diminish the unemployment costs of anti-inflationary monetary and fiscal measures (1978).

The intellectual climate of professional macroeconomics was inhospitable to *Prices and Quantities* when it was published. 'New classical' models relying on 'invisible hand' microfoundations were the dominant fashion. They are theoretically appealing but have trouble explaining the commonly observed facts of business fluctuations. No one knew those facts better than Okun, whose last published paper (1980) is a masterful litany of the many ways new classical business cycle theories fail to fit them. Fashions change and controversies fade. Okun's macroeconomics will be an important component of whatever new synthesis emerges from contemporary debate.

Arthur Okun was not only an effective adviser and participant in the making of economic policy;

he was also a scholar and scientist of *political economy* – the ancient name for our discipline suggests a broader scope of inquiry and concern that most economists essay today. Okun's reflections on the role of the academic policy adviser in government and on the politics and economics of macroeconomic management, published shortly after he returned to private life, are the most thoughtful of the genre (1970).

For his Godkin Lectures at Harvard (1975) Okun chose the broadest and most basic question of political economy: how democratic societies do, can, and should balance the ethical desirability of mitigating inequalities of well-being against the practical utility of the inequalities arising in free markets as incentives for efficient economic performance. Okun coined the metaphor 'leaky bucket' for losses in aggregate wealth incident to government interventions to transfer wealth from rich to poor. Citizens will disagree on the tolerable degree of leakage, he says, but both liberals and conservatives should face the trade-offs realistically. They should be able to agree on measures to plug leaks, exploiting opportunities to diminish inequality without impairing incentives (even if such reforms are not Pareto optimal). Okun suggests an agenda of such opportunities, focusing on measures to assure greater equality of opportunity. The book has already become a classic. In its erudition, logic, lucidity, and wisdom, and above all in its humanity, it truly reflects the qualities of its author.

Selected Works

1962. Potential GNP: its measurement and significance. In *Proceedings of the business and economic statistics section, American Statistical Association*, 98–103. Reprinted in (1983).
1970. *The political economy of prosperity.* Washington, DC: Brookings Institution.
1975. *Equality and efficiency: The big tradeoff.* Washington, DC: Brookings Institution.
1978. A reward TIP. In Senate, Banking, Housing and Urban Affairs, *Anti-inflation proposals*, Hearings, 95 Congress 2 Session, Washington, DC: Government Printing Office. Reprinted in (1983).
1980. Rational-expectations-with-misperceptions as a theory of the business cycle. *Journal of Money, Credit and Banking* Part 2, 12: 817–25. Reprinted in (1983).
1981. *Prices and quantities: A macroeconomic analysis*. Washington, DC: Brookings Institution.
1983. *Economics for policymaking: Selected essays of Arthur M. Okun*, ed. J.A. Pechman, with editor's preface. Cambridge, MA: MIT Press.

Okun's Law

Jesús Crespo Cuaresma

Abstract

Okun's law describes the empirical relationship between changes in unemployment and output at the macroeconomic level and has been regarded since its discovery by Arthur Okun (Potential GNP: its measurement and significance. In: Proceedings of the business and economics statistics. American Statistical Association, Washington, DC, p 98–104, 1962) as a building block of traditional macroeconomic models. This article discusses the interpretation of this relationship and summarizes recent developments in the econometric specification of Okun's law.

Keywords

Aggregate demand; Demand and supply shocks; Okun's coefficient; Okun's law; Output gap; Phillips curve; Production functions; Unemployment

JEL Classifications

E32; J21

The term 'Okun's law' refers to the empirical relationship between changes in unemployment

and output. It is a basic building block of traditional macroeconomic models, where the aggregate supply function is derived from combining Okun's law with the Phillips curve.

In Okun's original contribution (Okun 1962), the empirical relationship between unemployment and output is introduced in the context of the quantification of potential output and the measurement of the social costs of unemployment in terms of forgone production.

Okun (1962) presents estimates for the United States which are based on three alternative econometric specifications aimed at quantifying empirically the relationship between unemployment and output growth: (a) regressing (quarterly) changes in the unemployment rate on (quarterly) percentage changes in production (as proxied by Gross National Product, or GNP), (b) regressing the unemployment rate on percentage deviations from potential output, defined as the exponential trend in GNP and (c) regressing the (logarithmized) employment rate on a linear time trend and (logarithmized) GNP. The effect of output changes on unemployment is quantified by the estimated parameter associated with the output variable in each of these regressions, whose inverse is usually known as 'Okun's coefficient'. The results in Okun's contribution indicate that there exists roughly a three-to-one link between unemployment and output changes, in the sense that an increase/decrease of three percentage points in output (or the output gap, depending on the specification) is associated with a decrease/increase of one percentage point in unemployment. This rule of thumb is proposed as a 'subjectively weighted average' (Okun 1962, p. 100) of the estimates obtained from the three specifications. Estimations based on data including the post-oil crisis period, which can be found in most modern macroeconomic textbooks, tend to reveal an Okun's coefficient that is closer to two than to three.

Okun's arguments (see, for example, Okun 1962, p. 99) suggest that the link found between unemployment and output is not to be understood as a *ceteris paribus* relationship, but rather as capturing also the effects of simultaneous changes in labour force, hours worked and productivity (see also Friedman and Wachter 1974). Okun argues that a reduction in the unemployment rate would induce an increase in the labour force by persuading discouraged workers to seek work actively, and also presents estimates of the increase in hours worked per employed person caused by rising output. The analysis carried out in Okun's contribution, based on data for the United States in 1960, assigns approximately 56 per cent of the change in output to the effect of changes in total labour input measured in hours worked, while the rest is attributed to productivity increases. Prachowny (1993) approaches the quantification of the link between unemployment and output by proposing a specification based on a fairly general production function, where the independent effects of changes in unemployment, hours worked, capacity utilization and labour force on output can be estimated separately. In this setting, Okun's empirical specifications would be appropriate only if certain parameter restrictions on the production function are satisfied. Prachowny (1993) therefore proposes labelling Okun's law 'Okun's theory' and testing these restrictions directly on the data. The estimates of the direct effect of unemployment on output obtained using this specification are correspondingly smaller than in the original contribution by Okun, although the econometric modelling strategy used (based on estimating the production function in gap form and in first differences) is not without criticism. Attfield and Silverstone (1997) reconsider this approach using cointegration techniques and find estimates that are comparable with the original values in Okun (1962).

Obviously, the relationship observed between changes in output and changes in the unemployment rate is determined by the nature of the shocks hitting the economy. The usual interpretation of the relationship summarized by Okun's law refers to arguments based on shocks to aggregate demand. Blanchard and Quah (1989) emphasize the importance of identifying demand and supply shocks in order to estimate and interpret Okun's coefficient. Using a dynamic system formed by the unemployment rate and output growth, Blanchard and Quah (1989) assess the issue by

isolating supply and demand shocks and interpreting the responses of these two variables to each type of shock. The results suggest that the implied Okun's coefficient for demand shocks is slightly above two, while there is no such systematic short-run relationship between unemployment and output changes following a supply shock.

While many macroeconomic textbooks tend to emphasize that the relationship between short-run changes in output and unemployment is a robust and reliable empirical regularity, much of the literature dealing with Okun's law is aimed at evaluating the robustness of this link across countries (Kaufman 1988; Moosa 1997; Lee 2000), in time (Sheehan and Zahn 1980; Gordon 1984; Evans 1989), across econometric specifications (Weber 1995; Lee 2000) and across states of the business cycle – recessions versus expansions – (Lee 2000; Crespo Cuaresma 2003). The results of this branch of literature point towards the existence of asymmetric, country-specific Okun's coefficients, with a higher elasticity of unemployment to output changes in recessions than in expansions, and a lower elasticity in continental European countries compared with Canada, the United States and the United Kingdom. The estimates of Okun's coefficient appear to be sensitive to the specification and de-trending method used for retrieving the cyclical component of output and the unemployment rate. Furthermore, this empirical literature usually reports evidence of structural instability, with a break in Okun's coefficient taking place in the 1970s.

See Also

- ▶ Okun, Arthur M. (1928–1980)
- ▶ Phillips Curve
- ▶ Trend/Cycle Decomposition
- ▶ Unemployment

Bibliography

Attfield, C.L.F., and B. Silverstone. 1997. Okun's coefficient: A comment. *Review of Economics and Statistics* 79: 326–329.

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

Crespo Cuaresma, J. 2003. Okun's law revisited. *Oxford Bulletin of Economics and Statistics* 65: 439–451.

Evans, G.W. 1989. Output and unemployment dynamics in the United States: 1950–1985. *Journal of Applied Econometrics* 4: 213–217.

Friedman, B.M., and M.L. Wachter. 1974. Unemployment: Okun's law, labor force, and productivity. *Review of Economics and Statistics* 56: 167–176.

Gordon, R.J. 1984. Unemployment and potential output in the 1980s. *Brookings Papers on Economic Activity* 15: 537–564.

Kaufman, R.T. 1988. An international comparison of Okun's law. *Journal of Comparative Economics* 12: 182–203.

Lee, J. 2000. The robustness of Okun's law: Evidence from OECD countries. *Journal of Macroeconomics* 22: 331–356.

Moosa, I.A. 1997. A cross-country comparison of Okun's coefficient. *Journal of Comparative Economics* 24: 335–356.

Okun, A.M. 1962. Potential GNP: Its measurement and significance. In *Proceedings of the business and economics statistics*, 98–104. Washington, DC: American Statistical Association.

Prachowny, M.F.J. 1993. Okun's law: Theoretical foundations and revised estimates. *Review of Economics and Statistics* 75: 331–336.

Sheehan, R.G., and F. Zahn. 1980. The variability of the Okun coefficient. *Southern Economic Journal* 47: 488–497.

Weber, C.E. 1995. Cyclical output, cyclical unemployment and Okun's coefficient: A new approach. *Journal of Applied Econometrics* 10: 433–445.

Oligarchs

Sergei Guriev

Abstract

In several countries economic transition was accompanied by the emergence of 'oligarchs' – businessmen who amassed fortunes and used them to influence economic policies. At their height in 2003, a few oligarchs controlled much of Russia's economy, as did a similar elite in Ukraine. Oligarchs seem to run their empires more efficiently than other domestic owners. While the relative

weight of their firms in the economy is huge, it is not excessive by the standards of the global economy where most of them are operating. Policymakers should therefore focus on 'political antitrust' to prevent state capture and subversion of institutions.

Keywords

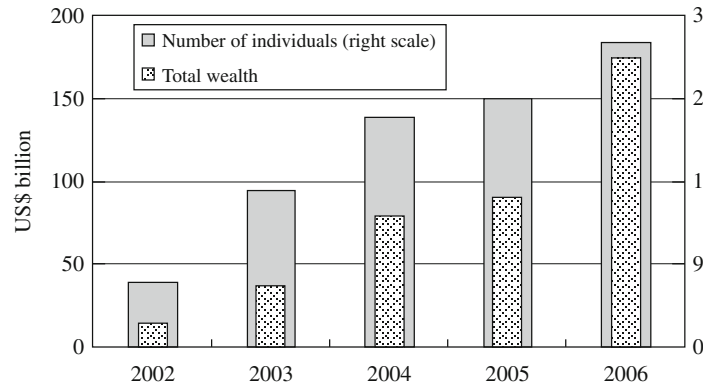
Arbitrage; Berezovsky, B.; Bribery; Democracy; Hold-up problem; Khodorkovsky, M.; Loans-for-shares auctions (Russia); Market power; Oligarchs; Ownership and control, separation of; Ownership concentration; Private property; Privatization; Putin, V.; Rent seeking; Tariffs; Total factor productivity; Transition and institutions; Vertical integration; World Trade Organization

JEL Classifications

P3

An oligarchy, as discussed in Plato's *Republic* and *Statesman* and Aristotle's *Politics*, is a form of government by a small group. Interestingly, while in Plato's works, *oligarchy* is used as a neutral term, and may include both aristocracy and plutocracy, Aristotle already provides the term with a negative connotation, defining oligarchy (similar to Plato's plutocracy) as a deviant form of the rule by a few (while aristocracy remains the correct one).

In its current meaning in transition economies, the term 'oligarch' denotes a businessman who controls sufficient resources to influence national politics. (The lists of oligarchs include only men; the richest Russian businesswoman, Moscow mayor's wife Elena Baturina, ranked outside the top 25 wealthiest Russians in 2004; she entered the *Forbes* billionaires list (Fig. 1) only in 2005 but remained the only woman in the list, ranked 27 out of 34 in 2006 (*Forbes* 2004–6). Such businessmen have played a substantial role in almost all transition countries, although most of the discussion of the role of oligarchs in transition has concerned Russia and Ukraine. The reason for this is also similar to the ideas of Plato and Aristotle, who classified oligarchy as an intermediate form of government between dictatorship/monarchy and democracy. On the one hand, EU accession countries in Central and Eastern Europe have succeeded in building accountable and democratic governments, thus limiting the role of oligarchs. On the other hand, members of the Commonwealth of Independent States (except Russia and Ukraine) have seen the concentration of power in the hands of a single politician rather than a group of rich businessmen. Also, Russian oligarchs have been more prominent than those in Ukraine, in terms of both their wealth (due to Russia's resource richness) and their substantial impact on politics. Actually, in the *Forbes* 2005 and 2006 lists the total wealth of all non- Russian billionaires from transition countries (including China but excluding Hong Kong) was less than that of the single richest Russian. Not surprisingly, Russian oligarchs have been studied in far more detail. This is why this article concentrates on the case study of Russia even though most issues are



Oligarchs,
Fig. 1 Numbers of Russians in the *Forbes* billionaires list, 2002–2006 (*Source: Forbes* (2002–6))

relevant to Ukraine and other transition countries. (See Aslund 2006, for a study of Ukrainian oligarchs; Gorodnichenko and Grigorenko 2005, provide a quantitative analysis.)

It is not clear who first used the term 'oligarch' to describe the newly emerged class of Russian tycoons. *Kommersant* (2003) refers to a pro-market politician Boris Nemtsov (then a governor of Nizhny Novgorod region, later to become a deputy prime minister) and a journalist, Alexander Privalov (then *Izvestiya* daily and *Expert* weekly), both introducing the term in 1994–5. It is also clear that the Russian elite's thinking of oligarchs has been affected by Jack London's *The Iron Heel* (1908), an anti-utopia on the rise of an oligarchy of robber barons, which was widely publicized in Soviet times.

Who Are the Oligarchs?

There is no complete list of Russian oligarchs. Given the multi-layered and nontransparent ownership structure of Russian companies, compiling such a list would be extremely difficult. On the other hand, any such list has to be constantly updated: there is substantial vertical mobility among Russia's richest. For example, out of seven or eight business groups that dominated President Yeltsin's Russia in the 1990s, two were destroyed by the 1998 crisis (SBS and Inkombank), one took a hit but survived to be later sold to fellow billionaires (Roskredit-cum-Metalloinvest), two have their leaders (Berezovsky and Gusinsky) in exile, and one (Khodorkovsky) in prison. Other problems are related to the vagueness of the definition of oligarchs. First, there are different views on how to measure tycoons' power rather than wealth (this is especially important for a comparison between oligarchs and US robber barons). Second, it is not clear whether to count public officials and CEOs of large public companies as oligarchs. In what follows, we stick to the definition of oligarchs as private owners, although certain CEOs of state-owned firms and family members of some government officials do resemble oligarchs in many respects.

The first list of oligarchs probably belongs to Boris Berezovsky (by all accounts, an oligarch himself) who, in his 1996 interview in the *Financial Times,* named seven bankers who controlled about 50 per cent of the productive assets of the Russian economy. Since then there have been numerous lists, some even endorsed by the oligarchs themselves. Still, all the oligarch rankings identify similar sets of individuals. Table 1 presents a list that was constructed based on a study of ownership concentration in a substantial subset of Russian economy by the World Bank's 2004 *Country Economic Memorandum* (CEM) for Russia (see Guriev and Rachinsky 2005, for a detailed description of the data-set and the project). The study refers to summer 2003 – the oligarchs' heyday. While this study has its limitations, it makes it possible to reach some conclusions on who the Russian oligarchs are, and why they matter.

How Important Are the Oligarchs?

First, the oligarchs do control a substantial part of Russian economy. In the CEM sample, they account for about 40 per cent of sales and employment – more than all other private owners combined, or more than federal and regional governments combined. (As of June 2006, quite a few of these oligarchs have seen their assets nationalized, so a more relevant figure would be 30 per cent.)

Cross-country comparisons of wealth concentration are usually based on the share of stock market capitalization controlled by a given number (often ten) of families. Certainly, it is not a perfect metric – after all, it doesn't include firms not listed on stock markets, and emerging markets are likely to provide at best an imperfect measure of value. But we are not aware of comparable data-sets on non-listed firms, so we have to rely on the data on the share of the stock market owned by the top ten families. By that measure, ownership concentration in modern Russia is higher than in any other country for which the data are available. The top ten families or ownership groups (a subset of Table 1) owned 60.2 per cent of Russia's stock market in June 2003. This percentage is much higher than in

Oligarchs, Table 1 Russian oligarchs as of summer 2003

Senior partner (s)	Holding company/ firm, major sector(s)	Employment, '000s (% sample)	Sales, in billions of roubles (% sample)	Wealth, in billions of US dollars	Other ranking[a]	RSPP bureau, head of committee/ taskforce (as of June 2004)
Oleg Deripaska	Base Element/ RusAl, aluminum, auto	169 (3.9)	65 (1.3)	4.5	P, BR, DS, K, F	B, Railroad reform
Roman Abramovich	Millhouse/Sibneft, oil	169 (3.9)	203 (3.9)	12.5	S, BR, DS, K, H,[b] F	
Vladimir Kadannikov	AutoVAZ, automotive	167 (3.9)	112 (2.2)	0.8	BR, K	
Sergei Popov, Andrei Melnichenko, Dmitry Pumpiansky	MDM, coal, pipes, chemical	143 (3.3)	70 (1.4)	2.9	F	B, Financial markets (Mamut[b])
Vagit Alekperov	Lukoil, oil	137 (3.2)	475 (9.2)	5.6	S, P, BR, DS, K, F	
Alexei Mordashov	Severstal, steel, auto	122 (2.8)	78 (1.5)	4.5	BR, DS, F	B, Customs and WTO accession
Vladimir Potanin, Mikhail Prokhorov	Interros/Norilsk Nickel, non-ferrous metals	112 (2.6)	137 (2.6)	10.8	B, S, P, BR, DS, K, F	B, Social and labour relations (Eremeev[b])
Alexandr Abramov	Evrazholding, steel	101 (2.3)	52 (1.0)	2.4	F	B
Len Blavatnik, Victor Vekselberg	Access-Renova/ TNK-BP, oil, aluminum	94 (2.2)	121 (2.3)	9.4	DS, F	B
Mikhail Khodorkovsk[c]	Menatep/Yukos, oil	93 (2.2)	149 (2.9)	24.4	B, S, P, BR, DS K, H, F	B, International affairs
Iskander Makhmudov	UGMK, non-ferrous metals	75 (1.7)	33 (0.6)	2.1	K	
Vladimir Bogdanov	Surgutneftegaz, oil	65 (1.5)	163 (3.1)	2.2	P, BR, DS, K, F	
Victor Rashnikov	Magnitogorsk Steel, steel	57 (1.3)	57 (1.1)	1.3		
Igor Zyuzin	Mechel, steel, coal	54 (1.3)	31 (0.6)	1.1		
Vladimir Lisin	Novolipetsk Steel, steel	47 (1.1)	39 (0.8)	4.8	F	B
Zakhar Smushkin, Boris Zingarevich, Mikhail Zingarevich	IlimPulpEnterprises, pulp	42 (1.0)	20 (0.4)	1		
Shafagat Tahaudinov	Tatneft, oil	41 (1.0)	41 (0.8)	2.9		

(*continued*)

Oligarchs, Table 1 (continued)

Senior partner (s)	Holding company/ firm, major sector(s)	Employment, '000s (% sample)	Sales, in billions of roubles (% sample)	Wealth, in billions of US dollars	Other ranking[a]	RSPP bureau, head of committee/ taskforce (as of June 2004)
Mikhail Fridman	Alfa/TNK-BP, oil	38 (0.9)	107 (2.1)	5.2	B, S, P, BR, DS, K, F	B, Judiciary reform
Boris Ivanishvili	Metalloinvest, ore	36 (0.8)	15 (0.3)	8.8	P	B, Land reform (Kiselev[b])
Kakha Bendukidze	United Machinery, engineering	35 (0.8)	10 (0.2)	0.3	BR, K	B, Budget and taxes
Vladimir Yevtushenkov	Sistema/MTS, telecoms	20 (0.5)	27 (0.5)	2.1	S, P, BR, DS, K, F	B, Industrial policy, Pension reform (Yurgens[b])
David Yakobashvili, Mikhail Dubinin, Sergei Plastinin	WimmBillDann, dairy/juice	13 (0.3)	20 (0.4)	0.2		
Total		**1,831 (42.4)**	**2,026 (39.1)**			

Sources: Employment and sales are from World Bank (2004) and Guriev and Rachinsky (2005). The percentages in parentheses are the shares of employment/sales of the World Bank's sample, which in turn covers a substantial share of the economy (yet, as some industries are not represented, the list misses a couple of important candidates, such as Alexander Lebedev of National Reserve Corporation). Wealth is the market value of the oligarchs' stakes in spring 2004, calculated by authors using Forbes (2004) and stock market data. Wealth includes stakes of all the partners identified by the survey. Each entry lists the leading shareholder(s) in a respective business group, the name of the holding company or the flagship asset, and one or two major sectors. Several individuals per group are reported only when there is equal or near equal partnership. Ranking is based on employment in the sample and may therefore be different from the actual, as the sample disproportionally covers assets of different oligarchs. Employment and sales are based on official firm-level data for 2001. The exchange rate was 29 roubles to the US dollar. RSPP = Russian Union of Industrialists and Entrepreneur, the leading lobbying organization for Russian business. Among other things, RSPP represented the private sector in multiple meetings with President Putin, including the first one where the 2000 pact was allegedly concluded. B = RSPP Bureau membership (in total, the RSPP Bureau includes the President and 24 members); we also list the RSPP committees/ taskforces the particular oligarchs are in charge of (in total there are 17 committees/taskforces in the RSPP)

[a]Other oligarch rankings. B: Berezovsky's Group of Seven (*Financial Times* 1996). BR: Boone and Rodionov (2002). DS: Dynkin and Sokolov (2002). F: *Forbes* (2004). H: Hoffman (2003). K: *Kommersant* (2003). P: Pappe (2000). S: Classified as oligarchs in Freeland (2000, pp. xv–xvii)

[b]Some RSPP committee chairs have retired from active business. Eremeev was an Interros executive prior to the appointment at RSPP. Hoffman discusses Berezovsky rather than Abramovich. In 2000–3, Abramovich took over most of Berezovsky's assets in Russia as Berezovsky went into exile. Kiselev was Metalloinvest Board Chairman at the time of appointment at RSPP. Mamut was MDM Board Chairman at the time of appointment at RSPP. Yurgens was a Sistema executive prior to the appointment at RSPP

[c]Khodorkovsky remained a Bureau member and a Committee Chair for a while even after he was imprisoned, indicted and even convicted

any country in Continental Europe, where the share of the ten largest families is less than 35 per cent in small countries and less than 30 per cent in all large countries. In the United States and the United Kingdom, this share is in single-digit percentages. (A less rigorous approach is to look at the *Forbes*

billionaires lists. Even though Russian companies are significantly undervalued relative to their OECD counterparts, *Forbes*, 2004, lists 26 billionaires in Russia; only the United States and Germany have more. The 26 Russian billionaires are worth $81 billion, or 19 per cent of Russia's annual GDP. The 26 richest US citizens are worth four per cent of US GDP; the total wealth of all US billionaires is less than seven per cent of US GDP.) In the East Asian countries before the 1997 crisis, the highest shares of the ten largest families were in Indonesia (58 per cent), Philippines (52 per cent), Thailand (43 per cent) and Korea (37 per cent). The numbers for Indonesia and Philippines include the holdings of the Suharto and Marcos families, each controlling 17 per cent of total market capitalization in the respective countries. In Russia, the personal wealth of ex-President Yeltsin and President Putin is considered to be very modest.

What Do Oligarchs Control?

Each group in Table 1 controls assets in multiple provinces of Russia and even other countries, and in several industries. Mostly, the oligarchs' conglomerates are horizontally and vertically integrated. (Only Abramovitch, Deripaska, MDM group, and Potanin control major assets in unrelated industries, but even in their empires a single industry accounts for most of the conglomerate's value.) Oligarchs do dominate the largest industrial sectors, in particular natural resources (especially oil and metals) and automotive. The only large sectors not controlled by oligarchs are natural gas, energy, and manufacture of machinery. The gas and energy sectors are dominated by federally owned monopolies Gazprom and RAO UES; machinery production is a diverse sector which is populated by defence equipment suppliers (controlled by the federal government), oligarch firms and smaller firms controlled by non-oligarch private domestic owners.

Do oligarchs exercise excessive market power in the sectors that they control? The sectors controlled by oligarchs are indeed those with the highest concentration ratios in Russia (Guriev and Rachinsky 2005). However, these are also tradable goods sectors that are subject to global competition. For example, consider the ten sectors where oligarchs control more than 20 per cent of total sales. Except for ore and automotives, all these sectors sell to the global market: they export 30 to 90 per cent of their output; indeed, these sectors account for half of total Russian exports. The first exception, ore production, is mostly owned by oligarchs' vertically integrated conglomerates, where ore is an input. The second exception, the automotive sector, is a classic example of interest group politics. Russian cars are not internationally competitive, and the industry has always relied on protection. Such protection was usually granted, especially in the period in the 1990s when the largest carmaker's CEO, Vladimir Kadannikov, served as the first deputy prime minister in charge of economic policy. Yet, even with high import duties and support for domestic producers through generous tax write-offs and subsidies, import penetration was 25 per cent and rising. As of 2000, Oleg Deripaska consolidated his control over the second largest car producer and almost all of the bus and truck production, and the lobbying for stronger protection reached new heights. Indeed, one of the main reasons Russia is not yet a member of the World Trade Organization is that the WTO requires lowering import duties for cars, and Russia's automotive lobby launched an aggressive (and a very successful) anti-WTO campaign. The lobbyists managed to install increasingly high tariffs on both used and new imported cars.

The large industries where oligarchs play a large role are also those with substantial economies of scale. Indeed, these are exactly the sectors where large business empires originated in many countries in the late 19th century and the early 20th century, including the United States, Japan and Sweden. But, except for the automotive sector, there seems little reason for concern that Russia's oligarchs have excessive market power. Although their conglomerates are large by Russian standards, they are certainly not excessive by global standards. Some oligarchs are important global players in their industries (especially in oil and metals), but none is a dominant market leader. Thus, there is no basis, on efficiency

grounds, for antitrust policies aimed at breaking up the oligarchs' companies. Instead, it is more important that Russian competition policy assure a level playing field for all owners without regard to their size and political influence.

How Did the Oligarchs Gain Control?

A common belief is that the oligarchs owe their fortunes to the 'loans-for-shares' auctions held in mid-1990s, which are widely regarded as the most scandalous episode of Russian privatization. In the classical loans-for-shares scenario, the government appointed a commercial banker to run an auction that would allocate a controlling stake of a large natural resource enterprise in exchange for a loan to the federal government that the latter never intended to repay. Not surprisingly, the auctioneer always awarded the stake to himself for a nominal bid (usually, slightly above a very low reserve price) by excluding all outside bidders. The scheme was designed to consolidate the bankers' support for Yeltsin's re-election campaign in 1996.

The conventional loans-for-shares story fits Abramovich (in 1995–7, a junior partner of Berezovsky), Khodorkovsky, and especially Potanin. The other two winners were the oil sector insiders Alekperov and Bogdanov, who obtained stakes in firms they already controlled. However, most of those listed in Table 1 did not become oligarchs through the loans-for-shares programme. Some of the 22 largest owners tried to participate in the loans-for-shares programme and even offered more competitive bids, but were excluded by those in charge of respective auctions; some even raised their concerns in public.

Most of the individuals listed in Table 1 are relatively young: nine of them are in their thirties, and 13 are in their forties. (Both mean and median individuals in Table 1 are 44 years old. Russian oligarchs are much younger than their American counterparts. In the *Forbes* 2004, list, the average age of the 25 richest Americans is 64 years; the average age of all 262 US billionaires is the same.) The older oligarchs have typically come from Soviet-era nomenklatura. Prior to transition, they were either managing their respective enterprises or working in government agencies supervising those enterprises. When Soviet-era enterprises were privatized, they successfully converted their de facto control into ownership rights. The younger entrepreneurs started from scratch in the late 1980s, building their initial wealth during President Gorbachev's partial reforms when the coexistence of regulated and quasi-market prices created huge opportunities for arbitrage. In 1992, as price liberalization and privatization began, most of them owned trading companies and/or banks. Thus, when privatization of industrial enterprises occurred, they had the financial capital available to purchase ownership in privatization auctions. Some of these entrepreneurs were neither industry nor government insiders; yet, they converted Soviet manufacturing enterprises into successful modern capitalist firms. Of course, a cynic might note that such companies are near the bottom of the list in Table 1 in terms of size, while the loans-for-shares winners dominate the top of the list.

Oligarchs' Dilemmas

Whatever the source of individual oligarchs' wealth, the Russian public still deems it illegitimate, believing that the oligarchs obtained their initial wealth through connections and furthered it by securing preferential treatment through exerting political influence. (In a July 2003 poll by ROMIR, an independent Russian research and polling agency, 88 per cent responded that all large fortunes were amassed in an illegal way, 77 per cent said that privatization results should be partially or fully reconsidered, and 57 per cent agreed that the government should launch criminal investigations against the wealthy; *Vedomosti* 2003.) This has created a fundamental problem for Russia's transition: promoting democratic values (that is, respecting the median voter's opinion) may undermine liberal values (private property rights in a substantial part of the economy). This conflict has created a window of opportunity for such a pragmatic politician as President Vladimir Putin, who has managed to play oligarchs and voters off against each other to consolidate his own political power.

Curbing the oligarchs' political influence was an essential part of Vladimir Putin's presidential campaign in 2000. In his open letter to voters, he promised to treat the oligarchs in the same way as other entrepreneurs; a few days later he announced that all interest groups would be kept at an 'equal distance' from his government. In the first meeting with the leading oligarchs on 28 July 2000, President Putin offered them the following pact. As long as the oligarchs paid taxes and did not use their political power (at least not against Putin), Putin would respect their property rights and refrain from revisiting privatization. This pact defined the ground rules of oligarchs' interaction with central and regional government during Putin's first term (2000–4). Although the pact could have never been written, even the general public was well aware of its existence. A poll by FOM (2000), an independent non-profit Russian polling organization, a week after the meeting showed that 57 per cent Russians knew about it.

Putin's threat to prosecute any oligarch who deviated from the pact was based on the median voter's support for expropriating the oligarchs. Putin carried out his threat in 2003, when the prominent oligarch Mikhail Khodorkovsky, the majority owner of the Yukos oil company, deviated from the pact by openly criticizing corruption in Putin's administration and supporting opposition parties and independent media. He and his partners were soon arrested or forced into exile, and their stakes in Yukos expropriated. It is not clear why Khodorkovsky did not stick to the pact. Perhaps he thought that supporting opposition parties rather than challenging Putin himself was not a violation. Almost certainly, he did not expect Putin to respond so decisively.

The expropriation of the Yukos shareholders certainly involved serious costs for Russian economy – the investment climate worsened and capital flight increased substantially. However, Putin clearly demonstrated that his priority was to establish his credibility even if this damaged his economic agenda. The Yukos affair has clarified the rules of the game between oligarchs and the Kremlin. Oligarchs have learned the risks associated with violating the pact, and so in the future they will be less likely to interfere in national politics. The Yukos affair effectively shifted the bargaining power from oligarchs to bureaucrats. Although outright expropriation of oligarchs will probably remain just a threat, their cash flows will be milked more intensively by bureaucrats in the form of kickbacks, donations to pet projects, and direct bribes (for a discussion of this 'contract' between bureaucrats and the entrepreneurs as a 'viability insurance contract', see Ickes 2005). This will in turn undermine oligarchs' property rights and incentives to invest. To sustain economic growth, Putin has to constrain rent-seeking by his own bureaucrats. This task is certainly not an easy one, given that democratic checks and balances are very weak. Moreover, neither government nor the oligarchs are interested in the development of democracy and civil society. (Actually, oligarchs may also benefit from imperfect property rights protection as there are economies of scale in private rent-seeking; see Glaeser et al. 2003; Rajan and Zingales 2003; Sonin 2003.) Bureaucrats do not like to cede their control, while oligarchs are afraid of the median voter's redistributive agenda.

The potential exit strategy for any individual Russian oligarch is to sell a large stake to a reputable foreign investor. Indeed, expropriating foreigners is harder for the state because they are more popular than oligarchs, and because of pressures from foreign governments. However, timing the exit properly is a complex problem. Selling too early would bring too little as the assets are initially undervalued. Delaying the sale in order to restructure the company and improve its transparency would raise the price, but would also increase the risk of expropriation by the Russian government. This expropriation may also occur through a seemingly market-based transaction. For example, the government can use public funds to pay the oligarch the market value of his assets in exchange for (hidden) substantial side payments to selected government officials or their pet projects. Given the threat of complete expropriation, this is an offer the oligarch cannot refuse.

Economic Performance of the Oligarchs

Do oligarchs create value or strip assets? Do they improve the performance of the firms they control or injure their performance?

Most oligarchic groups are horizontally or vertically integrated and are run by active majority owners, so the usual 'conglomerate discount' diseconomies of scale are unlikely to apply. A more important problem is, of course, the political risk of expropriation that shortens time horizons and reduces the incentives to invest.

On the other side, several arguments suggest that Russia's oligarchs might improve firm performance. First, the oligarchs' performance might be superior because they have successfully overcome the separation of ownership and control. An oligarch who owns a very large majority share should have strong incentives to restructure companies and to seek to improve the value of this asset, rather than for diverting cash flows and stripping the assets. Even if a firm was originally privatized to dispersed shareholders, its ownership structure was quickly consolidated through dilution and, in some cases, outright expropriation of outside investors, including government and foreigners. The current champions of transparency, Mikhail Khodorkovsky and Vladimir Potanin (now chairing Russia's National Council for Corporate Governance), kept expropriating outside investors until as recently as 1999. In our sample, oligarchs do control large stakes in their firms. In an average firm where the largest owner is the oligarch, he controls 79 per cent; in the case of non-oligarch private domestic owners, the corresponding figure is only 74 per cent. The difference is statistically significant but not necessarily economically important. The average degree of control exercised by smaller owners over their companies is also very high. Poor protection of minority shareholders rights has resulted in consolidation of control within most Russian companies. As a result, smaller owners are not investors that hold small stakes in large companies; rather, they hold large stakes in small companies.

Second, vertical integration can mitigate the risk of hold-up problems, where in a situation of relatively few buyers and sellers each party must be concerned that the other will attempt to renegotiate and seize a greater share of the joint surplus. Many oligarch empires have been built to overcome such hold-up problems: for example, all Russian major oil companies are vertically integrated; most steel producers own sources of coal and ore; some companies own ports, fleets of railroad cars and even railroad track. Third, in a situation with underdeveloped financial markets, external finance is costly; larger oligarch-run firms can benefit from their access to internal finance. They can create an internal financial market to finance expansion (see Khanna and Yafeh 2005, for the discussion of these two benefits for business groups in developing countries). Fourth, Russia lacks a clear rule of law, and the larger conglomerates are certainly more effective than small firms in influencing judicial and political decisions and protecting their property from the predatory 'grabbing hand' of federal and local governments.

There is still no convincing test of whether and how oligarchs affect the performance of their firms. Constructing such a test is a significant challenge. Preliminary results (Guriev and Rachinsky 2005) show that in terms of total factor productivity growth (with industry, region and size controlled for) oligarchs' firms do perform almost as well as foreign firms and better than other Russian-owned firms. Yet more empirical work is needed to control for endogeneity of oligarch ownership, and to study the long-term effects. In addition, more work is needed to produce a quantitative evaluation of the oligarchs' effect on social welfare.

Oligarchs and Russia's Future

While ownership concentration in Russia is higher than in other countries today, it does not seem unprecedented in historical perspective. Owners of Korean chaebols, Japanese zaibatsu, Sweden's and Italy's largest family controlled firms, and US 'robber barons' exercised a similar share of economic and political power. Also, in many of these countries the oligarchs' wealth was accumulated with substantial support from the state (in direct subsidies, tax breaks, land grants,

subsidized credit, and so forth) and was deemed illegitimate by a substantial share of the public at some points in history. Yet these countries have managed to build functioning market economies, although it took much longer for some of them to create functioning democracies. Therefore, it is not clear whether and how soon Russia will succeed in establishing legitimacy of private property rights and whether this will be accompanied by a transition to a sustainable democracy.

See Also

- ▶ Inequality (International Evidence))
- ▶ Policy Reform, Political Economy of
- ▶ Privatization
- ▶ Rent Seeking
- ▶ State Capture and Corruption in Transition Economies
- ▶ Transition and Institutions

Acknowledgment *This article draws substantially on* Guriev and Rachinsky (2005) *and mostly refers to the situation in Russia prior to the renationalization campaign that started in 2004.*

Bibliography

Aslund, A. 2006. *Revolution in orange: The origins of Ukraine's democratic breakthrough*. Washington, DC: Carnegie Endowment for International Peace.

Boone, P., and D. Rodionov. 2002. *Rent seeking in Russia and the CIS*. Moscow: Brunswick UBS Warburg.

Dynkin, A., and A. Sokolov. 2002. Integrated business groups in the Russian economy [in Russian]. *Voprosy Ekonomiki* 4: 78–95.

Financial Times. 1996. Moscow's Group of Seven. 1 November, p. 17.

FOM (Fond Obschestvennogo Mnenia [Public Opinion Foundation]). 2000. *Government and Large Business: A Poll of 1500 Russian Citizens Held on 5 August 2000* [in Russian]. Online. Available at http://bd.fom.ru/report/cat/societas/market_economy/economic_reform/private_enterprise/power_and_business/d001621. Accessed 16 June 2006.

Forbes. 2002–6. The world's billionaires. Online. Available at http://www.forbes. com/billionaires, consulted 16 June 2006.

Freeland, C. 2000. *Sale of the century: Russia's wild ride from communism to capitalism*. New York: Crown Business.

Glaeser, E., J. Scheinkman, and A. Shleifer. 2003. *Journal of Monetary Economics* 50, 199–222.

Gorodnichenko, Y., and Y. Grygorenko. 2005. *Are oligarchs productive? Theory and evidence*. Mimeo: University of Michigan.

Guriev, S., and A. Rachinsky. 2005. The role of oligarchs in Russian capitalism. *Journal of Economic Perspectives* 19(1): 131–150.

Hoffman, D. 2003. *The oligarchs: Wealth and power in the new Russia*. New York: Public Affairs.

Ickes, B. 2005. Economic pathology and comparative economics: Why economies fail to succeed. Presidential address. *Comparative Economic Studies* 47: 503–519.

Khanna, T., and Y. Yafeh. 2005. *Business groups in emerging markets: Paragons or parasites?* Discussion Paper No. 5208. London: CEPR.

Kommersant. 2003. *Who owns Russia?* [in Russian]. Moscow: Vagrius.

London, J. 1908. *The iron heel*. London/New York: Macmillan.

Pappe, Y. 2000. *The oligarchs*. Moscow: Higher School of Economics.

Rajan, R., and L. Zingales. 2003. *Saving capitalism from the capitalists: Unleashing the power of financial markets to create wealth and spread opportunity*. New York: Crown Business.

Sonin, K. 2003. Why the rich may favor poor protection of property rights. *Journal of Comparative Economics* 31: 715–731.

Vedomosti. 2003. Take away and divide: People's aspirations have not changed in 86 years [in Russian]. 18 July.

World Bank. 2004. *From transition to development: A country economic memorandum for the Russian Federation*. Moscow: World Bank.

Oligopoly

P. Sylos-Labini

Keywords

Advertising; Barriers to entry; Competition; Concentration; Differentiated oligopoly; Duopoly; Full cost principle; Great depression; Imperfect competition; Increasing returns; Innovation; Market power; Mixed oligopoly; Monopoly; Obstacles to entry; Okun's Law; Oligopoly; Price determination; Price leaders; Price rigidity; Price variation; Price wars; Productivity growth; Returns to scale; Specialization economies; Sraffian economics; Technical change; Verdoorn's Law

JEL Classifications
D4

No article entitled 'oligopoly' appeared in any edition of Palgrave's *Dictionary of Political Economy*. It is true that the simplest case of oligopoly, that is, duopoly, was considered more than a century and a half ago, by Cournot; but such an analysis was motivated by purely theoretical interests. The fact is that only in the 20th century and especially after the Second World War did this market form become important in economic reality, as a result of two processes of economic change: the process of concentration and the process of differentiation. In those branches where the former process has asserted itself – for example, steel, basic chemical products, cement, electricity – concentrated oligopoly with relatively homogeneous products has emerged; where the latter process has prevailed, we find differentiated oligopoly; in those branches where both processes have taken place simultaneously, then mixed oligopoly has emerged. In both processes innovations have played a major role, with the proviso that in the process of concentration innovations have given rise to economies of scale, whereas in the process of differentiation the most important role has been that of technological innovations implying economies of specialization; in this case, technological innovations are combined with commercial innovations. In fact, differentiated oligopoly can be found mainly in those activities in which quality competition, commercial services and advertising have had a particularly relevant role – non-durable consumer goods, such as textiles, tyres, canned foods, soft drinks and cigarettes are often produced in conditions of differentiated oligopoly.

In the past, when the standard of living of the masses of consumers was not much above the subsistence level, there was not much scope for the factors just mentioned. With the gradual increase of per capita income, consumers' preferences have acquired an increasing space. At the same time the possibility of advertising has been greatly enhanced by particular innovations – modern means of transportation and the so-called mass media, among which radio and television play a special role. Mixed oligopoly (concentration cum differentiation) is typical of several industries producing consumer durables such as automobiles, typewriters, refrigerators, radio and television sets, computers; mixed oligopoly can be found in several important service sectors such as banking and insurance. In addition a large number of non-durable consumers' goods and services – including commercial services – constitute the area where differentiated oligopoly prevails; it is well to point out that as a rule there is no difference between imperfect competition and differentiated oligopoly. Analytically, the former can be seen, as a rule, as a first and the latter as a second approximation; this standpoint becomes natural if we recognize that the imperfect markets are composed by a 'chain of oligopolistic groups'.

After careful reflection, we are bound to admit that in modern industry and in services, oligopoly, in its three varieties, is the rule and competition the exception –to be found in certain industries producing sufficiently homogeneous non-durable goods and in subsidiary activities. Competition, on the other hand, is the rule in most agricultural and mineral raw materials traded in international markets.

According to the traditional (neoclassical) conception, markets in competitive conditions are formed by a great number of firms, each of which is so small as to be unable to influence prices. Each firm, then, is bound to accept the market price and pushes output up to the point at which marginal cost – which, after a point, cannot but be increasing – equals price. In fact, the increasing marginal cost, that is, diminishing returns both in the short and in the long run are a necessary feature of traditional theory. In monopoly equilibrium is reached when the decreasing marginal income equals marginal cost. Indeed, according to that theory, only two market forms are worth consideration – competition and monopoly – the former being the rule, the second the exception (The analytical tools to be used for imperfect competition are those worked out for monopoly).

The whole analysis is statical and thus presupposes given technology. To work out theoretical

models consistent with dynamic analysis, we have to go back to the classical concept of competition, where freedom of entry and not the number and size of firms is crucial. If we adopt this concept, it is easy to shift from competition to non-competitive market forms, by considering obstacles of various relevance to entry. Clearly, when in a given market the obstacles to entry are serious, firms operating in that market are likely to be few; this, however, is to be seen, not as a preliminary datum, but as the likely (not necessary) result of the existence of those obstacles.

When the obstacles to entry are of little importance, then a super-normal profit will attract new firms: supply will increase and the price will fall, so that supernormal profit will tend to disappear: such a profit can persist when obstacles to entry are important.

Having chosen this approach, in a first approximation we have to distinguish, in price analysis, between agriculture and mining, on the one hand, where obstacles to entry as a rule are modest, and industry and services, on the other, where those obstacles are often considerable. Again, in the first approximation, we can state, with Ricardo, that in primary activities in the short run prices depend on demand and supply, whereas in the long run they depend on costs. If we refer to the short run and intend to work out an analysis susceptible of empirical verification, we realize that 'demand' can be variously interpreted; in the case of raw materials traded in the international markets, demand can best be represented by an index of world industrial production. In industry and services, instead, in the short run prices depend principally on changes in direct costs and, in the long run, on changes in total costs per unit.

The reason for this sharp difference as regards short-run variations of prices is as follows. In primary activities firms, owing to the relative freedom of entry, have no outstanding market power and cannot influence prices, which vary according to the variations of aggregate demand and aggregate supply. In the other activities, however, prices are to a non-negligible extent controlled by firms and, in particular, by those that act as price leaders. Starting from a price that is accepted by all firms – that is, from an 'equilibrium price' – the firms acting as leaders will modify it when the conditions of equilibrium change. There are, then, two analytical problems, conceptually different but strictly interrelated: the problem of price determination and that of price variations. In traditional terms, the former problem belongs to the area of statical analysis, the latter to that of dynamics. We can accept such a distinction provided that it implies no cleavage, that is, provided that we can pass without discontinuities from the analysis of price determination to that of price variations.

The problem of price determination implies the analysis of the equilibrium, which includes: the size of the market (that is, the position in a Cartesian diagram of the demand curve, a concept that becomes relevant when firms are no more conceived as atoms); the shape of the demand curve (that is, the elasticity of demand); technology, salaries and other administrative expenses; taxes; and the prices of durable and those of variable means of production. This is not the place to present a formal solution of the problem of price determination. Suffice it to say that the concept of entry-preventing price and elimination price are important analytical tools to be used in the construction of a theoretical model of price determination. Once the price reaches the level acceptable to all firms – the equilibrium level – each firm is in a position to calculate the markup, that is, the ratio between price and cost or, more precisely, direct cost. When the equilibrium conditions change, the price is to be changed. Normally this occurs without a price war, since such wars are costly and major firms are willing to undertake them if only the expected gains (net of risks) are higher than expected costs, an occurrence that does not appear to be frequent.

The analytical steps, then, are two: the first is to understand how the equilibrium price is arrived at; the second is to understand how it varies when the equilibrium conditions change. If in the first step the concepts of entry-preventing and elimination prices are essential, in the second step it is the 'full cost principle' that plays the key role. Empirical enquiries have consistently shown that this principle is generally followed by managers operating in non-agricultural activities. Yet for a long period it has been considered only as a rough rule of

thumb, without theoretical relevance. Probably the reasons are twofold. The first is that it contradicts the received doctrine, which is founded on marginal analysis and which, as a condition of equilibrium, assumes a rising marginal cost – the full cost principle, instead, which is based on the markup on direct costs, assumes the marginal cost to be constant and therefore equal to direct cost. The second reason is that that principle has been described as if it were a criterion to determine the price, not to modify it, but it can have a meaning only in the second case. Thus, Hall and Hitch (1939), in their pioneer empirical enquiry, report that 'prime (or "direct") cost per unit is taken as the base, a percentage addition is made to cover overheads . . . and a further conventional addition . . . is made for profit'. However, it is evident from this statement that the crucial theoretical problem is to explain the height of the two percentage additions – that can be unified into one percentage. Thus, given the cost elements, we have to explain the conditions that limit the discretionary powers of managers in choosing a given percentage and not another, that is, we have to explain the equilibrium conditions. Only after having explained the equilibrium price can the markup acquire a meaning. In other words, the full cost principle is theoretically meaningless as regards the problem of price determination and becomes meaningful as regards the problem of price variations: in fact, barring price wars, the markup appears to be the quickest and most rational way for firms, and particularly for price leaders, to arrive at a new equilibrium price when the equilibrium conditions vary.

The further question is to understand why direct cost and not total unit cost is taken as the term of reference to modify the price in the short run – say, year by year or even in shorter periods. The reason is that the changes in the prices of variable factors affect without much delay all firms, though not necessarily in the same proportions, whereas the changes in the other equilibrium conditions – size of the market, elasticity of demand, technology, salaries and other overhead costs – affect the firms at different degrees and in different times. These changes either affect prices in relatively long periods or do not affect them at all – substantial increases in overhead costs can be offset, not by price increases, but through productivity increases. To be sure, when these are insufficient, the increases in overhead costs can push some of the firms out of the market; this can also be the outcome of unfavourable changes in market conditions.

Changes in direct costs, then, tend to be shifted to prices in the short run. But even for this category of changes a sort of hierarchy is necessary: changes in the prices of raw materials (including the sources of energy) tend to be fully shifted on prices of finished products in both directions, since those changes tend very quickly to affect all firms. This is not so for changes in wage cost per unit, since this cost is given by the ratio between wages and productivity. Now, wage changes – if we except the areas of the so-called submerged economy – affect in a relatively short run all firms, whereas productivity increases due to organizational innovations and to technological changes determined by previous investment tend to take place at different rates in the different firms (declines in productivity are exceptional): only those changes in wage cost per unit of output have to be shifted onto prices that are common in both the upward and the downward direction. However, under contemporary conditions the shift in the downward direction will be more limited than that in the upward direction, since it is unlikely that the prices of finished industrial products in international markets will generally decrease; and it is international competition that, in industry, will limit the market power of the firms of a given country. Briefly, in the short run, the shift of changes in total direct costs will tend to be not only partial but also asymmetrical.

In the case of industrial products, then, short-run variations in prices depend on the variations of direct costs: demand does affect prices, but, as a rule, only in the long run and not in the same direction, as is the case in the short-run variations of prices under competitive conditions, but in the opposite direction, since the long-run expansion of demand makes the entry of new firms easier and opens the possibility of exploiting economies of scale. Thus, an expansion of demand tends, *ceteris paribus*, to reduce and not to raise the

price. In the short run demand increases have no significant direct effect on prices of industrial goods; they can have an indirect effect, that is, via the prices of raw materials, when demand pressure is so strong as to affect not only finished products but also raw materials. As for finished products, demand pressure tend to affect not prices but (consistently with the Keynesian conception) the level of activity.

If we pass from partial to general analysis and adopt the framework of a Sraffian model, we are bound to distinguish between basic and non-basic ('luxury') products. If we decide to consider not only competitive but also non-competitive markets, we have to drop either the assumption of a unique rate of profit or the assumption of a unique wage rate (for a given type of labour). In any case, prices enter into the conditions of simple reproduction. The conditions of expanded reproduction, that is, of accumulation – to use the Marxian expression – imply, in addition, that at least a share of the surplus be employed productively, that is, invested – the velocity of accumulation being determined by that basic product that has got the lowest surplus. It is important to point out that technological progress is essential not only in the case of accumulation but also in the case of simple reproduction, since mineral products tend gradually to exhaust themselves; it is essential also in the case of a growth proportional to the increase of population, not only due to the reason just mentioned, but also due to the necessity of offsetting the tendency of diminishing returns in agriculture.

If we adopt a Sraffian model of general analysis, the study of the effects of technological changes meets with several problems, certainly serious, but, in principle, not insurmountable; in fact, some important steps in this direction have already been made by Sraffa himself. That study, instead, seems to be precluded if we adopt a Walrasian model of general equilibrium that implies a strictly static framework, in which all firms operate in conditions of diminishing returns, that is, of increasing marginal costs. Now, barring special cases, increasing returns are to be related to changes in the methods of production, even in the short run: increases in the productivity of labour can take place as a consequence of quick readjustments of the labour force and of innovating investment carried out in previous periods. A long series of empirical observations – among which may be mentioned Dunlop's 1938 article on the movement of real wages, the 'Verdoorn Law' and 'Okun's Law' – show that increasing, not diminishing, returns dominate modern economies and, in particular, non-agricultural activities. Thus, to admit that it is not perfect but imperfect competition and oligopoly that is the rule seems to be the only way to reconcile theoretical models and empirical enquires in both partial and general analysis.

In the short run technical progress takes mainly the form of increases in productivity of means of production and, in particular of labour; in the long run one has also to consider the production of new goods, that in the short run represent a tiny fraction of the total. The diversification of output, which in fact conditions the growth of all firms, can assume either a prevailingly commercial character, in the case where the goods are already in the market, or also a technological character, if the goods or the process through which they are produced are new. In its turn, the expansion of demand represents the condition for the introduction of two important types of technological innovations – that is, new goods and new processes implying the exploitation of economies of scale – which, after all, is nothing but another way to re-propose the Smithian proposition according to which 'the division of labour is limited by the extent of the market'.

For the sake of simplicity, we limit ourselves to considering, as the index of technical progress both for the short and the long run, the increase in productivity of labour. The basic consequence of this increase is, at the aggregate level, a systematic divergence between the average variations of nominal incomes and the average variations of prices, with the fall of the relative prices of those goods produced in the most dynamic industries. Referring to average variations, the said divergence can take four different forms:

	Nominal incomes	Prices
(a)	Falling	Falling more rapidly
(b)	Constant	Falling

(*continued*)

	Nominal incomes	Prices
(c)	Rising	Constant
(d)	Rising	Rising more slowly

Cases of falling prices – (*a*) and (*b*) – were frequent in the 19th century, when the process of concentration and that differentiation in industry and services had not proceeded far enough and competition was still the rule in those sectors. In the 20th century case (*a*) occurred during the first four years of the Great Depression; but, in sharp contrast to what was normally occurring in the 19th century, the level of activity in industry and services fell much more than prices, whereas in agriculture the prices fell violently, but the level of activity remained approximately constant. The comparison with the great depression of the 19th century – which occurred in the years 1873–9 – is illuminating.

Putting aside services, which offer a picture similar to that of industry, the percentage changes in prices and production of agriculture and industry during the two great depressions (I and II) were as follows:

		United Kingdom		United States	
		Prices	Production	Prices	Production
Agriculture	I	−18	+3	+31	+4
	II	−44	0	−54	+2
Industry	I	−29	−5	−33	−5
	II	−21	−16	−23	−48

If we except the period of the great depression of the 20th century, which was in all senses an exceptional event, with productivity varying in a very irregular fashion, in the 20th century cases (*c*) and (*d*) – rising nominal incomes with constant prices or prices rising more closely – were the rule. Now, it is not indifferent that the fruits of technical progress have one type of consequence or the other on prices and incomes.

When prices of all goods fall, the means of production (Sraffa's basic products) become cheaper and this stimulates the expansion of all firms, including those that do not introduce innovations. On the other hand, when prices fall demand increases automatically in real terms.

Let us now consider what happens when productivity rises but prices do not fall. If, in such circumstances, nominal incomes do not rise, the whole increase in productivity tends to translate itself into a decreasing level of employment; to have at least a stable level of employment, nominal incomes should rise in proportion to the increase in productivity; and this is not an automatic process. It is unlikely that wages and salaries rise if there is not a systematic action of trade unions, unless the process of differentiation and the consequent increasing fragmentation in the labour market have become so widespread as to favour wage increases even without a generalized pressure of trade unions. On their side, non-labour incomes will increase only if investment or government expenditure increases, or both. Investment can increase only if new investment opportunities arise, due to technical innovations, whereas government expenditure can increase as a political decision. On the other hand, with stable prices, the firms that do not introduce innovations cannot receive the stimulus arising from the means of production becoming cheaper. As a result, the process of growth tends to become more and more unbalanced, unless a general expansion of demand – originated by innovations and/or by government – takes the place of the stimulus afforded by an overall fall in prices.

In short, owing to the obstacles to entry, in most non-agriculture activities the 'competitive mechanism' for the distribution of the fruits of technical progress (falling prices, stable nominal incomes) has been more and more substituted by the 'oligopolistic mechanism' (stable prices and increasing nominal incomes). In the new conditions, the process of growth requires increasing intervention of public powers, but not necessarily in the form of increasing public expenditure. That intervention can consist of taxation (to afford incentives or to put brakes), or can support the prices and the incomes of those activities, like agriculture, least affected by those two processes, or can promote the source of technological innovation, that is, scientific research, or – to give another important example – can create conditions favourable to development of small firms, not only with fiscal and credit incentives, but also by supplying real services – especially commercial and technical assistance. All these measures

of public powers can push up the growth of the volume of investment to the velocity required to avoid an increase of unemployment or gradually to reduce it to the frictional level.

If the countervailing influences of public interventions process are not strong enough, in the new conditions the process of growth tends to become more unbalanced not only from the standpoint of the different industries (since those that do not carry out innovations directly have no more the stimulus determined by the declining prices of the means of production they use), but also from the point of view of income distribution. In fact, the downward price rigidity tends to create special margins in certain industries or in certain firms. These rising margins do not necessarily become above-normal profits; they can become, too, above-normal wages or salaries, depending on the relative strength of the opposing parties. Instead of the above-normal incomes, the advantages for workers can also take the form of a greater stability of employment; similarly, the advantages of capitalists can take the form, rather than of above-normal profits, of more stable profits.

It seems that in recent times the process of differentiation has become more important than the process of concentration and the economies of specialization seem to have become more important than the economies of scale. This new development in industry has been promoted by at least three changes: (1) the growth of electronics and allied industries; (2) the reaction of increasing masses of workers in advanced countries against the monotony of assembly lines and other methods of mass production; (3) the growing differentiation in consumer preferences originated by the increasing per capita income. In services, differentiation has always been important and in recent times has become even more important; at the same time, services become the most important section of the economy in the so-called post-industrial societies. Considering the declining relative weight of agriculture and mining in advanced societies, we have to conclude that the area of flexible prices tends to shrink and that of rigid prices to expand – I mean flexibility or rigidity in the downward direction. In particular, the area of rigid prices tending to expand refers more and more to services and less and less to industry; this phenomenon, that has important consequences also on the overall behaviour of prices, up to now has received very little attention.

It remains true, however, that the increasing rigidity of prices of goods and services determines the need for an increase in demand large enough to avoid a decline in employment, if population grows. Now, with the diffusion of high education, with the space for a rapidly increasing number of goods opened up by the increasing per capita income, in recent times the potentialities of development have increased. But such potentialities can remain unexploited if they are left to spontaneous market forces; given the rate of interest, all depends on investment stimulated by technological innovations that promise to be profitable and that can be devised and carried out by private firms without the support or the stimulus afforded by public powers. If those investments are not enough to promote an increase of demand capable of generating an increase in income at least equal to that in productivity, unemployment gradually grows. It is well to emphasize that the main obstacles to a policy of economic growth arise not by diminishing returns, but either from the side of the public deficit, if the increasing supply of bonds pushes up the rate of interest; or from the side of the foreign deficit, which pushes up the value of foreign currencies, giving rise to a special kind of inflationary pressure. Such problems are aggravated by the fact that the two deficits, to some extent, reinforce each other: for instance, large firms tend to borrow abroad, owing to the high internal rate of interest. But these are matters that go beyond the limits of our theme.

See Also

▶ Game Theory

Bibliography

Andrews, P.W.S. 1949. *Manufacturing business*. London: Macmillan.

Bain, J.S. 1956. *Barriers to new competition*. Cambridge, MA: Harvard University Press.

Baumol, W.J. 1967. *Business behavior, value and growth.* Rev. ed. New York: Harcourt Brace and World.
Bhagwati, J.N. 1970. Oligopoly theory, entry-prevention, and growth. *Oxford Economic Papers* 22: 297–301.
Chamberlin, E.H. 1933. *The theory of monopolistic competition: A reorientation of the theory of value*. Cambridge, MA: Harvard University Press.
Dunlop, J. 1938. The movement of real and money wage rates. *Economic Journal* 48: 413–434.
Eichner, A.S. 1976. *The megacorp and oligopoly: Microfoundations of macrodynamics*. Cambridge: Cambridge University Press.
Fellner, W. 1949. *Competition among the few: Oligopoly and similar market structures*. New York: Knopf.
Galbraith, J. 1957. Market structure and stabilization policy. *Review of Economic and Statistics* 39: 124–133.
Hall, R.L., and C.J. Hitch. 1939. Price theory and business behaviour. *Oxford Economic Papers* 2: 12–45.
Kaldor, N. 1935. Market imperfection and excess capacity. *Economica* NS 2: 33–50.
Kalecki, M. 1943. Costs and prices. Repr. in M. Kalecki, *Selected essays on the dynamics of the capitalist economy 1933–1970*. Cambridge: Cambridge University Press, 1971.
Modigliani, F. 1958. New developments on the oligopoly front. *Journal of Political Economy* 66: 215–232.
Robinson, J. 1933. *The economics of imperfect competition*. London: Macmillan; New ed, 1969.
Sraffa, P. 1960. *Production of commodities by means of commodities*. Cambridge: Cambridge University Press.
Sylos-Labini, P. 1956. *Oligopoly and technical progress*. Trans., Cambridge, MA: Harvard University Press, 1969.
Sylos-Labini, P. 1984. *The forces of economic growth and decline*. Cambridge, MA: MIT Press.

Oligopoly and Game Theory

Hugo Sonnenschein

Oligopoly theory is concerned with market structures in which the actions of individual firms affect and are affected by the actions of other firms. Unlike the polar cases of perfect competition and monopoly, strategic issues are fundamental to the study of such markets. In this entry we will explain some of the central themes of oligopoly theory, both modern and classical, and emphasize the connection between these themes and developments in the noncooperative theory of games.

Section 1 presents a simple static oligopoly model and uses it to discuss the classic solutions of Cournot (1838), Bertrand (1883) and Stackelberg (1934). Section 2 contains an introductory account of a modern line of research into a class of dynamic oligopoly models. In these models, firms and consumers meet repeatedly under identical circumstances. An example is presented to illustrate the important result that a firm's behaviour in such situations can drastically differ from that in the static model. Section 3 is concerned with a 'folk theorem' which states that with free entry, and when firms are small relative to the market, the market outcome approximates the result of perfect competition. Novshek's Theorem (1980) gives a precise statement of this result, and in doing so provides an important bridge between oligopoly theory and the theory of perfect competition.

I.

We consider a market in which n firms ($n > 1$) produce a single homogeneous product. The quantity of output produced by the ith firm is denoted by q_i and the cost associated with production of q_i by $C_i(q_i)$. Demand is specified by an inverse demand function $F(\cdot)$: $F(Q)$ is the price when $Q\ (=\Sigma_i q_i)$ is the aggregate output of firms. Let q and q_{-i} denote the vectors $(q_1,\ldots, q_n)$ and $(q_1,\ldots, q_{i-1}, q_{i+1},\ldots, q_n)$ respectively. The profit of the ith firm is given by

$$\Pi_i(q) = F(Q)q_i - C_i(q_i).$$

The interdependence of firms' actions is reflected in the fact that the profits of the ith firm depend not only on its own quantity decision but also on the quantity decisions of all other firms.

A Cournot equilibrium is an output vector $\overline{q} = (\overline{q}_1,\ldots \overline{q}_n)$ such that

$$\forall i, \qquad \forall q_i, \qquad \Pi_i\ (\overline{q}) \geq \Pi_i\ (q_i, \overline{q}_{-i})$$

where $(q_i, \overline{q}_{-i})$ denotes the vector $(\overline{q}_1,\ldots, \overline{q}_{i-1}, q_i, \overline{q}_{i+1},\ldots, q_n)$. The equilibrium $\overline{q}$ is symmetric if $\overline{q}_1 = \cdots = \overline{q}_n$.

In the Cournot model, firms make quantity decisions. A single homogeneous good is

produced, which all firms sell at the same price. At equilibrium, no firm can increase its profit by a unilateral decision to alter its action. A Cournot equilibrium is illustrated in the following example.

Example 1: Let n firms have identical linear cost functions: $C_i(q_i) = cq_i$ for all i. Assume that the inverse demand function is linear: $F(Q) = a - bQ$, where $a, b > 0$, and $a > c$. Thus,

$$\Pi_i\ (q) = (a - bQ)q_i - cq_i.$$

At Cournot equilibrium $\overline{q}$,

$$\frac{\partial \Pi_i(\overline{q})}{\partial q_i} = 0 \qquad \text{for all } i.$$

Therefore,

$$a - b\sum_i \overline{q}_i - b\overline{q}_i - c = 0 \qquad \text{for all } i.$$

It follows that equilibrium is unique and symmetric and

$$\overline{q}_i = \frac{a - c}{b(n + 1)} \qquad \text{for all } i.$$

Equilibrium aggregate output is $(a - c)/b$ $(1 + 1/n)$; thus with two or more firms it is greater than monopoly output $(a - c)/2b$ but less than competitive output $(a - c)/b$. (The competitive output is defined by the condition that inverse demand price is equal to the constant per unit cost.)

It can be argued that Cournot incorrectly deduced from the fact that, in equilibrium, a homogeneous commodity can have only one price, the conclusion that an oligopolist cannot choose a different price from one charged by its competitors (see Simon 1984). Bertrand observed that if firms choose prices rather than quantities, then the Cournot outcome is not an equilibrium. For the case in which prices rather than quantities are the strategic variable, the analysis proceeds as follows. Assume that all n firms $(n > 1)$ have linear cost functions as described in Example 1 and that demand is continuous. Since the good being produced is homogeneous, a firm charging a price lower than that of other firms can capture the entire market. (To be specific we assume that all sales are shared equally among the firms that charge the lowest price.) Let $\overline{p}$ and $\overline{\Pi}$ denote price and individual profits respectively at the symmetric Cournot equilibrium. A firm can earn profits arbitrarily close to $n\overline{\Pi}$ (and hence, greater than Π) by lowering its price by a little from $\overline{p}$. The same argument can be used to show that in Bertrand equilibrium there is only one price at which sales are made. This price equals marginal cost and aggregate output is the competitive output. In Bertrand equilibrium, no firm can make a higher profit by altering its price decision.

An alternative equilibrium concept, due to Stackelberg, will be applied to the case of duopoly. There are two firms, labelled 1 and 2. The function $H_2(\cdot)$, called the reaction function of firm 2 (see Friedman 1977), is defined by

$$q_2 = H_2(q_1) \text{ if } \forall \tilde{q}_2, \quad \Pi_2(q_1, q_2) \geqslant \Pi_2(q_1, \tilde{q}_2).$$

The output vector $q = (\widehat{q}_1, \widehat{q}_2)$ is a Stackelberg equilibrium with firm 1 as the leader and firm 2 as the follower if firm 1 maximizes profit subject to the constraint that firm 2 chooses according to his reaction function; that is,

$$\forall q_1, \quad \Pi_i[\widehat{q}_1, H_2(\widehat{q}_1)] \geqslant \Pi_1[q_1, H_2(q_1)] \text{ and } \widehat{q}_2 = H_2(\widehat{q}_1)$$

In the model of Example 1, the Stackelberg equlibrium is

$$(\widehat{q}_i, \widehat{q}_2) = \left(\frac{a - c}{2b}, \frac{a - c}{4b}\right) \text{ and } \widehat{p} = \frac{a + 3c}{4}.$$

The Stackelberg equilibrium is interpreted as follows. The leader decides on a quantity to place on the market: this quantity is fixed. The follower decides how much to place on the market as a function of the quantity placed on the market by the leader. Again, equilibrium requires that neither firm can increase its profit by altering its decision.

Despite the fact that for the same model the Cournot, Bertrand and Stackelberg outcomes differ from each other, there is an important respect in which they are similar. In particular, they can all

be viewed as the application of the Nash equilibrium solution concept (see the entry on NASH EQUILIBRIUM) to games which differ with respect to the choice of strategic variables and the timing of moves. Thus, Cournot and Bertrand equilibria are Nash equilibria of simultaneous move games where the strategic variables are quantities and prices respectively. The Stackelberg equilibrium is the subgame perfect equilibrium of a game where firms make quantity choices but where the leader moves before the follower. This observation points to a general characteristic of oligopoly theory; the results are very sensitive to the details of the model. Nash equilibrium is the dominant solution concept in the analysis of oligopolistic markets and because its application is so pervasive one might expect substantial unity in the predictions of oligopoly theory. Unfortunately, as the preceding analysis makes clear, this is not so.

II. It was observed in Example 1 that aggregate output in Cournot equilibrium exceeds monopoly output. This holds generally and it implies that aggregate profit in a Cournot equilibrium is less than monopoly profit. Thus, there exists a pair of (identical) quantity choices for firms such that with these choices each firm earns a higher profit than in Cournot equilibrium. Since such choices do not form a Cournot equilibrium it would be in some firms' interest to deviate unilaterally from the choice assigned to it. In other words, without the possibility of binding contracts, the higher profit choices cannot be sustained, at least not in a static model. In this section, an extended example is presented to illustrate that if firms and consumers meet repeatedly, then it is possible for them to act more collusively than would be the case if they met only once. This result is very general and its importance for oligopoly theory was first pointed out by Friedman (see Friedman 1971).

There are two firms labelled 1 and 2. Each firm has three pure strategies *L, M* and *H* which can be thought of as representing 'low', 'middle' and 'high' quantities of output respectively. The payoffs are indicated in the matrix shown in Fig. 1, where (*L, L*), (*M, M*) and (*H, H*) may be thought of as the monopoly, Cournot and competitive outcomes respectively. In this game, (*M, M*) is the unique Nash equilibrium: given that one's opponent plays *M*, the best that he can do is play *M* himself.

		2's output		
		L	M	H
1's output	L	15,15	5,21	3,10
	M	21,5	12,12	2,5
	H	10,3	5,2	0,0

Oligopoly and Game Theory, Fig. 1

Consider now the game which is an infinite repetition of the game described above. The point that we wish to develop is that with repeated play it is possible to sustain outcomes that are much more collusive than (*M, M*). Strategies in the repeated game are more complicated than in the single period game. Specifically, the play of firm *i* in period *t* is a function of the 'history' of the game; i.e., of the plays of both firms in all periods preceding *t*. This allows a firm to 'punish' or 'reward' other firms. An outcome of the infinitely repeated game is a pair of infinite streams of returns, one for each firm. These infinite streams can be evaluated according to various criteria: two examples are considered. The stream $\{x_t\}_{t=0}^{\infty}$ is preferred to the stream $\{y_t\}_{t=0}^{\infty}$ according to the limit of means criterion if

$$\lim_{T\to\infty} (1/T)\sum_{t=0} (x_t - y_t) > 0.$$

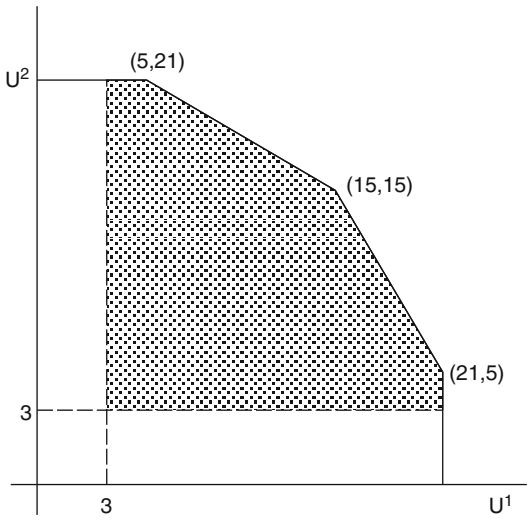
In the case where there is discounting, $\{x_t\}_{t=0}^{\infty}$ is preferred to $\{y_t\}_{t=0}^{\infty}$ if the former has a higher present value; that is, if

$$\sum_{t=0}^{\infty} \frac{x_t - y_t}{(1+r)^t} > 0,$$

where *r* is the discount rate.

Consider first the case where outcomes are evaluated according to the limit of means

criterion. The strategies in which both players choose M, no matter what the history, is easily seen to constitute an equilibrium. However, strategies in which both players choose L in every period (call this (L, L)) provided there has been no deviation also form a subgame perfect Nash equilibrium. If there is a deviation (L, L), then the equilibrium strategies call for players to play the subgame perfect Nash equilibrium (M, M). A firm contemplating a unilateral deviation from (L, L) at time t must weigh an immediate gain of 6 against a loss of at least 3 from $t + 1$ onwards. The deviation is unprofitable according to the limit of means criterion since a gain of 6 today becomes arbitrarily small when averaged over an increasingly large number of periods. The *mean* gain from the deviation is thus zero, while the mean loss from the deviation is 3. This argument can be used to demonstrate that any feasible payoff which dominates (M, M) can be realized by some equilibrium. (Strategies which involve reversion to Nash equilibrium forever cannot be used to characterize the entire set of subgame perfect Nash equilibria utility outcomes. In fact, the shaded area in Fig. 2 can be obtained). These ideas are developed further in Aumann–Shapley (1976), Friedman (1971) and Rubinstein (1979). See also Axelrod (1984).



Oligopoly and Game Theory, Fig. 2

It is considerably more difficult to characterize the set of subgame perfect equilibria in the case where outcomes are evaluated according to their present value. However, Abreu (1986) provides results which help to determine the amount of collusion that is possible with various amounts of discounting. Of course this amount depends on the interest rate. It also depends on punishments that are a good deal more subtle than the threat to repeat the single period Nash equilibrium in the event of any deviation. To introduce you to this work we return to Fig. 1 and consider first the case where $r = 1/4$. The threat of playing (M, M) forever if there is a deviation from (L, L), sustains (L, L) as an equilibrium. To see this, note that a firm by deviating gains 6 immediately and loses 3 forever, thereafter. This loss has a present value of $3/r = [3/(1/4)] = 12$, so deviation is not profitable. On the other hand, if $r = 3/4$ present value of the loss is $3/r = [3/(3/4)] = 4$, which is less than the gain from deviating. Therefore, deviation is profitable. But note that (L, L) can be sustained by a pair of subgame perfect Nash equilibrium strategies which are recursively defined as follows:

(a) The prescribed initial play is L for both players.
(b) If both players act according to the prescription in t, then they are both to play L in $t + 1$.
(c) If one or both do not play according to the prescription in t, then they are both to play H in $t + 1$.

To verify that this is a subgame perfect equilibrium, it has to be checked that no pattern of unilateral deviations is beneficial to a firm for any history of the game. The required argument is somewhat technical and is not given here (see Abreu 1986); however, we will show that no one-period deviation is profitable for any history of the game. There are two cases to consider:

(a) No firm has deviated in period $t - 1$. In this case, the other firm is considered to be playing L at t so that the gain from deviation at t, is at most 6. At $t + 1$, a loss of 15 $(= 15 - 0)$ will occur, which has a discounted value of

$$\frac{15}{1+r} = \frac{15}{1+(3/4)} = \frac{60}{7},$$

which is greater than 6.

(b) Some firm has deviated in period $t-1$. In this case, the equilibrium strategy requires both firms to play H in t. A firm by deviating (to L) can receive 3 in period t rather than 0; however the loss of 15 in the next period, as before, has present value 60/7, which is greater than 3.

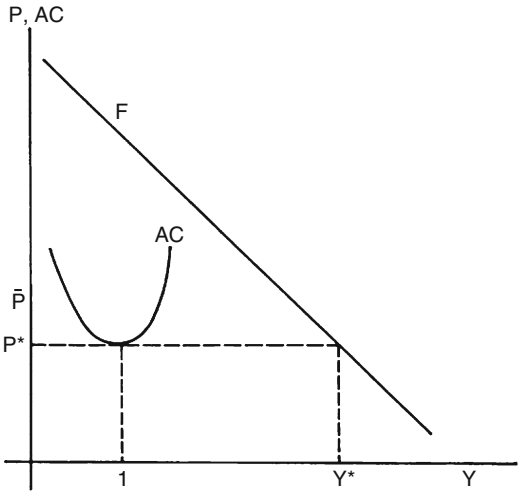
III. The theory of perfect competition assumes that all agents are price takers. We can improve our understanding of that theory by developing foundations for it that have firms behave strategically, in that they appreciate their market power, but nevertheless find themselves forced into actions that are well explained by the price taking assumption.

Consider the simple case where the demand function is linear and all firms have identical cost functions of the type $C = cq_i$. Recall from Example 1 that aggregate output in Cournot equilibrium is $(a-c)/b(1+(1/b))$ and the equilibrium price is therefore $a-(a-c)/1+1/n$. As the number of firms n increases, equilibrium price converges to c, which is the competitive price. This result does not generalize to the case of U-shaped average cost curves; furthermore, it has the defect that the number of firms in the market is fixed exogenously rather than being the result of a competitive process of free entry. These deficiencies are remedied in the work of Novshek.

Novshek's model

Novshek considers economies of the type described in Fig. 3.

In the figure, F denotes an inverse demand function and AC an average cost curve associated with the employment of any one of an unlimited number of available units of an entrepreneurial factor. The price P^* and the output Y^* are the (perfectly) competitive price and the (perfectly) competitive output respectively. An intuitive



Oligopoly and Game Theory, Fig. 3

argument for the convergence of equilibrium to P^* runs as follows. Suppose price $\overline{P}$ exceeds P^*. A firm can now enter the market and make a profit by producing at minimum average cost provided that it does not change prices by 'too much'. If the minimum efficient scale is small relative to the market, price will not change by 'too much' when the firm enters. Since there is an inexhaustible supply of potential entrants, $\overline{P}$ is not viable. Prices below P^* are not viable since firms are free to leave the market.

Novshek's theorem may be interpreted as a formalization of the intuitive argument presented above. The theorem states that there exists a quantity-setting Cournot equilibrium with entry when efficient scale is small relative to demand and that in this case the equilibrium output and price are approximately competitive. We conclude with a formal statement of the result. Assumptions: All firms have the same cost function C:

$$C(q_i) = 0 \quad \text{if } q_i = 0,$$

and

$$C(q_i) = C_0 + v(q_i) \quad \text{if } q_i > 0,$$

where $C_0 > 0$ and for all $q_i \geq 0$, $v' > 0$ and $v'' > 0$. Assume further that average cost is minimized uniquely at $q_i = 1$.

The inverse demand function $F(Q)$ is assumed to be twice continuously differentiable, with $F' < 0$ whenever $F > 0$, and there exists $Y^* > 0$ such that $F(Y^*) = C(1)$ (price equals minimum average cost). Definitions: An $\alpha(\alpha > 0)$ size firm corresponding to C is a firm with cost function $C_\alpha(q_i) = \alpha C(q_i/\alpha)$. Average cost for an α size firm is minimized at $q_i = \alpha$. For each α, C, and F, one considers a pool of available firms, each with cost function C_α, facing inverse market demand F.

Given C, F and α, an (α, C, F) market equilibrium with free entry is an integer n and an output vector $\overline{q} = (\overline{q}, \ldots, \overline{q}_n)$ such that (a) $\overline{q}$ is an n firm Cournot equilibrium (without entry), that is,

$$\forall_i = 1, \ldots, n, \qquad \forall q_i, \qquad \Pi_i(\overline{q}) \geqslant \Pi_i(q_i, \overline{q}_{-i}),$$

where $\Pi_i(\cdot)$ is the profit function for firm i described in Section 1 and (b) entry is not profitable, that is,

$$\forall q_i, F\left(\sum_{j=1}^{n} q_j + q_i\right) q_i - C_\alpha(q_i) \leqslant 0.$$

The set of all (α C, F) market equilibria with free entry is denoted by $E(\alpha\ C, F)$.

Novshek's theorem states that Cournot equilibrium exists provided that efficient scale is sufficiently small relative to demand, and furthermore, that it converges to the competitive output as efficient scale becomes small.

Novshek's theorem: Under the above hypotheses, for each C and F there exists $\alpha^* > 0$ such that for all $\alpha \in (0, \alpha^*]$, $E(\alpha, C, F)$ is non-empty. Furthermore, $\overline{q} \in E(\alpha, C, F)$ implies $\sum_{j=1}^{n} \overline{q}_j \in [Y^* - \alpha, Y^*]$ and so aggregate output and price approximate the perfectly competitive values P^* and Y^*.

It is perhaps reasonable to believe that the perfectly competitive result will also hold under conditions that allow for only a relatively small number of firms. No claim is made here that a large number of firms is necessary for firms to *act as if* they are unable to influence price. Novshek's Theorem, which relates well to the classical analysis of Cournot, provides a framework in which the perfectly competitive result obtains in the limit because in the limit firms cannot influence price.

See Also

▶ Game Theory

Bibliography

Abreu, D. 1986. External equilibria of oligopolistic supergames. *Journal of Economic Theory* 39: 191–225.

Aumann, R.J., and L. Shapley. 1976. Long term competition – A game theoretic analysis. Unpublished manuscript.

Axelrod, R.M. 1984. *The evolution of cooperation*. New York: Basic Books.

Bertrand, J. 1883. Théorie mathématique de la richesse social. *Journal des Savants* 48: 499–508.

Cournot, A. 1838. *Recherches sur les principes mathématiques de la théorie des richesses*. Paris: Hachette. Trans. by N.T. Bacon as *Researches into the mathematical principles of the theory of wealth*. New York: Macmillan, 1927.

Friedman, J.W. 1971. A non-cooperative equilibrium of supergames. *Review of Economic Studies* 38: 1–12.

Friedman, J.W. 1977. *Oligopoly and the theory of games*. Amsterdam: North-Holland.

Fudenberg, D., and E. Maskin. 1986. The folk theorem in repeated games with discounting and with incomplete information. *Econometrica* 54: 533–554.

Novshek, W. 1980. Cournot equilibrium with free entry. *Review of Economic Studies* 47: 473–486.

Rubinstein, A. 1979. Equilibrium in supergames with the overtaking criterion. *Journal of Economic Theory* 21: 1–9.

Simon, L. 1984. Bertrand, the Cournot paradigm and the theory of perfect competition. *Review of Economic Studies* 51: 209–230.

von Stackelberg, H. 1934. *Marktform und Gleichgewicht*. Vienna: Springer.

Olson, Mancur (1932–1998)

Joe A. Oppenheimer

Abstract

Mancur Olson was one of the small group of economists in the twentieth century who laid the foundation of rational choice theorizing about non-market behaviour. He demonstrated self-interested individuals have a great incentive to free ride rather than to contribute to the supply of a public good. He also

showed how self-interested group behaviour explained why nations tend to stagnate after periods of growth. Utilizing the notion of profit seeking political entrepreneurs, he argued the benefits of democratic systems were to contain the extractive costs imposed by government and to extend the time horizons for property rights.

Keywords

Arrow, K; Roving bandits; Baumol. W; Buchanan, J; Collective action; Constitutionalism; Theory of democracy; Dictatorship; Downs, A; Economic growth; Free rider problem; Interest groups; Kleptocracy; Non-market behaviour; Non-market economics; Olson, M; Political entrepreneur; Public goods; Samuelson, P; Social dilemmas; Time horizon; Tullock, G.: on constitutionalism; Von Neumann, J; Tragedy of the commons

JEL Classification

B31

Along with a handful of other economists of the twentieth century (Kenneth Arrow, James Buchanan, Anthony Downs, and John von Neumann), Mancur Olson laid the foundation for the adoption of rational choice theorizing about non-market behaviour in the social sciences. His work (1965) on the relationship between the rational choice of individuals and the performance of groups had a revolutionary impact on the fields of sociology and political science. In 1967 he left his first academic job at Princeton University to become Deputy Assistant Secretary of the US Department of Health, Education and Welfare. From there he went to the University of Maryland, where he held the position of Distinguished Professor of Economics, co-founded the University of Maryland's Center for Collective Choice, and founded the Institute for Research on the Informal Sector (IRIS).

Mancur Olson was born in January 1932 to a Norwegian-American farming family in North Dakota's Red River Valley. The valley contained some of the richest farmland in the state; the family grew mainly flax and did quite well. Neither his parents nor other members of that generation in the Olson family were educated beyond high school, but his father and his uncle were intellectually curious and questioning of society's arrangements. Mancur grew up on the farm and, as the eldest of three sons, he was permitted to be party to the adults' conversations about farming and social problems.

Throughout his life he recalled those early discussions regarding the shared interests of farmers in getting a fair price for their crops, the difficulties in their meeting other common concerns and the many references to the ability of the Scandinavian countries to overcome narrow interests to achieve both social justice and economic growth. He noted these as the part of his inheritance that motivated his life-long research interests in the problems of collective action, social justice and economic prosperity.

Mancur went to college at North Dakota State University on an Air Force Reserve Officer Training Corps (ROTC) scholarship. There he studied agricultural economics and had the good fortune to be mentored by Rainer Schickle (the father of the American composer, Peter Schickle). He won a Rhodes scholarship and went to Oxford, only to discover that Oxford dons could not imagine that a graduate (1954) from North Dakota's Agricultural College could qualify for entry into their graduate programme of Philosophy, Politics and Economics. So, unlike most of the other Americans coming from more prestigious institutions, he was required to get a second BA from Oxford before going on for an M.Phil.

At Oxford he met his lifelong companion and wife, Allison, who was also getting her M.Phil. (in history). The Olsons left Oxford together for the environs of Boston, where Allison had a job at Smith College and Mancur was to get a Ph.D. at Harvard. Two barriers were created. First, Mancur's chosen advisors, first Kenneth Galbraith and then also Otto Eckstein, left for Washington to work in the Kennedy administration. Further, Air Force officials discovered that Mancur had yet to do his service for his North Dakota Air Force ROTC contract, and they required him to leave Harvard to do military service. That service was performed between Rand, Brookings and the Air

Force Academy for two years, after which he was able to finish his work again at Harvard under the tutelage of Thomas Schelling. During this time their family grew and eventually Allison and Mancur had four children: Elicka, born in 1963, a veterinarian; Severn, born in 1967, a civil servant; Sander, born in 1969, a journalist; and Garth, who died in infancy.

Olson's major contribution to economics and to the social sciences more broadly was in the analysis of 'non-market' economics. He focused both on how individual non-market behaviour and political institutions (broadly understood) affected socio-political and economic outcomes. Many of his most important findings are encapsulated in his three major books *The Logic of Collective Action* (LCA) (1965), *The Rise and Decline of Nations: Economic Growth, Stagflation, and Social Rigidities* (RD) (1982), and *Power and Prosperity: Outgrowing Communist and Capitalist Dictatorships* (PP) (posthumous, 2001).

His first book, LCA, grew out of his dissertation, and focused on the non- Paretian outcomes one can expect from unorganized groups of individuals in their efforts to secure costly public goods (that is, goods where consumers cannot be excluded and consumption by one does not diminish consumption by another) such as air quality and peace. LCA built on the findings of William Baumol (1967) and Paul Samuelson (1954) who had shown that suboptimality was to be expected from rational self-interested behaviour regarding public goods. Olson expanded their arguments and generalized them by noting that the satisfying of virtually all shared interests is a form of public good, thereby selling the argument to the non-economist. The crux of the observation is simple: self-interested individuals have a great incentive to free ride rather than to contribute to the supply of a public good. Individuals will, after all, receive the good if others supply it. In LCA, Olson also tried to develop an argument that the size of the group was central to the analysis, but this was later shown to be erroneous (Frohlich and Oppenheimer 1970; Hardin 1982). The work spawned a paradigm shift in the study of group behaviour in both political science and sociology.

In RD, Olson built on LCA (and also the 1962 work of Buchanan and Tullock, who argued that one could evaluate constitutional rules by the externalities imposed upon losing subsets of the population by the extraction of resources for redistribution to the winners). In RD Olson argued that narrow-interested lobbying groups, designed to extract rewards from the general population via governmental action, clung to stable political systems much like barnacles to a ship's hull. Such extractive interests were shown to be more harmful the narrower the interests they represented. Newer political systems, built on cataclysmic changes in a society, were likely to be relatively free of such encumbrances and hence would lead to less wasteful extraction. Therefore, their economies would be more likely to exhibit substantial and sustained growth than would those associated with more established, stable political systems. He expanded the analysis (1990) to consider the comparative efficiency of the Scandinavian political systems' foundation on a coalition of a very few, very broad political interests. These welfare states were contrasted with welfare states in other industrialized countries built on a patchwork quilt of narrow, coalesced social-interest groups.

Coupled with Downs's 1957 work *An Economic Theory of Democracy,* LCA also sparked a reconsideration of political leaders as entrepreneurs (Salisbury 1969; Frohlich et al. 1971) as a way of solving the collective action problem. In the 1990s Olson himself began to mine the profit motive as a tool to understand the motivational characteristics of political leaders, and to reconsider the social gains from democracy. By assuming politics was necessarily based on coercive taxes, he considered the evolution of political systems as a hypothetical history from roving bandits to stationary bandits and then to kleptocratic political leaders constrained by the rules of succession and, more generally, competition. Roving bandits would take what they could. Stationary bandits, who controlled an area (for example, 'war lords' and mafiosi) would find it worth their while to ensure the prosperity of the population they exploited. Rules of succession, such as those that underlie monarchies, were shown to change the time horizon for

maximizing the extractive behaviour of the kleptocrat, thereby giving incentives to investments that had longer time horizons. Using the finding that narrower interests impose greater costs on society than wider ones, Olson (1993) and McGuire and Olson (1996) showed the general gain from democratic (majoritarian) systems to be the decrease in imposed external costs by the winning kleptocrats, as well as the extension of the time horizons for property rights. His last book, PP, built upon his kleptocratic entrepreneurial arguments and their relation to the time horizon of politicians. Long-term property and other rights were seen to be a key to the development of more complex financial markets that underlie modern economic development.

Olson's heritage is extraordinarily wide: the general interest in 'social dilemmas' grew directly out of his work via the translation of LCA into the language of *n* person game theory. His sure-handed encouragement of young scholars interested in non-market economics helped foster the multidisciplinary adoption of rational choice theoretic tools in the social sciences in general.

See Also

- ▶ Buchanan, James M. (Born 1919)
- ▶ Collective Action
- ▶ Public Goods
- ▶ Rational Choice and Political Science
- ▶ Tragedy of the Commons

Selected Works

1965. *The logic of collective action.* Cambridge, MA: Harvard University Press.
1982. *The rise and decline of nations: Economic growth, stagflation, and social rigidities*. New Haven: Yale University Press.
1990. *How bright are the northern lights? Some questions about Sweden.* Lund: Institute of Economic Research, Lund University.
1993. Dictatorship, democracy, and development. *American Political Science Review* 87, 567–76.
1996. (With M. McGuire.) The economics of autocracy and majority rule: The invisible hand and the use of force. *Journal of Economic Literature* 34, 72–96.
2001. *Power and prosperity: Outgrowing communist and capitalist dictatorships.* New York: Basic Books.

Bibliography

Baumol, W. 1967. *Welfare economics and the theory of the state*, 2nd ed. Cambridge, MA: Harvard University Press.
Buchanan, J., and G. Tullock. 1962. *The calculus of consent*. Ann Arbor: University of Michigan Press.
Downs, A. 1957. *An economic theory of democracy.* New York: Harper and Row.
Frohlich, N., and J. Oppenheimer. 1970. I get by with a little help from my friends. *World Politics* 23: 104–121.
Frohlich, N., J. Oppenheimer, and O. Young. 1971. *Political leadership and the supply of collective goods*. Princeton: Princeton University Press.
Hardin, R. 1982. *Collective action*. Baltimore: Johns Hopkins University Press.
Salisbury, R. 1969. An exchange theory of interest groups. *Midwest Journal of Political Science* 13: 1–32.
Samuelson, P. 1954. The pure theory of public expenditure. *Review of Economics and Statistics* 36: 387–389.

Oman, Economy of

Barry Turner

O

Keywords
Baiza; Gas reserves; Rial Omani

JEL Classification
O53; R11

Overview

Oil and natural gas (excluding petroleum products) contributed 41.0% to GDP in 2009; followed by manufacturing (including petroleum products) 10.3%; trade, restaurants and hotels, 10.2%; and finance and real estate, 9.9%.

Crude oil dominates the economy, accounting for 37% of exports in 2009, with China, Japan and

India taking 32%, 17% and 11% respectively. Government attempts to diversify the economy have focused on tourism, shipping and investment in infrastructure. There are also plans to increase natural gas production as a share of gross domestic product to 10% by 2020. Oman holds 0.5% of the world's liquefied natural gas supply.

Growth from 2005 to 2009 averaged 7.1%, supported by high oil prices and accelerated growth in non-hydrocarbon sectors including trade, transport and communications. Higher global commodity prices, domestic demand growth (prompted by fiscal stimuli) and strong private sector credit growth raised inflation to over 12% in 2008 although it has fallen since then.

Public debt was 5.6% of GDP in 2012, while unemployment stood at 15%. The government's eighth Five Year Plan (for 2011 until 2015) aims for GDP growth at a minimum of 3% per year, with RO 12 bn. earmarked for investment in the natural gas sector. It is hoped that development of gas-based and non-hydrocarbon industries will reduce unemployment.

Currency

The unit of currency is the *Rial Omani* (OMR). It is divided into 1,000 *baiza*. The rial is pegged to the US dollar. In July 2005 foreign exchange reserves were US$4,511m. and gold reserves totalled 1,000 troy oz (291,000 troy oz in April 2002). Total money supply was RO 1,067 m. in May 2005. Inflation was 12.6% in 2008, 3.5% in 2009, 3.3% in 2010 and 4.0% in 2011.

In 2001 the six Gulf Arab states—Oman, along with Bahrain, Kuwait, Qatar, Saudi Arabia and the United Arab Emirates—signed an agreement to establish a single currency by 2010. However, Oman withdrew from the scheme in 2007.

Budget

In 2008 revenues were RO 7,829.4 m. and expenditures RO 7,556.7 m. Oil revenue accounted for 67.5% of revenues in 2008; current expenditure accounted for 58.5% of expenditures.

Performance

Real GDP growth was 3.9% in 2009, 5.0% in 2010 and 5.4% in 2011. Total GDP in 2011 was US$70.0 bn.

Banking and Finance

The bank of issue is the Central Bank of Oman, which commenced operations in 1975 (*President*, Hamood Sangour Al Zadjali). All banks must comply with BIS capital adequacy ratios and have a minimum capital of RO 20 m. (minimum capital requirement for foreign banks established in Oman is RO 3 m.). In 2002 there were 15 commercial banks (of which nine were foreign) and three specialized banks. The largest bank is BankMuscat SAOG, with assets of RO 1.3 bn.

Total foreign debt was US$3,472 m. in 2005.

There is a stock exchange in Muscat, which is linked with those in Bahrain and Kuwait.

See Also

- ▶ Energy Economics
- ▶ International Monetary Fund
- ▶ Islamic Economic Institutions
- ▶ Islamic Finance
- ▶ Oil and the Macroeconomy
- ▶ Organization of the Petroleum Exporting Countries (OPEC)

Oncken, August (1844–1911)

Jürg Niehans

Oncken was born in Heidelberg on 10 April 1844 and died in Schwerin (Mecklenburg) on 10 July 1911. After studies in Munich, Heidelberg and Berlin, Oncken first became a landowner in

Oldenburg. Behind his scholarly interest in physiocracy was a life-long interest in agriculture. He began his academic career as university lecturer in economics and statistics at the Vienna School of Agriculture. In 1878, after a brief interlude at the Aachen Institute of Technology he accepted an appointment as professor of economics at the University of Bern, where he taught a wide range of courses until his retirement (because of failing eyesight) at the end of 1909.

As a general economist, Oncken has little claim to our attention. He never had a correct understanding of things like, say, diminishing returns, and he remained an unsophisticated advocate of protection, particularly for agriculture (1901a), applauding Henry Carey as the greatest living economist (1874). As an historian of economic thought, however, he was one of the leading lights between 1870 and 1920.

In his earliest historical paper (1874) Oncken criticized Adam Smith, in the spirit of German economics of that time, for his 'materialism' and his radical 'laissez faire' doctrines. In *Adam Smith and Immanuel Kant* (1877) he confessed that these criticisms did not survive a careful reading of the original sources. Instead he now stressed the similarities between those two giants of moral philosophy.

In Bern, Oncken's interests shifted to the Physiocrats. The result was a series of masterpieces of archival detective work and historical interpretation. It begins with a paper on the relationship between the Physiocrats and their disciples in Bern (1886a). In the following monograph (1886b), Oncken traces the maxim 'laissez faire' to d'Argenson (and not to Boisguillebert, as Stephan Bauer states in the *Encyclopaedia of the Social Sciences*) and further back to the time of Colbert, while 'laissez passer' was later added by de Gournay in a conversation with Mirabeau. In this context, Oncken puts forth the startling conjecture (not reiterated in (1902)) that the *Tableau Economique* was originally printed in support of a bid by Quesnay for the premier ministership.

Oncken's edition of Quesnay's writings (1888) became fundamental for all further work in this field. The sought-for completeness, however, eluded Oncken, because his very publication set off a renewed search of archives, culminating in Bauer's discovery of an early (but still not the first) version of the *Tableau Economique* (published by the British Economic Association) and of the article 'Hommes' in 1890. A further article, 'Impôts', was later published by Schelle. On the other hand, Oncken's collection includes non-economic writings not available in the 1958 edition, as well as the basic biographical sources. A first, hand-written draft of the *Tableau* was later reproduced in Oncken's *History of Political Economy* (1902).

Oncken himself made use of much of the newly discovered material in a succession of essays on Quesnay's life (1894–6) and the history of physiocracy (1893a, b, 1897a). He was well aware that the time was not ripe for a definitive biography, but for brilliance of historical scholarship Oncken's essays are unsurpassed. It is regrettable that, being available only in German and in inaccessible journals, they are usually not given the credit they deserve.

With respect to the circumstances under which the *Tableau Economique* was first printed, we do not seem to have progressed much beyond Oncken. The story that the most famous single page in the history of economics was typeset and printed by a bored Louis XV with his own hands, Oncken regarded as a fable, mainly because of its incompatibility with the known facts about the King's character. Schelle, however, chose to treat the story, despite its implausibility, as historical fact and his view was still accepted by Jacqueline Hecht in 1958.

Of the *History of Political Economy* (1902), only the first volume appeared, dealing with the time before Adam Smith. The first half, reaching from antiquity to mercantilism, is today of little interest. The second half, treating the Physiocrats and their predecessors, is still a valuable source of historical information about men, books and ideas, making an effective case for Quesnay as the 'founder' of economic science.

Oncken later returned to Adam Smith by defending him, not without some polemics, against his detractors of the Schmoller School (1897b, 1898). In another paper (1909), he also pointed out that Smith did not borrow from Ferguson, but had valid reasons for feeling that Ferguson had borrowed from his lecture notes.

Selected Works

1870. *Untersuchung über den Begriff der Statistik*. Leipzig.
1874. *Adam Smith in der Kulturgeschichte. Ein Vortrag*. Vienna.
1877. *Adam Smith und Immanuel Kant. Der Einklang und das Wechselverhältniss ihrer Lehren über Sitte, Staat und Wirthschaft. Vol. 1: Ethik und Politik*. Leipzig.
1886a. *Der ältere Mirabeau und die ökonomische Gesellschaft in Bern*. Berner Beiträge zur Geschichte der Nationalökonomie No. 1. Bern.
1886b. *Die Maxime Laissez faire et Laissez passer, ihr Ursprung, ihr Werden. Ein Beitrag zur Geschichte der Freihandelslehre*. Berner Beiträge zur Geschichte der Nationalökonomie No. 2. Bern.
1888. ed. *Oeuvres économiques et philosophiques de François Quesnay, fondateur du système physiocratique*. Berner Beiträge zur Geschichte der Nationalökonomie No. 3. Frankfurt/Paris.
1893a. *Zur Geschichte der Physiokratie*. Jahrbuch für Gesetzgebung, Verwaltung und Volkswirtschaft im Deutschen Reich 17. Leipzig.
1893b. *Ludwig XVI und das physiokratische System*. Zeitschrift für Litteratur und Geschichte der Staatswissenschaften 1. Leipzig.
1894–6. *Zur Biographie des Stifters der Physiokratie, François Quesnay*. Zeitschrift für Litteratur und Geschichte der Staatswissenschaften (from 1896: *Vierteljahrsschrift für Staats- und Volkswirtschaft, für Litteratur und Geschichte der Staatswissenschaften aller Länder*), vols 2, 3, 4. Leipzig.
1895. Political economy in Switzerland. *Economic Journal* 5: 133–137.
1896–8. Letter from Switzerland. *Economic Journal* 6: 308–314 (1896); 7: 228–293 (1897); 8: 269–273 (1898).
1897a. Entstehen und Werden der physiokratischen Theorie. In *Vierteljahrsschrift für Staats- und Volkswirtschaft, für Litteratur und Geschichte der Staatswissenschaften aller Länder*, vol. 5. Leipzig.
1897b. The consistency of Adam Smith. *Economic Journal* 7: 443–450.
1898. Das Adam Smith-Problem. *Zeitschrift für Socialwissenschaft* 1. Berlin.
1901a. *Was sagt die Nationalökonomie als Wissenschaft über die Bedeutung hoher und niedriger Getreidepreise?* Offprint from *Monatliche Nachrichten zur Regulierung der Getreidepreise*. Berlin.
1901b. Quesnay, François. In *Handwörterbuch der Staatswissenschaften*, 2nd ed, vol. VI. Jena.
1902. *Geschichte der Nationalökonomie. Erster Teil: Die Zeit vor Adam Smith*. Leipzig.
1909. Adam Smith und Adam Ferguson. *Zeitschrift für Socialwissenschaft* 12. Leipzig.

Online Platforms, Economics of

Alexander White

Abstract

Following the Internet's widespread adoption, much economic work has studied 'online platforms': firms that mainly interact with consumers in cyberspace. This article surveys such work, focusing on the ways in which traditional economic models have been adapted to incorporate novel aspects made relevant by the Internet. This literature can be divided roughly into two categories: broad-brush study of the competition between platforms and more fine-grained study of the ways in which users and platforms interact with one another. The former focuses on extending oligopoly theory to include 'consumption externalities'; the latter extends auction and search theory to a world of precisely measureable actions.

Keywords

Consumption externalities; Economics of the internet; Electronic commerce; Multi-sided platforms; Network effects; Search engines; Social networks; Sponsored search auctions; Two-sided markets

JEL Classification

D40; D43; D44; L10; L14

Introduction

The objective of this article is to articulate as clearly as possible the ways in which more traditional economic theory has been adapted in order to inform the study of Internet-connected intermediaries, or 'online platforms'. As Varian (2005, p. 12) remarks, 'Recent literature that aims to understand the economics of information technology is firmly grounded in the traditional literature. As with technology itself, the innovation comes not in the basic building blocks, but rather the ways in which they are combined'.

The building blocks whose combination this article focuses on are the following. First, it analyses the incorporation of consumption externalities into monopoly and oligopoly theory, selectively reviewing the recent two-sided markets literature and its precursors. In doing so, it presents a broad-brush picture that is useful for understanding the fundamental similarities and differences between online platforms and traditional firms, particularly regarding their pricing incentives. It then considers finer-grained perspectives of search engines and electronic commerce, surveying research that mixes elements of auction and search theory to understand how intermediaries sell advertisements to firms and to what extent they seek to match the latter with consumers in a frictionless way.

A Bird's Eye View: Oligopoly with Consumption Externalities

The theory of 'network effects', including the recently developed theory of 'two-sided markets' or 'multi-sided platforms', can be seen as attempting to extend classical oligopoly theory to incorporate features that are both prevalent among and, to some extent, novel to online platforms. A particularly salient point of distinction between traditional firms and online platforms is that, while the former are more likely to sell goods (or services) whose quality or performance depends largely on the way the *firm itself* produces it, the latter are more likely, one way or another, to sell *connections* between different economic agents who potentially benefit from interacting with one another. For example, while a traditional firm might sell shirts or plumbing services, an online platform might offer to link job seekers with employers or to provide the technical means to connect video game developers with gamers. Thus, consumers' perceptions of the quality of an online platform depend on the set of connections it offers as well as on more traditional factors.

In view of this 'connecting' role that they frequently play, a natural class of models for studying online platforms turns out to be that which generalises one version or another of existing oligopoly models by allowing for 'consumption externalities'. Whereas traditional oligopoly models assume that the utility each consumer derives from purchasing a good depends solely on the price of the good and, in some cases, on certain production choices the seller has made, platform models also allow this utility to depend on the consumption choices of *other consumers*.

The first literature studying the economics of network effects is not about 'online platforms' *per se*, as it was written before the Internet came of age, in the 1970s and 1980s. Instead, it largely considered issues such as ownership of telephone networks and standards for videocassettes and computer keyboards (Rohlfs 1974; Katz and Shapiro 1985; Farrell and Saloner 1985; David 1985). Nevertheless, the study of online platforms owes a significant debt to this literature, as it provides a fundamental building block for the subsequent literature on multi-sided platforms that is more explicitly focused on online industries. Moreover, it is worth noting that, while the telephone and home video industries may not have been 'online' at the time that the above literature was written, such goods are increasingly furnished using the Internet; consider, for example, Skype's Internet voice service or Netflix's streaming video service.

A Toy Model of a Platform

The crucial building block established by the earlier literature on network effects is the

formalisation of the aforementioned idea that each consumer 'cares about' the choices made by other consumers. To see, in the simplest way possible, the form that such a formalisation takes, consider the following very stylised model of a monopoly social network. Here (unlike what one observes with Facebook, for example), assume that the social network charges a monetary price to each consumer who chooses to become a member.

Consumer i's expected payoff from joining the social network is given by $v_i + \beta\widehat{N} - P$, where vi denotes the component of i's valuation for the network that is independent of the choices of other consumers, $\widehat{N}$ denotes the total number (more formally, the measure) of consumers expected to join the network, β captures the strength of i's valuation for 'interaction' with other consumers, and P denotes the price. For now, assume that β takes on the same value for all consumers, whereas vi is allowed to vary from one consumer to another and is distributed according to a continuous density function, $f(v)$. Meanwhile, assume that the social network's profits are given by $PN - c(N)$, where N denotes the number of consumers that indeed do join, and the continuous and increasing function $c(N)$ denotes the cost to the network of serving N consumers. The timing is such that, first, the social network announces P, and, second, each consumer decides whether or not to join.

For a given price, P, the set of consumers that choose to join are those for whom the inequality $v_i \geq P - \beta\widehat{N}$ holds. Therefore, the social network's demand can be written as

$$N\left(P, \widehat{N}\right) = \int_{P-\beta\widehat{N}}^{\infty} f(x)dx. \tag{1}$$

To proceed, it is convenient to assume that all consumers correctly anticipate the demand level that the network seeks to attract, so that $\widehat{N} = N$. (Justification for this assumption is considered below.) Using this assumption and the demand function in expression (1), it is relatively straightforward to derive the optimal price for the profit-maximising social network, given by

$$P = c' + \frac{N}{-\frac{\partial N}{\partial P}} - \beta N. \tag{2}$$

To interpret the price prescribed by Eq. (2), it is useful to compare it with the price that maximises total surplus, which can be derived in a similar fashion, given by

$$P = c' - \beta N. \tag{3}$$

According to (3), it is socially optimal for price to differ from marginal cost by βN Thus, if, as makes sense in this example, $\beta > 0$, meaning that consumers positively value the presence of others on the social network, then it is socially optimal to offer a 'discount' to each consumer, compared with traditional marginal cost pricing, equal to the total value that the consumer adds to the network: this total value is βN, because each consumer gains β from the presence of another consumer and there are N consumers. (This follows closely in the spirit of Pigou (1912).)

Turning to the formula in (2), for the network's optimal price, the only difference is the 'markup' term, $\frac{N}{-\frac{\partial N}{\partial P}}$, which is positive since $\frac{\partial N}{\partial P} < 0$, and which is precisely the same as the markup term that appears in the traditional monopoly pricing formula (Cournot 1838). This reflects the standard trade-off of infra-marginal gain versus marginal loss that faces a firm with market power when it raises its price. As is the case for a total surplus-maximising social planner, the profit-maximising network finds it optimal to discount its price by βN, since, for each additional consumer that joins, the network can extract the value this new consumer creates by raising its price by β, while still retaining N consumers.

Therefore, in this model that incorporates consumption externalities into the traditional monopoly model in the simplest way possible, no additional distortion arises between socially and privately optimal pricing. Such lack of additional distortion, however, is driven crucially by the assumption that β is the same for all consumers. The following discussion illustrates another distortion that arises when β varies

across consumers, but first, it considers a more basic form of heterogeneity among a platform's users.

A Richer Model: Diverse Groups of Platform Consumers

An important feature of many online platforms, which may seem to undermine their resemblance to the model described above, is the fundamental difference between the different types of agents that they connect to one another. For example, while, to a first approximation, one may think of a social network as a platform that connects users to each other, all of whom fall into some common category, in the examples given above of an employment or gaming platform, the assumption is inaccurate. Job seekers look especially for open positions, but not so much for other job seekers. Gamers may value both the opportunity to play new games and the ability to play them with other gamers, but their valuations for these two things cannot reasonably be conflated into a single valuation for interacting with other 'consumers'.

The literature on multi-sided platforms, pioneered especially by Caillaud and Jullien (2003), Evans (2003), Parker and Van Alstyne (2005) and Rochet and Tirole (2003), extends the type of model discussed above, allowing it to be applicable in a much broader set of environments in which consumers fall into many different categories and have differential valuations for interacting with one another. The basic model assumes that there are s groups or 'sides' of consumers. For example, consider a simple extension representing an employment platform and assume that $s = 2$, where side w represents workers seeking jobs and side e represents employers looking to fill positions. Here, the payoff to worker i of joining the platform is given by $v_i + \beta^w \widehat{N}^e - P^w$, where β^w denotes the marginal impact that the presence of an additional employer has on workers' valuations for joining the platform, $\widehat{N}^e$ denotes the number of employers that workers expect to join, and P^w denotes the price that the platform charges workers. Employers' payoffs are, analogously, given by $v_i + \beta^e \widehat{N}^w - P^e$. The platform's profits are now given by $P^w N^w + P^e N^e - c(N^w, N^e)$.

Continuing with the same form of analysis discussed above, consider the privately and socially optimal prices. On the workers' side of the market, these are, respectively,

$$P^w = \frac{\partial c}{\partial N^w} + \frac{N^w}{-\frac{\partial N^w}{\partial P^w}} - \beta^e N^e \tag{4}$$

and

$$P^w = \frac{\partial c}{\partial N^w} - \beta^e N^e \tag{5}$$

Note, moreover, that the corresponding expressions for prices charged to employers simply have the e and w indices reversed.

These pricing formulae follow the same logic as those of the first example, with one crucial modification. The 'discount' that workers receive, with respect to the analogous prices in a model with no consumption externalities, is given by $\beta^e N^e$: *employers*' valuations for the presence of an additional worker times the number of people that join the platform. In the case of the socially optimal price given by (5), $\beta^e N^e$ is the relevant quantity, because, in this example, it measures the total externality that a worker has on other consumers, as only employers 'care' about how many workers are present. For the same reason, in the case of the privately optimal price of Eq. (4), $\beta^e N^e$ measures the total additional profit that the platform can earn from its 'other' consumers, by virtue of serving one more worker. In a more general setting, with numerous groups of consumers and more complex externalities from one group to another, or within groups, the above formulae can be readily extended, in accordance with these general principles.

One particularly relevant insight that a two-sided model gives (but that a model with just one group of consumers does not) is the fact that it can be optimal for a profit-maximising platform to charge one group a negative price, i.e., to *pay* one type of consumer to join. This occurs when one group's own demand for the platform is relatively elastic compared to the

positive externality that their presence has on the other group. An oft-cited, albeit brick-and-mortar, example of such a phenomenon is night-club pricing, whereby women are charged a negative price in the form of free entry and complementary drinks, thus attracting more women and allowing the nightclub to extract a higher cover charge from men. A similar phenomenon occurs on many online platforms, although, in practice, it is often not feasible to charge one group a strictly negative price, so, instead, a price of zero is offered to the consumers that generate a high positive externality. One such example is that of search engines, such as Google and Microsoft's Bing, which are free for web users, whose presence in large numbers increases advertisers' willingness to pay to appear among the search results.

Rich Heterogeneity Within Groups

As mentioned above, another issue arises when consumers *within* a given group have heterogeneous valuations for externalities. For simplicity, reconsider the social network example with just one group of consumers. However, following Weyl (2010), who extends the model of Rochet and Tirole (2006), assume that consumer i's payoff is given by $v_i + \beta_i \widehat{N} - P$. Here, not only vi but also βi varies across consumers, and they are distributed according to a continuous joint density function, $f(v, \beta)$. Under this setup, consumers who choose to join the network are those for whom $v_i \geq P - \beta_i \widehat{N}$ holds, giving rise to demand,

$$N\left(P, \widehat{N}\right) = \int_{-\infty}^{\infty} \int_{P-y\widehat{N}}^{\infty} f(x, y)dxdy.$$

The expression for the network's privately optimal price then becomes

$$P = c' + \frac{N}{-\frac{\partial N}{\partial P}} - \widehat{\beta}N, \tag{6}$$

where $\widehat{\beta} \equiv E[\beta_i | v_i = P - \beta_i N]$, while the expression for the socially optimal price is given by

$$P = c' - \overline{\beta}N, \tag{7}$$

where $\widehat{\beta} \equiv E[\beta_i | v_i \geq P - \beta_i N]$.

In economic terms, $\widehat{\beta}$ is the average valuation among the set of *marginal* consumers, i.e., those consumers who are indifferent between joining the network or not, for the externality created by one more consumer. In contrast, $\overline{\beta}$ is the analogous quantity averaged over the entire set of consumers that join the network. Thus, when consumers are allowed to be heterogeneous in their valuations for externalities, a second source of distortion arises between privately and socially optimal pricing incentives. On the one hand, regarding total surplus maximisation, as (7) illustrates, it is still optimal to offer consumers a discount, with respect to marginal cost, equal to the total externality they create, measured here by $\overline{\beta}N$.

On the other hand, as (6) shows, the profit-maximising network has an incentive to offer such a discount only to the extent that it can recoup the loss that the discount provokes by increasing the rent it can extract from its entire set of N consumers who value the network more highly when there is an additional consumer. When adding a marginal consumer, in order to hold fixed the size of its demand at N, the network increases its price by an amount which will not incite an additional flow of users either into or out of the network. This amount is precisely $\widehat{\beta}$, the average of those consumers who are marginal, because the marginal set of consumers are the only ones who, in response to small price changes, are prone to reversing their decision of whether or not to join.

Note that $\widehat{\beta}$ may be either larger or smaller than $\overline{\beta}$. Thus, unlike the traditional markup distortion, which always pushes the privately optimal price to be higher than the socially optimal one, the distortion arising from heterogeneity in valuations for externalities can push prices to be either too high or too low. Moreover, as Weyl (2010) notes, this distortion, based on differences in valuations for a good's characteristics between marginal and infra-marginal consumers, closely mirrors the one studied by Spence (1975) in his model of a traditional,

quality-choosing monopolist. Veiga and Weyl (2012) explore this connection further, developing a model that allows consumers to *contribute* externalities to the network in a heterogeneous way.

Modelling Platform Competition: Challenges and Proposed Solutions

The above discussion focuses on monopoly platforms, and a detailed discussion of competition is beyond the scope of this article. A challenge facing the analysis of models of competition with consumption externalities is the presence of multiple equilibria, which arise from two sources. One source, touched on above (when the assumption that $\widehat{N} = N$ was posited), is the possibility of multiple equilibria in the game played by consumers after the platform has already set its prices. In the case of monopoly, this problem can be assumed away somewhat innocuously, because platforms can use contingent pricing to eliminate consumers' coordination problems. (See Weyl's (2010) discussion of 'Insulating Tariffs' as well as Ambrus and Argenziano (2009), who take an alternative approach that refines consumers' possible reactions to given prices.) The other source of multiplicity arises in the price-setting game played by competing platforms. As a well-known article by Armstrong (2006) shows, if platforms compete with one another using arbitrary forms of contingent pricing, then the equilibrium of their strategic interaction with one another is severely underdetermined. White and Weyl (2012) propose *Insulated Equilibrium* as a joint solution to these two multiplicity problems, using it to analyse the impacts of consumer heterogeneity on pricing in a competitive environment. See Reisinger (2012) for an alternative solution to the indeterminacy in the price-setting game.

The above discussion illustrates some of the main issues that arise when consumption externalities are incorporated into oligopoly theory in order to make it useful for the broad-brush study of online platforms that connect consumers to one another. However, this discussion is by no means exhaustive. For example, in many circumstances dynamic considerations, such as those studied by Cabral (2011), are of great importance. In this category, one may include the study of time-dependent pricing strategies as well as the issue of which and how many platforms survive in their respective markets. Furthermore, this discussion ignores the possibility of a platform engaging in second-degree price discrimination, an issue that Gomes and Pavan (2012) concentrate on. Another interesting issue is the impact of consumers patronising multiple competing platforms (known in the literature as 'multi-homing') and potentially interacting multiple times (Athey et al. 2012). Finally, for a broad, recent survey of the multi-sided platforms literature, focusing on applications, see Rysman (2009).

Detailed Views of Interaction on and Via Online Platforms

The previous section takes a 'bird's eye view'; however, there are also many more detailed issues related to online platforms that can be better understood by 'zooming in'. While attention to detail typically comes with a loss of general applicability, two topics have proved to be of especially broad interest. These are the 'sponsored search' auctions that search engines such as Google and Bing use to sell advertisements and the techniques that Internet sellers use to sustain profit margins in an environment that would seem to favour perfect price competition.

Sponsored Search Auctions

In practice, sponsored search auctions differ from 'textbook' auctions in two particularly important ways. First, even though the size of advertisers' bids largely determines whether they appear in a more prominent slot near the top of the search engine's results page or in a more obscure one near the bottom, the auction mechanisms dictate that the total payment that an advertiser makes depends on the number of times that users *click* on that particular advertiser's link. Second, because the items being auctioned are ads, which

are, in effect, opportunities for sellers to connect with web surfers, the interplay between the auction mechanism and users' surfing behaviour matters, both for descriptive purposes and for evaluating the welfare associated with different mechanisms.

The first articles in economics to study the impact of using per-click or 'Generalised Second-Price' auction mechanisms are Edelman et al. (2007) and Varian (2007). While the exact details of the auction are both complex and proprietary to search engines, two stylised features of these auctions are the following.

a. Auctions are conducted on a keyword-by-keyword basis. So, for example, an advertiser can bid one amount for an ad slot appearing among the results a user sees after searching for 'shoes' and a different amount for an ad slot a user sees after searching for 'boots'.
b. In each auction, the highest-bidding advertiser receives the most prominent slot and pays the amount bid by the second-highest bidder each time a user clicks on the former's ad. The second-highest bidder receives the second-most prominent slot and pays the third-highest bid for each click it receives, and so on.

The aforementioned articles show that, despite the second-price 'flavour' of these auctions, they are not special cases of the Vickrey–Clarke–Groves (VCG) mechanism, which, in view of a famous result of Green and Laffont (1979), implies that it is not a dominant strategy for advertisers to bid their true valuations for clicks. They further show that such auctions have multiple equilibria, all of which yield at least as much revenue to the search engine as would a VCG mechanism. In more recent work, Athey and Nekipelov (2012) modify the model somewhat, relaxing the assumption that advertisers literally submit a separate bid in every single auction and showing that this pins down a unique equilibrium.

Both Athey and Ellison (2011) and Chen and He (2011) explicitly integrate user behaviour into models of sponsored search. A basic insight of these models is that the value for an advertiser of receiving a slot at the top of the search results page stems not necessarily from its intrinsic 'prominence'. Instead, the value can come from surfers' anticipation that advertisers who bid more in an auction are also more likely to have websites that are worth visiting, thus creating a positive feedback loop. Athey and Ellison (2011) further show that, unlike in auctions for traditional goods, in search auctions, reserve prices can be welfare-improving, as they screen out low-quality sites that would be a waste of users' time to visit.

Price Competition and Obfuscation

As Ellison and Ellison (2009, p. 427) remark, 'When Internet commerce first emerged, one heard a lot about the promise of "frictionless commerce." Search technologies would have a dramatic effect by making it easy for consumers to compare prices at online and offline merchants'. However, many would argue that, in its current, more mature state, online shopping is sometimes rather complicated, with goods' prices and characteristics often not disclosed in a transparent way.

Numerous articles consider different aspects of this issue. Notable examples include the relatively early paper by Baye and Morgan (2001) focusing on brick-and-mortar firms' decisions about whether or not also to advertise online, Hagiu and Jullien (2011), who examine an intermediary's incentives not to eliminate the search frictions of its users, and Ellison and Wolitzky (2012), who adapt the classic model of Stahl (1989) to consider the incentives facing the sellers of goods themselves to provoke such frictions. Also, on this issue, see numerous articles published in volume 121, Issue 556 of the *Economic Journal*, described in an introduction by Wilson (2011).

Further Issues

This article focuses on some of the ways in which traditional components of microeconomics have been combined to build theories that speak to a world in which many important firms are 'online

platforms'. It does not begin, however, to address many of the fascinating issues involving such firms that the economics literature has studied. In particular, it ignores a large body of empirical work that examines matters including regulation of online privacy (Goldfarb and Tucker 2011a), circumstances in which people substitute between online and offline platforms (Goldfarb and Tucker 2011b) and the effect of such substitution on broader social trends (Gentzkow and Shapiro 2011), the dynamics of pro-social behaviour in large online communities (Zhang and Zhu 2011), and the ways in which online sellers experiment (Einav et al. 2011), to name a few. Finally, for a more comprehensive survey, which expands on many of the topics discussed in this article, the reader is referred to Levin (2012).

See Also

- ▶ Computer Industry
- ▶ Internet, Economics of the
- ▶ Information Technology and the World Economy
- ▶ Network Goods (Empirical Studies)
- ▶ Open Source Software, a Brief Survey of the Economics of

Acknowledgments I thank Catherine Tucker and Glen Weyl for their helpful comments, and I gratefully acknowledge the financial assistance of Project 20121087916 supported by the Tsinghua University Initiative Scientific Research Program and Project 71203113 supported by the National Natural Science Foundation of China.

Bibliography

Ambrus, A., and R. Argenziano. 2009. Asymmetric networks in two-sided markets. *American Economic Journal: Microeconomics* 1: 17–52.

Armstrong, M. 2006. Competition in two-sided markets. *RAND Journal of Economics* 37: 668–691.

Athey, S., and G. Ellison. 2011. Position auctions with consumer search. *Quarterly Journal of Economics* 126: 1213–1270.

Athey, S., and S. Nekipelov. 2012. A structural model of sponsored search advertising auctions. *Mimeo*.

Athey, S., E. Calvano, and J.S. Gans. 2012. The impact of the Internet on advertising markets for news media. *Mimeo*.

Baye, M.R., and J. Morgan. 2001. Information gatekeepers on the internet and the competitiveness of homogeneous product markets. *American Economic Review* 91: 545–574.

Cabral, L. 2011. Dynamic price competition with network effects. *Review of Economic Studies* 78: 83–111.

Caillaud, B., and B. Jullien. 2003. Chicken & egg: Competition among intermediation service providers. *RAND Journal of Economics* 34: 309–328.

Chen, Y., and C. He. 2011. Paid placement: Advertising and search on the internet. *Economic Journal* 121: F309–F328.

Cournot, A. 1838. *Researches into the mathematical principles of the theory of wealth* (1971 edn). (trans: Bacon, N.). New York: Augustus M. Kelley.

David, P.A. 1985. Clio and the economics of QWERTY. *American Economic Review* 75: 332–337.

Edelman, B., M. Ostrovsky, and M. Schwarz. 2007. Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. *American Economic Review* 97: 242–259.

Einav, L., T. Kuchler, J. Levin and N. Sundaresan. 2011. Learning from seller experiments in online markets. *Mimeo*.

Ellison, G., and S.F. Ellison. 2009. Search, obfuscation, and price elasticities on the internet. *Econometrica* 2: 427–452.

Ellison, G., and A. Wolitzky. 2012. A search cost model of obfuscation. *RAND Journal of Economics*, in press.

Evans, D.S. 2003. The antitrust economics of multi-sided platform markets. *Yale Journal of Regulation* 20: 325–431.

Farrell, J., and G. Saloner. 1985. Standardization, compatibility, and innovation. *RAND Journal of Economics* 16: 70–83.

Gentzkow, M., and J.M. Shapiro. 2011. Ideological segregation online and offline. *Quarterly Journal of Economics* 126: 1799–1839.

Goldfarb, A., and C. Tucker. 2011a. Privacy regulation and online advertising. *Management Science* 57: 57–71.

Goldfarb, A., and C. Tucker. 2011b. Search engine advertising: Channel substitution when pricing ads to context. *Management Science* 57: 458–470.

Gomes, R., and A. Pavan. 2012. Many-to-many matching design. *Mimeo*.

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

Hagiu, A., and B. Jullien. 2011. Why do intermediaries divert search? *RAND Journal of Economics* 42: 337–362.

Katz, M.L., and C. Shapiro. 1985. Network externalities, competition, and compatibility. *American Economic Review* 75: 424–440.

Levin, J. 2012. The economics of internet markets. In *Advances in economics and econometrics: Tenth world congress*, vol. 1, ed. D. Acemoglu, M. Arellano, and E. Dekel. New York: Cambridge University Press.

Parker, G.G., and M.W. Van Alstyne. 2005. Two-sided network effects: A theory of information product design. *Management Science* 51: 1494–1504.

Pigou, A. 1912. *Wealth and welfare*. London: Macmillan.
Reisinger, M. 2012. Unique equilibrium in two-part tariff competition between two-sided platforms. *Mimeo*
Rochet, J.-C., and J. Tirole. 2003. Platform competition in two-sided markets. *Journal of the European Economic Association* 1: 990–1029.
Rochet, J.-C., and J. Tirole. 2006. Two-sided markets: A progress report. *RAND Journal of Economics* 37: 645–667.
Rohlfs, J. 1974. A theory of interdependent demand for communications service. *Bell Journal of Economics and Management Science* 5: 16–37.
Rysman, M. 2009. The economics of two-sided markets. *Journal of Economic Perspectives* 23: 125–143.
Spence, A.M. 1975. Monopoly, quality, and regulation. *Bell Journal of Economics* 6: 417–429.
Stahl, D.O. 1989. Oligopolistic pricing with sequential consumer search. *American Economic Review* 79: 700–712.
Varian, H.R. 2005. Competition and market power. In *The economics of information technology: An introduction*, ed. H.R. Varian, J. Farrell, and C. Shapiro, 1–48. Cambridge: Cambridge University Press.
Varian, H.R. 2007. Position auctions. *International Journal of Industrial Organization* 25: 1163–1178.
Veiga, A., and E.G. Weyl. 2012. Multidimensional product design. *Mimeo*.
Weyl, E.G. 2010. A price theory of multi-sided platforms. *American Economic Review* 100: 1042–1072.
White, A., and E.G. Weyl. 2012. Insulated platform competition. *Mimeo*.
Wilson, C.M. 2011. Advertising, search and intermediaries on the internet: Introduction. *Economic Journal* 121: F291–F493.
Zhang, M., and F. Zhu. 2011. Group size and incentives to contribute: A natural experiment at Chinese wikipedia. *American Economic Review* 101: 1601–1615.

Open Field System

Donald N. McCloskey

The open field system was the arrangement of peasant agriculture in northern Europe before the twentieth century into scattered strips communally regulated but privately owned. The system shares features with much peasant agriculture worldwide, especially in its scattering of strips. Dissolved gradually by 'enclosure' (Turner 1984), first in England and Scandinavia and later in France (Grantham 1980), Germany (Mayhew 1973), and the Slavic lands (Blum 1961), it has been seen as an obstacle to agricultural development. The system is most thoroughly documented in England (Gray 1915; Ault 1972; Baker and Butlin 1973; Yelling 1977; and hundreds of local studies). The English case has long been disproportionately important because it has provided a rich set of myths for other cases of traditional agriculture and reform. (The Russian version, the *mir*, is important for the same reason; but its unique feature – the periodic redistribution of the strips among families – arose in the eighteenth century out of the need to pay taxes, not out of the ancient community of cousins.)

The scattering of strips within two or three large, unfenced (hence 'open') fields, perhaps a thousand acres each, implied common grazing on the stubble: fencing of the typical landholder's seven or so plots, an acre or so each, was otherwise too expensive. The common grazing implied in turn common decisions on what was to be grown and when. The grazing herd forced all the villagers to plant and harvest on a common schedule.

The word 'common' has led to a misunderstanding of the system by economists and geographers unfamiliar with the history (Hardin 1968; Baack and Thomas 1974; Cohen and Weitzman 1975). The 'commons' famed in nursery rhyme and academic fantasy were the waste land suitable only for grazing, usually absent or tiny in the open field regions, and to be distinguished from the main fields, the ploughed lands grazed 'in common' after the harvest (confusingly named 'the common fields'). The 'common' grazing and 'common' cropping did not mean that cattle and sheep were socialized or that cultivation was accomplished in communal gangs. The commonness was in coordination, not in ownership; in regulation, not in reward. Land, labour and capital were wholly private and rent-earning, not (as economists have imagined) 'common pools' or 'fisheries'. The inefficiency of open fields, therefore, was not the inefficiency of a primitive socialism but of an imperfect capitalism.

The inefficiencies of open fields arose from spillovers and lack of specialization (the loss of

land in boundaries and the loss of time in commuting from strip to strip were unimportant). Court records of quarrels between neighbours, and the poetry of the time, speak eloquently of the inconvenience of propinquity. In *Piers Ploughman* (*c*1378) Avarice boasts 'If I go to the plough, I pinch so narrow/ That a foot's land or a furrow to fetch I would/ Of my next neighbour, take of his earth;/ And if I reap, overreach, or give advice to him that reap/ To seize for me with his sickle what I never sowed.' Three centuries later, after voluntary enclosure had narrowed the system (which anyway had not existed in highland areas), Thomas Tusser recommended 'several' (that is, consolidated) farming over 'champion' (that is, open field), because 'Good land that is several, crops may have three,/ In champion country it may not so be:/ . . ./There common as commoners use,/ For otherwise shalt thou not choose.' Although the open field village as a whole could introduce novelties, the lone villager bound by the decision of the commoners could not. The system lasted in the Midlands of England into the eighteenth century, the last of it dissolved slowly by special acts of Parliament. Arthur Young was typical of this latter age, and of historians looking back, in scorning the inefficiencies of the system, railing against 'the Goths and Vandals of open-field farmers'.

With a complete set of markets, as A. Smith, R. Coase, and K. Arrow have explained, the Goths and Vandals would have traded away their inefficiencies. An explanation of open fields must depend therefore on some trade being blocked. The oldest explanation, imagining a spirit of fellowship within the primitive Germanic community, asks 'Who laid out these fields? The obvious answer is that they were laid out by men who would sacrifice economy and efficiency at the shrine of equality' (Maitland 1897). Evidence has accumulated since the nineteenth century that the fields in question were not laid out at once and the men laying them out did not worship at the shrine of equality. Yet even if they did, and did lay out the fields, they could have exchanged their scattered strips to achieve rational holdings later. An egalitarian explanation of *persistent* open fields depends therefore on a failure in the market for land. Here again, however, the evidence testifies to the contrary: villages in medieval England and in much of Europe had in fact a cheap and active market in parcels of land.

The same difficulty lies in the way of any other explanation of scattering. Scattering has been explained as arising also from egalitarian inheritance (Dovring 1965), common ploughing (Seebohm 1883), common grazing (Dahlman 1980), scheduling of harvest work (Fenoaltea 1976), and diversification of local risks (McCloskey 1976). These depend on market failures respectively in land, ploughing services, grazing rights, labour, and insurance. None of these is immune from the criticism, and few have faced it.

Insurance has been tested most thoroughly. The scattering of strips strikes the eye of an economist as diversification. Anthropologists, trained to take seriously the reasons proffered by their people, report the Hopi scattering corn lands to diversify against floods and Swiss peasants diversifying across altitudes. Furthermore, the amount of local variation in England was great: a wet year flooded clay lands in the valley while the chalk hills drained; infestations of insects and the paths of hailstorms were local. The portfolio that a peasant bought by having scattered strips can be calculated from medieval evidence of yields and modern evidence of agronomical experiments. The optimal number of plots proves to be roughly the same as the observed number.

The insurance argument, like the rest, can be criticized for ignoring a market, the market in this case for insurance (Fenoaltea 1976). It may well be that scattering was a form of insurance, but most social institutions anyway have an element of insurance, more so in the fourteenth century than now. A peasant could insure by sharecropping, by entering an extended family, by taking loans from the landlord, by purchasing liquid assets, and by storing grain. At the margin, however, the return from each form of insurance would be the same as any other. The scattering of strips incurred costs of about 15 per cent of output. The one other form of insurance whose costs are easily calculable is storage of grain (McCloskey and Nash 1984). A year's worth of grain storage in

the fourteenth century cost 40 per cent of the value of the crop, largely because interest rates were 30 per cent a year (by 1600, after interest rates had fallen, the cost was only 15 per cent: enclosure proceeded apace). For insurance, at least, we have a measure of the great imperfection in its market and therefore an explanation of the persistence of open fields.

Precise conclusions aside, the recent explanations all agree on a picture of the medieval peasant differing sharply from the romantic one drawn by nineteenth-century German scholarship. The new picture is market saturated (Popkin 1979) and individualistic (Macfarlane 1978); at any rate it is more so than the 'natural economy' once thought to prevail in medieval Europe and the 'moral economy' now thought to prevail in poor countries today.

See Also

- ▶ Common Land
- ▶ Common Property Rights
- ▶ Feudalism

Bibliography

Ault, W.O. 1972. *Open-field farming in medieval England: A study of village by-laws*. London/New York: Allen and Unwin/Barnes and Noble.

Baack, B.D., and R.P. Thomas. 1974. The enclosure movement and the supply of labor during the Industrial Revolution. *Journal of European Economic History* 3(2): 401–423.

Baker, A.H.R., and R.A. Butlin (eds.). 1973. *Studies of field systems in the British Isles*. Cambridge: Cambridge University Press.

Blum, J. 1961. *Lord and peasant in Russia: From the ninth to the nineteenth century*. Princeton: Princeton University Press.

Cohen, J., and M.L. Weitzman. 1975. A Marxian model of enclosures. *Journal of Development Economics* 1(4): 287–336.

Dahlman, C. 1980. *The open field system and beyond: A property rights analysis of an economic institution*. Cambridge: Cambridge University Press.

Dovring, F. 1965. *Land and labor in Europe in the 20th century*, 3rd ed. The Hague: Nijhoff.

Fenoaltea, S. 1976. Risk, transaction costs, and the organization of medieval agriculture. *Explorations in Economic History* 13(2): 129–151.

Grantham, G. 1980. The persistence of open field farming in nineteenth-century France. *Journal of Economic History* 40(3): 515–531.

Gray, H.L. 1915. *English field systems*. Cambridge, MA: Harvard University Press.

Hardin, G. 1968. The tragedy of the commons. *Science* 162: 1243–1248.

McCloskey, D.N. 1975. The persistence of common fields. In *European peasants and their markets*, ed. W.N. Parker and E.L. Jones. Princeton: Princeton University Press.

McCloskey, D.N. 1976. English open fields as behavior towards risk. *Research in Economic History* 1: 124–170.

McCloskey, D.N., and J. Nash. 1984. Corn at interest: The cost and extent of grain storage in medieval England. *American Economic Review* 74(1): 174–187.

Macfarlane, A. 1978. *The origins of English individualism*. Oxford: Basil Blackwell.

Maitland, F.W. 1897. *Domesday book and beyond*. Cambridge: Cambridge University Press.

Mayhew, A. 1973. *Rural settlement and farming in Germany*. New York: Barnes and Noble.

Popkin, S.L. 1979. *The rational peasant: The political economy of rural society in Vietnam*. Berkeley: University of California Press.

Seebohm, F. 1883. *The English village community*. London: Longmans & Co.

Turner, M. 1984. *Enclosures in Britain, 1750–1830*. London: Macmillan.

Yelling, J.A. 1977. *Common field and enclosure in England 1450–1850*. London: Macmillan.

Open Source Software, a Brief Survey of the Economics of

Chaim Fershtman and Neil Gandal

Abstract

The open source model is a form of software development in which the source code is made available, free of charge, to all interested parties; further users have the right to modify and extend the program. Open source software (OSS) methods rely on developers who reveal the source code under an open source licence. Under certain types of open source licence, any further development using the source code must also be publicly disclosed. In this brief survey, we will focus on several key aspects of open source software.

Keywords
Digital content; Intrinsic motivation; Licences; Open source software; R&D

JEL Classifications
L17; O31

Introduction

The open source model is a form of software development in which the source code is made available, free of charge, to all interested parties; further users have the right to modify and extend the program. Open source software (OSS) methods rely on developers who reveal the source code under an open source licence. Under certain types of open source licence, any further development using the source code must also be publicly disclosed.

The open source model has become quite popular and is often referred to as a movement with an ideology and enthusiastic supporters – see for example Stallman (1999) and Raymond (2000). At the core of this process are two interesting phenomena: unpaid volunteers do a non-trivial portion of the development of open source programs and, unlike commercial software, open source software is not sold or licensed for a fee.

Having unpaid volunteers develop 'free' software is a puzzling phenomenon for economists. (Boldrin and Levine (2009) argue that from a historical perspective, the 'open source' model of development is the norm for many industries. In this entry, we will focus on the open source phenomenon in software. See the final section for extensions of open source methodology to other applications.) What are the incentives that drive contributors to invest time and effort in developing these open source programs, which are not sold or licensed for a fee? Intrinsic motivation may provide a partial explanation and suggests an analogy between academia and the open source movement. While publication plays an important role in academia, the analogy in the OSS world is being included in the 'list of contributors' of different projects. Being listed as a contributor may enhance the reputation of a programmer and can be instrumental in the job market. Additional incentives to develop open source software come from 'self-use' benefits and the enhancement of other (potentially proprietary) products in the market.

In this brief survey, we will focus on several key aspects of open source software. Much of the empirical work we review in this survey paper comes from high-quality data on open source software projects which are publicly available. Since most open source development takes place in the public domain (by which we mean publicly available 'via the Internet'), data on many aspects of open source development are often available at various forges or platforms. These forges typically host many independent software projects. SourceForge, the largest forge, had more than 240,000 projects and 2.6 million registered users as of August 2010. Analysing the open source data available at SourceForge.net has already provided insight on worker motivation, the tradeoffs between intrinsic and monetary motivation, and the effect of the form of licensing on the contributions of developers (see Lerner and Tirole (2005b) and Fershtman and Gandal (2007), which are discussed below).

In the next section we examine motivation of programmers, while the following section examines the types of licensing employed in open source projects. The next two consider changes in the open source model, beginning with firm participation in the open source process and then reviewing some changes in the institutional structure of open source.

Open source development leads to very different incentives for R&D than the traditional proprietary development model – see Maurer and Scotchmer (2006) for a detailed analysis. Hence examining open source successes and failures may shed some light on the R&D process itself. We briefly examine this issue in the penultimate section. In the final section we briefly discuss the extensions of open source software model to digital content.

Finally, this is a short review; hence we focus on the topics we consider to be most important. Several books provide detailed reviews of open source software: see Dibona et al. (1999, 2006)

and Lerner and Schankerman (2010). Excellent early survey articles include Lerner and Tirole (2005a) and von Krogh and von Hippel (2006).

Motivation of Programmers

Theoretical Research on the Motivation of Programmers

Early research on the open source phenomenon was primarily theoretical and focused on the motivation of unpaid programmers to work on open source projects. Several explanations regarding motivation have been offered in the literature: Lerner and Tirole (2002) argue that developers of open source programs acquire a reputation that is eventually rewarded in the job market, while Harhoff et al. (2003) argue that end users of open source benefit by sharing their innovations. Ghosh et al. (2002) argue that open source development is more like a hobby than a (paying) job. Johnson (2002) develops a model of open source software as voluntary provision of a public good – but for such a model one needs to assume that the primary motivation of developers is the 'consumption' or use of the final program. (Johnson (2006) presents a model in which the OSS organisation structure is superior to that of proprietary development as it minimises transaction costs and avoids agency problems.)

Empirical Research on the Motivation of Programmers

Using survey methods, Hars and Ou (2001) and Hertel et al. (2003) find that peer recognition and identification with the goals of the project are the main motivations for developers who contribute to open source software projects. In particular, Hars and Ou's (2001) survey conducted among OSS programmers revealed that peer recognition was an important motivating factor for 43% of the respondents, while community identification was a key factor for 28% of the respondents. Similarly, Hertel et al.'s (2003) survey of 141 contributors to the Linux kernel project found that a prime motivating factor is 'identification with Linux kernel'.

Hann et al. (2004) empirically examined the Apache HTTP Server Project and found that contributions were not correlated with higher wages, but that a higher ranking within the Apache Project was indeed positively correlated with higher wages. Using a Web-based survey, Lakhani and Wolf (2005) found that intrinsic motivations help induce developers to contribute to OSS. Chakravarty et al. (2007) found that the motivation of OSS programmers depends both on private motivations (like future monetary payoffs or ego) and social motivations (like altruism).

Licensing of Open Source Software

Like other products based on intellectual property, the intellectual property in software is typically 'licensed' for use, not sold outright. This is the case regardless of whether the software is proprietary or open source. Even though open source software is distributed freely without payment, the programs are distributed under licensing agreements. There are several different types of open source licence. The main difference is the degree of restrictions they entail.

Reciprocal (or viral) licences require that modifications to the program also be licensed under the same licence as the original work. Examples of reciprocal licences are the GNU General Public License (GPL) and the GNU Lesser General Public License (LGPL). The most popular open source licence is the GPL. If a software program is distributed under a GPL, the source code must be made available to users. Further, programs that incorporate code from a software project employing a GPL also must ensure that the source code is available. The GPL is, hence, a very restrictive licence and it is difficult to develop commercial products under a GPL licence (the LGPL is also quite restrictive, but less so than the GPL).

More permissive (non-viral) licences enable redistribution under a small set of rules. Under these licences, the software can be modified without making the new source code available publicly as long as the proper attribution is given. Examples of such licences include the Berkeley Software Development (BSD) license, the

Apache License and the Mozilla Public License. Commercial products can be developed using software licensed under a BSD-type licence as long as credit for the underlying code is given to the copyright holder(s).

Many open source programs employ restrictive licences that would seem to hinder commercial development, since these licences require that all 'future' software using the relevant code must also be in the public domain.

Several papers in the literature have empirically examined the effect of different licences. Bonaccorsi and Rossi (2002) surveyed Italian firms that use open source software and found that, on average, firms that employ software with restrictive licences supply fewer proprietary products than firms that employ software with less restrictive licences.

The remaining papers we survey in this section come from the very detailed data that are publicly available at SourceForge. Project-level data include the 'names' of contributors, their role in the project, who contributed each part of the code, when the development took place, the stage of development, communications among project members, how bugs were fixed, how many times the project was downloaded the intended audience, type of licence, operating system etc.

Lerner and Tirole (2005b) examine the choice of licences using the database of open source projects from the SourceForge web site. They find that open source projects that run on commercial operating systems and projects that are designed for developers tend to use less restrictive licences, while projects that are targeted for end users tend to use more restrictive licences.

Fershtman and Gandal (2007) find that output per developer is much higher in OSS projects with less restrictive licences. This is striking, since the type of licence does not technically affect the writing of the code. This result is consistent with the hypothesis that the main motivation of programmers to contribute to restrictive OSS projects is to be included in the 'list of contributors': programmers have a strong motivation to contribute until the threshold level, and weak motivation to contribute above that level. Comino et al. (2007) find that the more restrictive the licence, the lower the probability that the project will reach an advanced development stage.

Changes in the Open Source Model: Firm Participation

Increased Firm Participation in Open Source Projects

The degree of reliance on unpaid programmers has changed over time. Today, more of the work on open source projects is done by contributors who work for firms. Using a sample of 100 open source projects hosted at Sourceforge.com, Lerner et al. (2006) find that the share of corporate contributors is higher in larger open source projects, where large means more lines of code.

Open Source and Proprietary Software in Same Market

Several open source products have had great success. Indeed, in most software markets, open source and proprietary products compete side by side. In many of these markets, open source products have a non-trivial market share, as the following examples show:

- Web browsers: according to W3Counter, in April 2011, Firefox had 29.5% of the web browser market. (The market data are from W3Counter; see http://www.w3counter.com/globalstats.php, accessed 19 May 2011.)
- Web servers: Apache has been the dominant system in this market for many years. As of May 2011, Apache served approximately 63% of all websites (see https://en.wikipedia.org/wiki/Apache_HTTP_Server, accessed 19 May 2011).
- Server operating system market: according to IDC (2008), as cited by Llanes and De Elejalde (2009), Linux had 13.7% of the server operating system market. (According to W3Counter, in the overall operating system market, Linux held a 1.41% share in April 2011; see http://www.w3counter.com/globalstats.php.)
- According to Trefis (see https://www.trefis.com/company?article=12891#, accessed 31

January 2011), MySQL, a database management system, had approximately a 20% market share in database installations worldwide in 2010.

Recent theoretical work examines this phenomenon as well. See for example Casadesus-Masanell and Ghemawat (2006), Economides and Katsamakas (2006), Athey and Ellison (2009), and Llanes and De Elejalde (2009).

Towards Mixed-Source Strategies

A key change over time in the open source model is that many proprietary firms now initiate open source projects themselves, in addition to supplying programmers. Indeed, many proprietary firms now use a mixed source model, that is, a model in which some of their products are proprietary and are distributed under traditional licences, while others are open source and distributed under an open source licence. Such a mixed source strategy enables firms to benefit from the advantages of both open source and proprietary development. One key advantage to open source software development is that because the code is developed in the public domain, problems (bugs) can be found and solved quickly.

In a huge survey of more than 2300 companies in 15 countries, Lerner and Schankerman (2010) found that more than 25% of all firms surveyed develop both open source and proprietary software. Using data on 73 Finnish software companies, Koski (2005) empirically examined which factors affect whether the firm releases its product using an OSS or proprietary licences. She found that the more service oriented the firm is, the more likely it will be to offer products using OSS licences.

Institutional Changes in the Open Source Model

(This section draws heavily from Greenstein (2011) and comments and suggestions made by Greenstein.)

Rules for participation and governance in open source software projects have changed over time. Initially, open source projects were rather informal organisational processes. While some open source projects still allow unrestricted participation, many do not. In addition to rules regarding participation, open source projects typically have rules for deciding versions, and rules about reuse.

The institutional setting in which open source development takes place has also evolved over time. Sourceforge, which we discussed above, is clearly not the only setting in which open source development occurs. Indeed, Sourceforge is an ideal platform when an open source project lacks an institutional home. But, there are many important cases in which open source projects are hosted within an institutional setting. Linux operates within a consortium supported by many firms – and senior personnel receive salaries from the organisation. In other cases, firms sponsor open source projects – WebKit, which received financing from Apple, is an example (see West and O'Mahoney (2008) for work in this area).

Open source has also become a part of standard development by standard-setting organisations (SSOs.) The Internet Engineering Task Force (IETF) is essentially both an open source organisation and an SSO (Bradner, 1999).

Open Source Software and Incentives for R&D

(This section draws from Maurer and Scotchmer (2006).)

Incentives for engaging in R&D are quite different for open source software than in traditional proprietary software. Under the latter development method, products are often protected by patents and copyrights, which do not typically require disclosure of the source code. Hence intellectual property laws provide protection against imitation. Since open source software is typically put into the public domain, open source software would not provide innovation incentives when the goal is to prevent imitation.

However, as Bessen and Maskin (2006) note, imitation of a discovery can be desirable in a

world of sequential/cumulative and complementary innovation because it helps the imitator develop further inventions. Since a non-trivial amount of software innovation is either sequential/cumulative or complementary (or both), this suggests that the open source development method may be socially preferable. Interestingly, Maurer and Scotchmer (2006) argue that open source development can also be privately preferable to traditional intellectual property protection when innovation is either sequential/cumulative or complementary. Open source development also has implications for the cost of R&D. Open source development can be thought of as 'pooled' R&D, which typically implies cost savings – see West and Gallagher (2006) Firms share code to test software, fix bugs and make improvements – see Rossi and Bonaccorsi (2005). Without open source, they would have to do this independently, which would imply duplicated costs.

Empirical research in this area is at a nascent stage. Using the data at SourceForge, Fershtman and Gandal (2011) find empirical support for the existence of knowledge spillovers among open source projects. The paper shows that the structure of the project network is associated with project success and that there is a positive association between project closeness centrality and project success. This suggests the existence of both direct and indirect project knowledge spillovers among open source software projects.

Open Source More Broadly Defined: Digital Content

(This section draws heavily from comments and suggestions made by Greenstein.)

Open source has spread well beyond the field of software development. Digital content is one area where open source has made major impacts. Creative Commons, which developed a way to help creators of content grant various degrees of copyright permissions to their work, is one of the most important outgrowths of the open source movement. Creative Commons licenses enable those who develop content to choose among a range of copyright protection, from 'all rights reserved' (full protection), via 'some rights reserved', to 'no rights preserved'. Several key institutions use Creative Commons licenses. Wikipedia, the incredibly successful online encyclopedia, started with a variant of a GPL licence for text, and then adopted 'Creative Commons' methodology (The 'wiki' concept was developed in 1995 by a software engineer named Ward Cunningham. Wikis were developed in order to fix bugs in software development, but are now applied to many other applications – see Greenstein (2011). Wikipedia recently celebrated its tenth birthday. According to *The Economist*, it now has over 17 million articles (3.5 million in English). The content is created and edited by users. It was ranked as the Internet's top research site in 2005, and consistently has been and continues to be one of the most popular websites. Currently it is used by a staggering 400 million users each month. (See 'Wiki birthday to you – a celebration of an astonishing achievement, and a few worries', *The Economist*, 13 January 2011.) Some YouTube and Flickr users share their content using Creative Commons licenses. The success of Wikipedia and other digital content providers using open source methodology shows that the open source model continues to evolve and will likely continue to be an important part of the digital economy.

See Also

- ▶ Computer Industry
- ▶ Information Technology and the World Economy

Acknowledgments We are especially grateful to Shane Greenstein for many comments and suggestions that significantly improved the manuscript. We are also grateful to Jacques Lawarree, and Nick Tsilas for very helpful comments and suggestions. An academic research grant from Microsoft is gratefully acknowledged. Any opinions expressed are those of the authors.

Bibliography

Athey, S., and G. Ellison. 2009. *Dynamics of open source movements*. Mimeo.

Bessen, J., and E. Maskin. 2006. *Sequential innovation, patents, and innovation*. http://econpapers.repec.org/paper/clanajeco/321307000000000021.htm

Boldrin, M., and D.K. Levine. 2009. Market structure and property rights in open source. *Washington University Journal of Law & Policy* 30: 325.

Bonaccorsi, A., and C. Rossi. 2002. *Licensing schemes in the production and distribution of open source software: An empirical investigation*. http://opensource.mit.edu/papers/bnaccorsirossilicense.pdf

Bradner, S. 1999. The internet engineering task force. In *Open sources: Voices from the revolution*, ed. C. DiBona, S. Ockman, and M. Stone. Sebastopol: O'Reilly Media Inc.

Chakravarty, S., E. Haruvy, and F. Wu. 2007. The link between incentives and product performance in open source development: An empirical investigation. *Global Business and Economics Review* 9: 151–169.

Casadesus-Masanell, R., and P. Ghemawat. 2006. Dynamic mixed duopoly: A model motivated by Linux vs Windows. *Management Science* 52(7): 1072–1084.

Comino, S., F. Manenti, and M. Parisi. 2007. On the success of open source projects. *Research Policy* 36: 1575–1586.

DiBona, C., S. Ockman, and M. Stone, eds. 1999. *Open sources: Voices from the revolution*. Sebastopol: O'Reilly Media Inc.

DiBona, C., D. Cooper, and M. Stone. 2006. *Open sources 2.0: The continuing revolution*. Sebastopol: O'Reilly Media Inc.

Economides, N., and E. Katsamakas. 2006. Two-sided competitions of proprietary vs. open source technology platforms and implications for the software industry. *Management Science* 52(7): 1057–1071.

Fershtman, C., and N. Gandal. 2007. Open source software: Motivation and restrictive licensing. *International Economics and Economic Policy* 4: 209–225.

Fershtman, C., and N. Gandal. 2011. Direct and indirect knowledge spillovers: The 'social network' of open source projects. *RAND Journal of Economics* 42: 70–91.

Ghosh, R., R. Glott, B. Krieger, and G. Robles. 2002. Free/libre and open source software: Survey and study, Part IV: Survey of developers. *FLOSS Report*, International Institute of Infonomics, University of Maastricht, The Netherlands. http://www.infonomics.nl/FLOSS/report/Final4.htm

Greenstein, S. 2011. Innovative conduct in U.S. commercial computing and Internet markets. In *Handbook on the economics of innovation,* ed. B. Hall and N. Rosenberg, Forthcoming.

Hann, I., J. Roberts, S. Slaughter, and R. Fielding. 2004. *An empirical analysis of economic returns to open source participation*. http://wwwrcf.usc.edu/hann/publications_files/economic_returns_to_open_source_participation.pdf

Harhoff, D., J. Henkel, and E. von Hippel. 2003. Profiting from voluntary spillovers: How users benefit by freely revealing their innovations. *Research Policy* 32: 1753–1769.

Hars, A., and S. Ou. 2001. Working for free? Motivations for participating in open source projects. *International Journal of Electronic Commerce* 6: 25–39.

Hertel, G., S. Niedner, and S. Herrmann. 2003. Motivation of software developers in open source projects: An internet-based survey of contributors to the Linux kernel. *Research Policy* 32: 1159–1177.

Johnson, J. 2002. Open source software: Private provision of a public good. *Journal of Economics & Management Strategy* 11: 637–662.

Johnson, J. 2006. Collaboration, peer review and open source software. *Information Economics and Policy* 18: 477–497.

Koski, H. 2005. OSS production and licensing strategies of software firms. *Review of Economic Research on Copyright Issues* 2: 111–125.

Lakhani, K., and R. Wolf. 2005. Why hackers do what they do: Understanding motivation and efforts in free open source projects. In *Perspectives on free and open source software*, ed. J. Feller/Open, B. Fitzgerland, S. Hissam, and K. Lakhani. Cambridge, MA: MIT Press.

Lerner, J., P. Parag, and J. Tirole. 2006. The dynamics of open-source contributors. *American Economic Review Papers and Proceedings* 96: 114–118.

Lerner, J., and M. Schankerman. 2010. *The comingled code: Open source and economic development*. Cambridge, MA: MIT Press.

Lerner, J., and J. Tirole. 2002. Some simple economics of open source. *Journal of Industrial Economics* 52: 197–234.

Lerner, J., and J. Tirole. 2005a. The economics of technology sharing: Open source and beyond. *Journal of Economic Perspectives* 19: 99–120.

Lerner, J., and J. Tirole. 2005b. The scope of open source licensing. *Journal of Law, Economics, and Organization* 21: 20–56.

Llanes, G., and R. De Elejalde. 2009. *Industry equilibrium with open source and proprietary firms*. Mimeo, http://www.hbs.edu/research/pdf/09-149.pdf

Maurer, S., and S. Scotchmer. 2006. Open source software: The new intellectual property paradigm. *Economics and Information Systems* 1: 285–322.

Raymond, E. 2000. *The cathedral and the bazaar*. http://www.catb.org/~esr/writings/cathedral-bazaar/cathedral-bazaar/

Rossi, C., and A. Bonaccorsi. 2005. Intrinsic vs. extrinsic incentives in profit-oriented firms supplying open source products and services. *First Monday*, Issue 10, No. 5. http://pear.accc.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/1242

Stallman, R. 1999. The GNU operating system and the free software movement. In *Open sources: Voices from the revolution*, ed. C. DiBona, S. Ockman, and M. Stone. Sebastopol, CA: O'Reilly Media Inc.

von Krogh, G., and E. von Hippel. 2006. The promise of research on open source software. *Management Science* 52: 975–983.

West, J., and S. Gallagher. 2006. Challenges of open innovation: The paradox of firm investment in open source software. *R&D Management* 36: 315–328.

West, J., and S. O'Mahoney. 2008. The role of participation architectures in growing sponsored open source communities. *Industry and Innovation* 15(2): 145–168.

Open-market Operations

Stephen H. Axilrod and Henry C. Wallich

An open-market operation is essentially a transaction undertaken by a central bank in the market for securities (or foreign exchange) that has the effect of supplying reserves to, or draining reserves from, the banking system. Open-market operations are one of the several instruments – including lending or discount-window operations and reserve requirements – available to a central bank to affect the cost and availability of bank reserves and hence the amount of money in the economy and, at the margin, credit flows.

Theory and Function

A distinctive feature of open-market operations is that they take place at the initiative of the monetary authority. They provide a means by which a central bank can directly and actively affect the amount of its liabilities for bank reserves, increasing them by purchases of securities and decreasing them by sales. With reserve provision at the initiative of the central bank, open-market operations facilitate control of the money supply and, from a short-run perspective, the pursuit of a stabilizing economic policy by the central bank and, from a longer-run perspective, of an anti-inflationary policy.

By contrast, in the operation of a central bank's lending function, the provision or liquidation of reserves is at the initiative of private financial institutions. To the extent that this facility is employed too actively or not actively enough and is not offset by the central bank through open-market operations, or by an appropriate discount rate, control of the volume of reserves, and, ultimately, the money supply, is weakened.

When open-market operations play the primary role in monetary-policy implementation, such as in the United States, the discount window still serves an important function in the monetary process. Indeed, in the short run, demands at the discount window are not independent of the amount of reserves supplied through the open market. For example, as a central bank restrains reserve growth by holding back on security purchases, some of the unsatisfied reserve demand will at least for a time shift to the discount window. In general, changes in the demand for borrowing will in practice provide some offset to provision of reserves through open-market operations. This may make it more difficult for a central bank to control bank reserves precisely through open market operations. However, precise control is probably not desirable in the short run because demands for, and needs for, money and credit in dynamic, highly active economies are quite variable. In that sense, the discount window provides a safety valve through which reserves can be provided to maintain a suitably elastic currency and to avert disorderly market conditions.

An open-market purchase essentially replaces an interest-earning asset on the books of banks (either a government security or a loan to some entity holding a government security) with a claim on the central bank – that is, with a reserve balance that has been created for the purpose of acquiring the security. This reserve balance is then 'excess' to the banking system. In the process of converting these non-interest-earning excess balances into interest-earning assets, banks will in turn make loans or purchase securities. That will tend to keep interest rates lower than they otherwise would be and lead to an expansion of the money supply through the well-known multiplier process as the original excess reserves turn into required reserves. The associated amount of money will be a multiple of the amount of reserves, with the multiple depending on the required reserve ratio and on the amount of excess reserves banks in the end want to hold at given levels of interest rates.

The power of open-market operations as an instrument of policy does not, however, depend in its essentials on banks being required to hold reserves or being required to hold a high or low fraction of deposits as reserves at the central bank. That might affect to a degree the precision of the relationship between open-market operations and

money. But the power of open-market operations to influence the economy derives essentially from the ability of the central bank to create its own product – whether it takes the form of bank reserves, clearing balances, or currency in circulation – without the need to take account of the circumstances that influence ordinary business decisions, such as costs of materials, profit potential, and the capacity to repay debt incurred. Even in the unlikely event that the banks, in the absence of reserve requirements, chose to carry zero reserves and to rely entirely on the discount window, the central bank would have large liabilities outstanding in the form of currency through which it could exert pressure on banks by means of open-market operations. In the United States about three-quarters of the central bank's assets reflect currency liabilities.

In a sense, the product of open-market operations – the non-borrowed portion of the monetary base (roughly the sum of the central bank's deposit or reserve liabilities plus currency in circulation) – can be viewed as being created from outside the economy. It is 'outside' money, exogenous to the economic process, but capable of strongly influencing that process. If the central bank continues to create a product for which there is no need or which the participants in the economy do not wish fully to accept, the economy will devalue that product; excessive money creation will cause the price of money relative to other products to fall. That effectively occurs through a rise in the general price level domestically and devaluation of the currency internationally.

Open-market operations in those countries which have sufficiently broad and active markets so that they can be the central instrument of policy are of course attuned to the nation's ultimate economic objectives and to the intermediate guides for over-all monetary policy used to accomplish these objectives; these guides may encompass money supply, interest rates, or exchange rates. In implementing policy on a day-to-day basis, however, open-market operations require additional guides in those cases where the intermediate objectives of policy are quantities, such as the money supply, that are not directly controllable through the purchase or sale of securities.

In most countries, operations are guided on a day-to-day basis by some view of desirable tautness or ease in the central money market, as judged by an appropriate short-term interest rate, complex of money-market rates, or degree of pressure on the banking system. As money-market and bank reserve pressures change, the banking system, financial markets generally, and the public make adaptations – through changes in interest rates broadly, lending terms and conditions, liquidity and asset preferences – that with some lag lead to attainment of money supply objectives or economic goals more broadly.

It has been argued, chiefly by those who would like policy to focus more or less exclusively on a money supply intermediate target, that open-market operations should be guided not by money-market conditions or the degree of pressure on the banking system but by the total quantity of reserves or monetary base. Because open-market operations are at the initiative of the central bank, they are construed as especially well suited to attainment of such quantitative reserve objectives. Reserves or the monetary base as a guide are thought to bear a more certain relationship to a money-supply intermediate guide than do money-market conditions because the former depend on the multiplier relationship between reserves or the base and money and not on predicting how markets and asset holders will react to a given change in interest rates. However, in practice the multiplier relationship itself is variable (in part because of varying reserve requirements or reserve balance practices behind differing deposits in measures of money) and is not independent of interest rates; for instance, rates affect the demands for both excess and borrowed reserves.

Techniques

A variety of techniques are available to implement open-market operations. Securities can be purchased or sold outright. The securities may be short- or long-term, although because short-term markets are generally larger and more active most transactions take place in that market.

Open-market transactions may also be undertaken through, in effect, lending or borrowing operations – by purchasing a security with and agreement to sell it back, say, tomorrow or in a few days, or by selling a security with an agreement to buy it back shortly. Whereas outright transactions take place at current market rates on the securities involved, these combined purchase and sale transactions (termed repurchase agreements) yield a return related to the going rate on collateralized short-term loans in the money market. Repurchase agreements have the advantage of greater flexibility. When they run out, after being outstanding overnight or for a few days only, reserves are withdrawn or provided automatically. Outright purchases or sales create or asorb permanent reserves requiring more explicit action to reverse.

Open-market operations are generally conducted in governmental securities, since that is usually the largest and most liquid market in the country. In the United States, domestic operations are confirmed by law to US government or federal agency securities and all operations must be conducted through the market; purchases cannot be made directly from the government. A large, active market is essential if the central bank is to be able to effect transactions at its own initiative when and in the size required to meet its day-to-day objectives.

The traditional responsibility of central banks for maintaining the liquidity of markets and averting disorderly conditions affects methods of open market operations. In the United States, the bulk of day-to-day open-market operations are undertaken to offset variations in such items as float, the Treasury cash balance, and currency in circulation that affect the reserve base of the banking system. On average per week, such factors absorb or add about 4 per cent (the equivalent of $1 1/2 billion) of the reserve base in the United States. Without offsetting open market operations – sometimes termed 'defensive' operations – typically undertaken through repurchase transactions, money-market conditions and rates would tend to vary sharply from day to day, unduly complicating private decision-making and possibly frustrating the central bank's purposes with respect to controlling the growth of the money supply or the level of interest or exchange rates.

Open-market operations conceptually can be employed to affect the yield curve – for example, to maintain short-term rates while exerting downward pressure on long-term rates. By shifting the composition of its portfolio, the central bank can change the supply of different maturities in the market. An effort to do this in the United States in the early 1960s was not clearly successful. In part, this may be because such an operation requires active cooperation by the Treasury. More fundamentally, most economists have come to believe that expectations so dominate the term structure of interest rates that the central bank, even if aided by a like-minded governmental debt management, would have to engage in massive changes in the maturity structure of securities in the market in order to produce more than a small impact on the shape of the yield curve.

Apart from operations in governmental securities, some central banks undertake open-market operations in foreign exchange. This may be done in an effort to influence the course of exchange rates while at the same time offsetting any effect on bank reserves or other money-market objectives – termed sterilized intervention. However, in certain countries the foreignexchange market may be the chief avenue available for open-market operations, as is typically the case for small countries in which foreign trade represents a large fraction of their gross national product and exchange-market transactions a large portion of total activity in the open market. In such cases, open-market operations in foreign exchange also tend to affect the bank reserve base either because there is little scope to offset them through domestic markets or because there is little desire to do so if the central bank has a relatively fixed exchange rate objective.

Relation to Governmental Budgetary Deficits

There is no necessary relationship between open-market operations and the financing of budgetary deficits. In a country like the United States open-

market purchases are undertaken only as needed to meet money supply and overall economic objectives; they are not increased because a budgetary deficit is enlarged nor decreased when a deficit diminishes. The government must meet its financing needs by attracting investors in the open market paying whatever market interest rate is necessary.

An enlarged deficit would itself lead to increased open-market purchases only if the monetary authority deliberately adjusted its objectives to permit an expansion of bank reserves and money to help finance the government, in which case the deficits would indirectly, through their influence on monetary-policy decisions, lead to inflationary financing. This occurred as a means of war finance during World War II in the United States when the central bank purchased government securities from the market at a fixed ceiling price, thus in effect monetizing the debt; price controls were employed in an effort to suppress the inflation.

But since the early 1950s, debt finance by the US government has had to meet the test of the market unaided by central-bank open-market purchases. Confidence that the central bank will not finance the government deficit is essential to a sound currency. A financial system in which the central bank can refuse to fund the deficit also can provide a powerful incentive to keep deficits from burgeoning.

The typical instances of central-bank monetization of the debt in recent decades have occurred in countries – usually not highly developed ones – with persisting large budgetary deficits who are unable to attract private investors at home or abroad because interest rates offered are artificially low, or the domestic market is undeveloped, or because of a lack of confidence in the security and the currency domestically or on the part of foreign investors. The central bank is then more or less forced to acquire securities directly from the government, automatically creating reserves and money, and leading to inflation and perhaps hyperinflation as the process continues. In those cases, a halt to monetization of the debt through central bank purchases depends essentially on greatly reducing, if not eliminating, budgetary deficits.

It must be recognized, to be sure, that a government deficit may lead to pressures on interest rates that the central bank, usually ill-advisedly, may wish to resist. By encouraging expansion of bank credit and money supply, through open-market operations or otherwise, the central bank may indirectly finance a deficit.

The number of countries in which effective open-market operations can be conducted is surprisingly limited. Required is a securities market sufficiently deep so that the central bank can make purchases and sales sufficient to achieve its reserve objectives without significantly affecting the price of the securities. Otherwise it will be constrained by fear of unintended price effects and would in any event be engaging more in interest-rate manipulation than in control of reserves. Such comparatively price-neutral operations, if possible at all, are feasible usually at the short rather than at the long end. It requires a market for Treasury bills or similar instruments such as exists in, for instance, the United States, the United Kingdom and Canada, but that does not at this time in, for example, Germany and Japan. In the United Kingdom open market operations, in former years, were conducted in Treasury bills; today, commercial bills are primarily employed. Thus, central banks have varying capacities to undertake open market operations; in some cases, where markets for short-term instruments are limited, operations may not be entirely at the initiative of the central bank, nor at a market price in contrast to one set by the central bank (although the price set by the central bank may be based on market conditions).

See Also

► Money Supply

Bibliography

Bank of England. 1984. *The development and operation of monetary policy, 1960–1983*, 156–164. Oxford: Clarendon Press.

Board of Governors of the Federal Reserve System. 1984. *The Federal Reserve System: Purposes and functions*, 7th edn. Washington, DC.

Federal Reserve Bank of New York Quarterly Review 10(1), Spring 1985, 36–56. (Reports for earlier years are available in the same publication.

Meek, P. 1982. *US monetary policy and financial markets*. New York: Federal Reserve Bank of New York. Monetary policy and open market operations in 1984.

Operations Research

Ilan Vertinsky

Abstract

Operations research (OR) is both a profession and an academic discipline. It involves the application of advanced analytical methods to improve executive and management decisions. This survey highlights the types of OR models and techniques in common use. It explores the roots of OR and its theoretical and professional evolution, and presents the current trends which shape its future.

Keywords

Allocation problems; Chance-constrained programming; Computational methods; Critical path method; Dantzig, G.; Data mining; Dual method; Dynamic pricing; Dynamic programming; E-commerce; Financial engineering; Game theory; Globalization; Graph theory; Information technology; Integer programming; Inventory theory; Kantorovich, L. V.; Lattice programming; Linear programming; Marketing engineering; Markov processes; Multi-criteria programming; Network flow optimization; Operations research; Polynomial algorithms; Polynomial submodular set functions; Probability theory; Production control theory; Program evaluation and review technique; Quadratic programming; Quesnay, F.; Queuing theory; Revenue management methods; Simplex method for solving linear programs; Simulation; Stochastic programming; Supermodularity; Von Neumann, J.; Walras, L

JEL Classifications

C44

Operations research is commonly referred to as OR. In the United Kingdom, where the first formally recognized group of practitioners was formed, it is called 'operational research'. Other names, such as 'management science', 'operational analysis' and 'systems analysis', are frequently used as synonyms.

Definitions of OR abound. The differences among these definitions reflect important dimensions of conflict in philosophy and perception of the field among the members of the various communities identifying themselves as operations researchers. It is instructive, therefore, to examine some of these definitions to identify areas of agreement about the distinctive characteristics of the field as well as those dimensions which are a cause of tension.

The Operational Research Society of the UK, the oldest OR professional society, developed the following official definition (Dando and Sharp 1978, p. 940):

> Operational Research is the application of the methods of science to the complex problems arising in the direction and management of large systems of men, machines, materials and money in industry, business, government and defense. The distinctive approach is to develop a scientific model of the system, incorporating measurements of factors such as chance and risk, with which to predict and compare outcomes of alternative decisions, strategies or controls. The purpose is to help management determine its policy and actions scientifically.

Its current website suggests that OR is the discipline of 'applying advanced analytical methods to make better decisions' (the OR Society). While these definitions see OR as an eclectic, problem-centred approach where scientific methods are employed to help management, definitions proposed in the United States view OR as a science or as a distinctive methodology providing scientific bases for decision-making. The constitution of the Operations Research Society of America (ORSA) referred to OR as 'the science of operations research' (House 1952, p. 28). This view was incorporated in 1982 in the Decision and Management Program of the U.S. National

Science Foundation that referred to the emergence of a combined theoretical and empirical science of operational and managerial processes (Little 1986). The current website of the Institute for Operations Research and Management Science (INFORMS) which has succeeded ORSA, refers to OR as the 'science of better'.

The scientific view of OR sees its goals as (*a*) the development of models of operations that represent the causal relationship between controlled variables, uncontrolled variables and system performance, and (*b*) the development of the computational means for identifying levels of controlled variables in ways that help managers of a system achieve systems outputs as close as possible to the ones they desire.

A broader and more proactive variant of the scientific view of OR was proposed in the first major textbook to be published on OR (Churchman et al. 1957). It suggested that the goal of OR is an overall understanding of optimal solutions to executive-type problems in organizations. This comprehensive goal of OR implies a normative prescriptive role with boundaries emancipated from mere reactive problem-solving.

Examination of the various definitions of OR establishes the following features upon which there is almost general agreement: (*a*) OR focuses upon executive and management-type decisions in organized systems; (*b*) a distinct feature of the methodologies used in OR is the development of quantitative models which relate controllable and uncontrollable variables to system performance measures; and (*c*) the outputs of OR models are solutions, that is, suggested levels of control variables that meet some prescribed restrictions. In addition, OR attempts to identify those solutions that are 'better' than others or are 'best' given an objective function and the validity of solutions ought to be tested empirically. OR, however, is concerned not only with the derivation of solutions but their relevance to management practice and their implementation.

While OR definitions reflect the ideals and aspirations of many leaders of OR communities, a commonly held view is that OR is a collection of techniques (National Academy of Sciences 1976). In fact, the success in practice of some OR techniques, such as linear programming and the proliferation of accessible optimization packages, is responsible for this perception. In part, it is also a reflection of imbalances in the work of OR academics. Examination of the content of OR journals and textbooks, for example, would support such a proposition. Indeed, much of the academic effort since the mid-1980s has focused on articulating the supporting mathematical theories of OR models, the development of alternative models and computational methods with a glaring absence of empirical testing (Denizel et al. 2003). This, however, was more a result of a natural progression of the life cycle of the field than a paradigm shift.

Disagreements in the OR communities exist with regard to the following questions. First, what is the level of generality that OR models can attain (that is, what are the prospects of OR becoming a science of operations as opposed to an approach to problem-solving in specific organizational contexts)? Second, what is the degree of comprehensiveness of OR missions, in particular the degree to which a systems approach should characterize OR activities (that is, focus of OR methodologies upon overall effects of a proposed solution on an organization rather than a narrower problem-solving focus)? Third, what is the role of interdisciplinary teamwork in OR?

OR Models and Techniques

Models and computational techniques are key elements in the OR methodology. Models in OR, as opposed to models developed by mathematicians, derive their legitimacy from the real world (as in other sciences) and from their potential uses. Thus one can classify OR models and techniques according to the type of management problems or decision areas they deal with.

Some of the characteristic problem areas that have stimulated OR modelling include:

- Allocation problems
- Inventory problems
- Queuing problems

- Scheduling problems
- Competitive problems
- Renewal and replacement problems
- Search problems
- Revenue management
- Supply chain management
- Financial and marketing engineering
- Data mining

Each of these problem areas is characterized by some typical structures which have stimulated the development of certain classes of mathematical models as well as their supporting mathematical theories. Often, however, a type of mathematical model developed for a specific problem area can be used to model processes with similar structures in other problem areas.

Let us consider, for example, allocation problems. These are the typical economic problems of allocating scarce resources between competing demands so as to maximize net benefits. The allocation, however, must satisfy some prescribed constraints. The first primitive mathematical programs were formulated by economists late in the 18th century. The typical structure of mathematical programs is the maximization or minimization of an objective function subject to a set of constraints. The properties of the objective function and the special structures of the constraints determine the methods and difficulty of finding optimal values. For example, *linear programming* postulates a system with a linear objective function, linear constraints and non-negative control variables. Thus sub-areas of mathematical programming designate the mathematical structure of the optimization problem at hand: *integer programming* requires integer solutions; *quadratic programming* postulates a quadratic objective function; *stochastic programming* assumes that stochastic parameters describe the objective function; *chance-constrained programming* assumes that the restrictions on a problem are given as probabilities of satisfying each constraint, and so on.

An interesting allocation problem arises in situations with multiple decision units with separate conflicting objectives, when rules for trade-offs or reconciliation of conflict are not given. This problem led to the emergence of *multi-criteria programming*, a technique that postulates several objective functions subject to a set of joint constraints.

As we indicated, while mathematical programming emerged as a means of dealing with allocation problems, its applications cut across most areas of OR endeavour.

Queuing theory evolved primarily to help design service policies to deal with congestion and waiting lines. The theory has its roots in probability theory. The application of the theory demonstrates well a problem which characterizes many OR models – limited empirical validity. Indeed, in many practical situations, the probability distributions which characterize arrival and service time depart from those postulated by the basic theory. In such cases problems become analytically intractable and simulation techniques are used.

Inventory and production control theory can be divided into the tractable but unrealistic deterministic cases, and the more problematic stochastic cases. The theory has contributed important insights as to the shape of optimal policies, but specific solutions to problems arising in the real world are typically obtained by simulation.

Simulation is indeed the most prolific OR technique. It is used in practice especially to model stochastic processes and provide solutions to analytically difficult or intractable problems. A computer model representing the system provides the vehicle for low-cost, fast experimentation with alternative patterns of control variables.

Competitive problems have led to the emergence of *game theory.* While the theory has had some important applications (for example, designing optimal stable policies of inspections associated with international nuclear-testing restrictions), its restrictive assumptions with respect to the rationality of players have limited its usefulness for modelling many competitive business situations. *Gaming* and simulation techniques are often used to improve strategic decisions in competitive situations.

Scheduling problems are typically modelled as *network flow optimization* problems. Two techniques have received great attention and have

been employed widely in project planning: the program evaluation and review technique (PERT) and the critical path method (CPM). Network flow optimization is used extensively to deal with many transportation and communication problems.

An important area of OR modelling is the area of Markov and related processes. In a Markov process, knowledge of the present makes the future independent of the past. Markov chains have been used extensively in manpower planning.

Dynamic programming is a method of analysing multi-stage decision processes in which each decision in a sequence depends upon those preceding it as well as exogenous factors. The technique reduces significantly the computational effort by eliminating the need to enumerate and consider the consequences of all possible decision sequences. The method is used in a wide variety of problem areas.

In the 1980s *revenue management methods* combining accurate demand forecasts with intelligent dynamic pricing were developed for and adopted by airlines, and their use spread to other sectors. In the 1990s increasing globalization and the emergence of complex business networks of suppliers and producers created the need for better *supply chain management*. Advances in computing power, communications and operation research methods created new modelling opportunities responding to the challenge of finding best overall combinations of suppliers, transportation, production, warehousing and inventory. Recent modelling efforts in this domain incorporate 'game like' situations in cooperative networks where incentives of different participants may be misaligned. The late 1990s saw the emergence of e-commerce and powerful information technology applications in business generating high volumes of customer data. Large, high-quality data stimulated the development of *data mining* techniques to use the data to improve business strategies and operations. The proliferation of personal powerful computers created opportunities for the development of OR applications for a variety of business functions. *Financial and marketing engineering* are examples of OR applications to traditional functional fields of business.

The lack of definite boundaries as to what constitutes OR makes it difficult to determine whether some techniques originating in other fields, but frequently used by OR practitioners, should be designated as OR techniques. Statistical analysis, forecasting methodologies and evaluation techniques are good examples.

The Roots of OR

The beginnings of OR can be traced to the emergence of the executive function and the complex organization brought about by the Industrial Revolution of the 19th century. The mathematical roots of OR can be traced earlier to the work of Quesnay (1759), who formulated primitive mathematical programming models. This fundamental work was followed by the work of Walras (1883), and by the work of von Neumann (1937) and Kantorovich (1939).

The roots of empirical OR can be traced to the scientific management movement. The work of Taylor, Gantt, Emerson and other pioneers of scientific management began around 1885. They proposed that scientific methods of analysis and measurement could and should be used in production management and business decisions. In 1909, Erlang, a Danish mathematician, published his study of traffic congestion in a telephone network, pioneering the modelling of queues. In 1916 Lanchester published his 'N-square law', assessing the fighting power of opposing forces. The theory was tested retrospectively against Admiral Nelson's plan of the battle of Trafalgar.

The appearance of OR as an organized activity is associated with preparation in the UK for the Second World War. In 1936 the British government decided to set up radar stations. The need to study the operational use of radar chains in order to increase their ability to detect aircraft led to the establishment of a study group of scientists called 'the operational research group'. Their success led to the adoption of OR by other branches of the military. In 1942 an OR section was established by the US Air Force. OR was soon adopted by other branches of the US military. Under the aegis of the US Air Force, a team of economists and mathematicians began in 1947 to model the military structure and the economy. During this period

Dantzig (1963) developed the simplex method of solving linear programs.

The Evolution of OR

The diffusion of OR to the industrial world was slow. Only in the early 1950s did the tools and methods of OR begin to be used outside the military. The first important industrial application of OR was the use of linear programming to schedule a petroleum refinery (Charnes and Cooper 1961).

The Operational Research Club of Britain was formed in 1948, and the Operations Research Society of America (ORSA) was established in 1951. Other national societies for OR soon followed, and in 1957 the International Federation of Operational Research Societies was formed. Books, journals and university programmes specializing in OR proliferated in the 1960s. A gradual process of change in the membership of most OR communities started, bringing a shift towards a higher proportion of university-based members. While the ORSA constitution saw as one of its major missions the establishment and maintenance of professional standards of competence in OR, the evolution of the field caused more emphasis to be placed upon the academic mission of the development of methods and techniques of OR. The tension between practice and theory of OR indeed originated in the 1950s and 1960s. It is interesting to note that this period is viewed by some as the best of times for OR (Miser 1978) and by others as the worst of times (Churchman 1979).

The period saw some of the most exciting mathematical developments since the simplex algorithm. Examples are the important paper by Kuhn and Tucker (1951) laying the foundations of nonlinear programming; the paper by Gomory (1958) presenting a systematic computational technique for integer programming; the works of Bellman (1957) developing dynamic programming; the seminal book by Ford Jr. and Fulkerson (1962) articulating network flow optimization; and the volume edited by Arrow et al. (1958) on the mathematical theory of inventory and production processes. Other important developments during the period were the articulation of decision analysis (see, for example, Raiffa 1968), the development of stochastic programming and chance-constrained programming (see, for example, Charnes and Cooper 1959) and the development of the dual method (Lemke 1954) and the linear complementarity algorithm (Lemke 1965).

Yet, despite these developments, Churchman (1979, p. 13) called the period 'dreary', lamenting the separation of theoretical developments from application, describing OR modelling as a 'study of the delights of algorithms; nuances of game theory; fascinating but irrelevant things that can happen in queues'.

The 1970s presented OR with an important mathematical theory – a theory focusing on its bounds rather than promises: the theory of NP-completeness. The theory presents a framework for the identification of bounds on computational efficiencies (Cook 1971; Karp 1972). Important breakthroughs in the early 1980s were associated with possible improvements on the simplex algorithm in solving linear programs – the development of polynomial algorithms by Khachian (1979, 1980) and Karmarkar (1984).

The 1980s also saw a breakthrough development in the inventory management field. Roundy (1985, 1986) found a simple heuristic and proved that it yields schedules within two per cent of the optimal solution; this work anticipated also the coordination problems characterizing supply chain management. The 1990s saw articulation of the general theory of supermodularity and lattice programming pioneered by Veinott, Edmonds and Topkis (see Topkis 1998). The theory provides fundamental insights to certain classes of optimization problems and issues related to monotone comparative statics, fundamental in economic analysis. The development of polynomial submodular set functions – an unresolved problem remaining – was solved simultaneously by Iwata, Fleischer and Fujishige and Schrijver (see Fleischer 2000).

The new millennium also saw a breakthrough in graph theory – the characterization of the strong perfect graphs by Chudnovsky, Robertson, Seymour and Thomas (see Cornuéjols 2003).

Perhaps more important to the future of operations research has been the great progress

achieved since the mid-1990s in computational methods. Advances in computing machinery, software improvements and development combined to increase the practical significance of the various OR methods. The increased speed of computation and the huge increases in computer memory capacity have made it possible to solve much larger problems and use entirely different solution strategies (Bixby 2002). Improved software also allowed also better interface with users, increasing the accessibility of OR methods to a wider population of users.

The scope of OR was enlarged while the cohesiveness of its communities reduced. Fragmentation was identified by many as an explanation of the declining memberships of many OR and management science professional societies. OR appeared to some observers to be 'in danger of losing its identity as a recognized activity and being assimilated into other fields of endeavor' (Bonder 1979, p. 218). Thus, while the power of OR methods and their use increased, the period since the mid-1980s has witnessed some trends which are threatening the identity of OR as a distinct profession.

The Future of OR

The apparent divorce of OR theory from practice and empirical testing led some leaders in the OR community to wonder whether 'the future of OR is past'. The microcomputer revolution has increased the benefit–costs ratios of OR methods and increased the direct access of general business users to OR. OR groups and practitioners, however, have lost some of their unique advantages as gatekeepers to the application of OR methods. Much of the diffusion of OR methods to the industry is now accomplished through the sales of packaged programs, and is marketed through demonstration CDs. Many users of OR methods in business do not consider themselves OR practitioners. Thus, the dispersion of OR practice in business has resulted in a loss of professional identity (Geoffrion 1992).

Loss of professional identity reduces the flow of new recruits to the profession and limits the career opportunities of OR professionals. The success of OR methods may, therefore, entail the decline of the profession. The sustainability and health of the profession depends on its ability to adopt new business models that fit the new environment, turning threats to opportunities for growth.

See Also

- ▶ Computer Science and Game Theory
- ▶ Convex Programming
- ▶ Graph Theory
- ▶ Linear Programming

Bibliography

Arrow, K., S. Karlin, and H. Scarf. 1958. *Studies in the mathematical theory of inventory and production*. Stanford: Stanford University Press.

Bellman, R. 1957. *Dynamic programming*. Princeton: Princeton University Press.

Bixby, R. 2002. Solving real-world linear programs: a decade and more of progress. *Operations Research* 50: 3–15.

Bonder, S. 1979. Changing the future of operations research. *Operations Research* 27: 209–224.

Charnes, A., and W. Cooper. 1959. Chance-constrained programming. *Management Science* 6: 73–79.

Charnes, A., and Cooper, W. 1961. Management models and industrial applications of linear programming, vols. 1 and 2. New York: Wiley.

Churchman, C. 1979. Paradise regained: A hope for the future of systems design education. In *Education in systems science*, ed. B. Bayraktar. London: Taylor and Francis.

Churchman, C., R. Ackoff, and E. Arnoff. 1957. *Introduction to operations research*. New York: Wiley.

Cook, S. 1971. The complexity of theorem-proving procedures. *Proceedings of the Association for Computing Machinery Annual Symposium on the Theory of Computing* 3: 151–158.

Cornuéjols, G. 2003. The strong perfect graph theorem. *Optima: The Mathematical Programming Society Newsletter* 70 (June): 2–6.

Dando, M.R., and R. Sharp. 1978. Operational research in the UK in 1977: The cases and consequences of a myth? *Journal of the Operational Research Society* 29: 939–949.

Dantzig, G. 1963. *Linear programming and extensions*. Princeton: Princeton University Press.

Denizel, M., B. Usdiken, and D. Tuncalp. 2003. Drift or shift? Continuity, change, and international variation in knowledge. *Operations Research* 51: 711–720.

Fleischer, L. 2000. Recent progress in submodular function minimization. *Optima: The Mathematical Programming Society Newsletter* 64 (September): 1–11.
Ford, L. Jr., and D. Fulkerson. 1962. *Flows in networks*. Princeton: Princeton University Press.
Geoffrion, A. 1992. Forces, trends and opportunities in MS/OR. *Operations Research* 40: 423–445.
Gomory, R. 1958. Essentials of an algorithm for integer solutions to linear programs. *Bulletin of the American Mathematical Society* 64: 275–278.
House, A. 1952. The founding meeting of the society. *Journal of the Operations Research Society of America* 1: 18–32.
INFORMS (Institute for Operations Research and Management Science). About operations research. Online. Available at http://www.informs.org/index.php? c = 49&kat = + About + Operations + Research&p = 17|. Accessed 22 Aug 2006.
Kantorovich, L. 1939. *Mathematical methods of organising and planning production*. Leningrad University [in Russian]. Trans. R. Campbell and W. Marlow, Management Science 6(1960): 366–422.
Karmarkar, N. 1984. A new polynomial time algorithm for linear programming. *Combinatorica* 4: 373–395.
Karp, R. 1972. Reducibility among combinatorial problems. In *Complexity of computer computations*, ed. R. Miller and J. Thatcher. New York: Plenum Press.
Khachian, L. 1979. A polynomial algorithm in linear programming. *Doklady Akademii Nauk SSSR* 224: 1093–1096.
Khachian, L. 1980. Polynomial algorithms in linear programming. *Zhurnal vychisditel'noi matematiki i matematicheskoi* 20: 51–68.
Kuhn, H., and A. Tucker. 1951. Nonlinear programming. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, ed. J. Neyman. Berkeley: University of California Press.
Lemke, C. 1954. The dual method of solving the linear programming problem. *Naval Research Logistic Quarterly* 1: 36–47.
Lemke, C. 1965. Bimatrix equilibrium points and mathematical programming. *Management Science* 11: 681–689.
Little, J. 1986. Research opportunities in the decision and management sciences. *Management Science* 32 (1): 1–13.
Miser, H. 1978. The history, nature and use of operations research. In *Handbook of operations research*, ed. J. Moder and S. Elmaghraby, vol. 1. New York: Van Nostrand Reinhold.
National Academy of Sciences. 1976. *Systems analysis and operations research: A tool for policy and program planning for developing countries. Report of an ad hoc panel*. Washington, DC: National Academy of Sciences.
Pocock, J. 1956. Operations research: A challenge to management. In *Operations research*. Special report no. 13. New York: American Management Association.
Quesnay, F. 1759. Tableau économique. In *Paris*, ed. M. Kuczynski and R. Meek, 3rd ed. London: Macmillan. 1972.
Raiffa, H. 1968. *Decision analysis: Introductory lectures on choices and uncertainty*. New York: Addison-Wesley.
Roundy, R. 1985. Effective integer ratio lot-sizing for one warehouse multi-retailer systems. *Management Science* 31: 1416–1430.
Roundy, R. 1986. Effective lot-sizing rule for a multi-product multi stage production/inventory system. *Mathematics of Operations Research* 11: 699–727.
The OR Society. *What operational research is*. Online. Available at http://www.orsoc. org.uk/orshop/(aamlgbmxg1m44xb520nsqw45)/orcontent.aspx?inc = about.htm. Accessed 22 Aug 2006.
Topkis, D.M. 1998. *Supermodularity and complementarity*. Princeton: Princeton University Press.
von Neumann, J. 1937. A model of general economic equilibrium. In *Ergebnisse eines mathematischen Kolloquiums* 8, ed. K. Menger [in German]. Trans. as 'A model of general equilibrium'. *Review of Economic Studies* 13(1945–6): 1–9.
Walras, L. 1883. *Théorie mathématique de la richesse sociale*. Lausanne: Corbaz.
Wolfe, P. 1959. The simplex method for quadratic programming. *Econometrica* 27: 382–398.

Ophelimity

Nicholas Georgescu-Roegen

Ophelimity is a term coined by Vilfredo Pareto (*Cours*, I) from the Greek ωφελίμος (beneficial) to denote 'the attribute of a thing capable of satisfying a need or a desire, legitimate or not'. His reason, invoked by others as well (e.g., Fisher 1906), was that 'utility' usually opposes 'perniciousness' which economic value does not exclude: weapons, addictive drugs, and the like are commodities. But his action had a root in the interminable controversies that surrounded the economic significance of 'utility' ever since the naturalization of that term in political economy.

Utilitas (utilitatis) with its original Latin meaning of usefulness, benefit, advantage, had been used throughout the Middle Ages by political and philosophical writers. In the early 16th century David Hume in a few places used 'utility' as a correlation of pleasure. But in economics, *utilità* was first used by Ferdinando Galiani in his admirable 1751 *Della moneta* with the specific

meaning of 'the aptitude of a thing to procure us felicity'. Galiani thus conceived utility as a physicalist attribute. The introduction of 'utility' as a technical term in English political science was the lifetime work of Jeremy Bentham (1838, in *Works*, I). Like Galiani, he first defined it as 'that property of any object, whereby it tends to produce benefit, advantage, pleasure, good, or happiness'. But in the same breath he equated 'utility' with happiness, a psychic attribute, through his fundamental principle of utility – namely, *the greatest happiness of the greatest number*. Ultimately, Bentham was disturbed by this terminological tangle and protested that he did not find 'a sufficiently manifest connection between the idea of *happiness* and *pleasure* on the one hand, and the idea of *utility* on the other'. Significantly, late in life he even admitted that '*utility* was an unfortunately chosen word', and blamed Etienne Dumont (his former French promoter) for it, saying that he was 'bigoted, old, and indisposed' to novelty. On this Bentham was both unjust and ignorant: in French (as in Italian, too) there is only *utilité* for both utility and usefulness. The difference between the physicalist and the psychic concept was admirably pinpointed in the rather forgotten 1833 lecture of W.F. Lloyd, who argued that 'the utility [usefulness] of corn is the same after an abundant harvest as in time of famine' whereas value [utility] expresses 'a feeling of the mind [which] is variable with the variations of the external circumstances'.

After Lloyd at least, the Anglophone economists lost a great opportunity to remedy the muddle originating from the French language. Instead, as if virtually everyone had believed with Plato's Cratylus that every thing has a 'natural' name, they kept proposing one term after another with the hope of hitting upon it. Expressions of dissatisfaction with 'utility' were *de rigueur*. Senior (1836), for example, rejected the suggestions of 'attractiveness and desirableness' as even more objectionable than 'utility', by which he proposed to denote a feeling of the mind. Most instructively, as late as 1898 Marshall judged that 'Ophelimity,... Agreeability, Enjoyability, Desirability, etc., are not faultless [but] it seems best for the present to adhere to Utility in spite of its faults.'

In his review of Pareto's *Cours*, Fisher also did not fail to criticize 'utility' for its resilient ambiguities, yet defended its use on the basis of its long tradition. But the best proof that the epistemological skeleton was still there is Fisher's own motions of ten years later. In his epochal 1906 monograph he finds fault with 'utility' because 'useful' is the opposite to 'ornamental', because of the awkward 'disutility', and because of the extraneous phrase 'public utility'. In the end he decided to use 'desirability' in preference to 'utility'. But in a later paper Fisher (1918) revealed how absorbing was that baptizing preoccupation. After arguing that his preferred 'desirability' would not, any more than 'utility', do away with the ethical incongruity of including 'undesirable articles, such as whiskey and prostitution', he cast a strong vote for a word he just coined: 'wantability'. And there is no little piquancy in his final proposal, 'wantab' for the unit of wantability.

The fate of 'ophelimity' itself in the hands of Pareto has been tortuous and inconsistent, yet greatly beneficial to the development of economic thought. Pareto was the greatest culprit for those accidents. However, Pareto's principal reason for this terminological innovation was to distinguish by 'ophelimity' the attribute of things possibly desired by an individual from the attribute of things beneficial to society, to the human race, for which he proposed to retain the old term, 'utility'. To wit, a gun belongs to the first, but not to the second category (save in special circumstances), whereas the air, the sunlight though useful to the human race have no ophelimity. Pareto did not discard 'utility'. In its new sense, the term is a pillar of his monumental *Mind and Society* (1916). The snags emerged rather on the economic track.

As mentioned at the outset, according to Pareto's earliest definition ophelimity was a physicalist attribute. Moreover, it was a quantitative one, subject to all the laws of quantity. This idea somehow remained in a fold of his mind even after he came to reject measurability completely. Denoting by one word a relational phenomenon (between a mind and an object), Pareto inevitably fell in the same pitfall as Bentham: without much ado he described ophelimity 'as a properly

subjective and fundamental' attribute. And he went on to define ophelimity in a completely analogous way with Jevons's final degree of utility, and followed with a surprising, yet instructive footnote (1896: § 25n). Arguing in continuation that the increment of ophelimity corresponding to an increment dx_i of commodity X_i is the *quantity* $f_i(x_1, x_2, \ldots, x_n)$, he noted that a function $f_i(x_1, x_2, \ldots, x_n)$ such that $\partial F/\partial x_i = f_i$, for all i's does not always exist. If F exists, as it does if f_i is a function of xi alone, then it measures total ophelimity. To this mathematical remark Fisher (1896) strongly objected, even though it was in perfect order. Its sin was other: it foisted general mathematics upon a particular structure. If total ophelimity *is* a quantity then the existence of F is part and parcel of that postulate. Pareto, however, carefully observed that if F does not exist, 'the ophelimity enjoyed by the individual depends . . . also on the possible combinations', in this way anticipating by ten years the issue of integrability (Pareto 1909, App. §14). Abiding throughout the *Cours* by the cardinal (purely quantitative) nature of ophelimity, Pareto defined weighted elementary ophelimity by f_i/p_i, where p_i is the market price of X_i and then established the famous theorem for the maximum of consumer ophelimity first proved by H. H. Gossen.

But in a letter of 28 December 1899 to Maffeo Pantaleoni (1960), Pareto set forth a novel idea that was to transform radically not only economics but also the other disciplines of man. It was the idea that an individual or any organized group of individuals always chooses as a matter of fact from accessible alternatives that which is preferred to any other, that which has the greatest ophelimity. An important link in this conception is the case of an individual completely unable to choose any alternative. Apart from a delightful drawing of the so-called Buridan's ass between two plates of fruit, Pareto did not elaborate upon this point. Like virtually all after him, he took for granted that between 'preference' and 'non-preference' there must be 'indifference'. Yet the existence of indifference in this case ought to have been explicitly postulated (Georgescu-Roegen 1936), for Pareto's new edifice of indifference curves was founded on it.

> Edgeworth proceeded from considerations of pleasure and its measure to arrive at the indifference curves. I go the reverse way; the indifference curves [which] are the result of experience are my starting point. I proceed from known to unknown (1966, VIII),

a neat description of his new theory that was repeated in all essays after 1900. And he rightly pointed out that the issue of whether or not the utility, the *rareté*, and even the ophelimity (!) are measurable is now idle: there are no more such things that must be measured. In *Manuel* (App.) he showed that we can give arbitrary (but increasing) *indices* to the indifference varieties, each index serving for ordering the ophelimities of the involved commodity combinations. He could thus stress that he moved from utility, to ophelimity, and finally, to indices that free economic theory of all 'metaphysical' ingredients. But then $f(x,y,\ldots,z)$ being an index, an increasing function $F(f)$ of it would serve as well.

What followed has been hard to understand. In all his later theoretical contributions Pareto continued to treat ophelimity as a cardinal entity, just as utility was by his predecessors. In *Manuel* (App.) as well as in the two Encyclopedia articles (1966, VIII) he assumed that any second partial derivative of an ophelimity index has an invariable sign. Curiously also, this peculiar error was detected only years later by Sir John (Hicks and Allen 1934), who simply observed that

$$\partial^2 F(f)/\partial x \partial y = F'\left(\partial^2 f/\partial x \partial y\right) + F''(\partial f/\partial x)(df/\partial y) \qquad (1)$$

hence, the signs of $\delta^2 F/\partial x \partial y$ and $\partial^2 f/\partial x \partial y$ are not necessarily the same. (An analogous statement is true for $\partial^2 F/\partial x^2$.)

Economic theorists have ever since been more respectful of this indeterminacy, albeit not in every case. The exception concerns another innovation of Pareto, the maximum of ophelimity of a community, which he first defined directly with the aid of his box and the now famous condition of Paretian optimum (1906, iii, §116, vi, §32). For the mathematical condition (App. 89) he proposed

$$[F] = \left(\Delta F^1/F_i^1\right) + \left(\Delta F^2/F_i^2\right) + \ldots + \left(\Delta F^n/F_i^n\right) = 0, \quad (2)$$

where F^k is the total ophelimity of the individual k and F_i^k is the elementary ophelimity of X_i The leading idea was that all terms of (2) are homogeneous, each representing an increment of X_i that would increase ophelimity by ΔF. But the operational meaning of $[F]$ is still very obscure. Pareto's clarifications were utterly unsatisfactory wherever he dealt with it (1909, App.; 1916; 1966, VIII). The very few commentators have not improved the situation. None seems to have raised the issues of ophelimity indeterminacy which would naturally come up in connection with $[F]$. The claim of M. Allais (1968, p. 405) that he has computed $d^2[F]$ is unavailing, for he has not considered the functional transformation $F(f)$. However, if $[F]$ is applied to every commodity, the ophelimity indeterminacy is eliminated. Consider the case of two individuals and two commodities; from $[\phi(x, y)] = 0$, $[\psi(w, z)] = 0$ and $x + w = a$, $y + z = b$, it follows

$$\phi_x/\phi_y = \psi_w/\psi_z, \quad (3)$$

which is the equation of the contract curve – the locus of Paretian optima. It may be well to note that his procedure cannot be applied for the optimal distribution of a single commodity. For Bananaland the optimal distribution of bananas requires cardinal and additive utility, a curious result.

See Also

- Pareto as an Economist
- Pareto, Vilfredo (1848–1923)
- Utility

Bibliography

Allais, M. 1968. Pareto, Vilfredo: Contributions to economics. In *International Encyclopedia of the social sciences*, vol. 11, ed. D.L. Sills, 405. New York: Macmillan.

Bentham, J. 1838–43. *The works of Jeremy Bentham*, 11 vols. ed. J. Bowring. New York: Russell, 1962.

Fisher, I. 1896. Review of *Cours d'économie politique*, Tome I, par Vilfredo Pareto, Lausanne: F. Rouge. *Yale Review*, November.

Fisher, I. 1906. *The nature of capital and income*. New York: Macmillan.

Fisher, I. 1918. Is 'utility' the most suitable term for the concept it is used to denote? *American Economic Review* 8: 335–337.

Georgescu-Roegen, N. 1936. The pure theory of consumer's behavior. *Quarterly Journal of Economics* 50: 545–593. Reprinted in N. Georgescu-Roegen, *Analytical Economics: Issues and Problems*, Cambridge, Mass.: Harvard University Press, 1966.

Hicks, J., and R.G.D. Allen. 1934. A reconsideration of the theory of value, Part I. *Economica* 1: 52–76.

Lloyd, W. 1833. *A lecture on the notion of value as distinguishable not only from utility, but also from value in exchange*. Reprinted in *Economic history* (a supplement to the *Economic Journal*) 1, May 1927, 169–183.

Marshall, A. 1898. *Principles of economics*, 4th ed. New York: Macmillan.

Pareto, V. 1896. *Cours d'économie politique professé à l'université de Lausanne*, vol. I. Lausanne: F. Rouge.

Pareto, V. 1909. *Manuel d'économie politique*. Trans. from Italian by A. Bonnet. Paris: Marcel Giard, 1909.

Pareto, V. 1916. *The mind and society*, 4 vols, ed. A. Livingston. New York: Harcourt/Brace, 1935.

Pareto, V. 1960. *Lettere à Maffeo Pantaleoni: 1890–1923*, 3 vols, ed. G. de Rosa. Rome: Banca Nazionale del Lavoro.

Pareto, V. 1966. *Oeuvres Complètes*, 15 vols, ed. G. Busino. Geneva: Droz.

Senior, N. 1836. *An outline of the science of political economy*. New York: Augustus M. Kelley, 1951.

Oppenheimer, Franz (1864–1943)

Nicholas Georgescu-Roegen

Franz Oppenheimer, the son of a rabbi, was born in 1864, in a Berlin suburb. At first he studied medicine which he practised for some time after earning an MD from Berlin University (1895). But his attraction to political and economic matters was already evidenced by his 1896 work about the agrarian settlement reforms. After several other similar works, in 1908 Oppenheimer obtained a doctoral diploma in the field of his new devotion. As was then the custom, he began as a non-salaried lecturer at Berlin University

(1909), from where in 1919 he moved to a chair of sociology and economics at Frankfurt University. The political developments in Germany prompted him during 1933 to move to France, then to Palestine, and finally to the United States. He died in Los Angeles in 1943.

Oppenheimer's intellectual mark was a theory of the state combined with an economic programme based on the ownership of land. He took up a theme entertained by Ludwig Gumplowicz and also aired by S.N. Patten, according to which the state as a social institution originated only 'through the conquest and subjugation' of a peaceful, classless community by a migrant, warrior tribe (1907). Not primitive accumulation through differentiation – Marx's 'fairy tale', but outright conquest was the origin of the state (1903). 'There are two fundamentally opposed ways whereby man [obtains] the necessary means of existence', the economic, by 'one's own labor', and the political, by 'the forcible appropriation of the labor of others'. It is against this complex that Oppenheimer endeavoured, with limited success, to trace the evolution of the state from its genesis to its modern constitutional form. The other signal tenet of Oppenheimer was that all mankind's evil comes from the unequal ownership of land (1896, 1898), probably an influence of H.H. Gossen's well-known programme for land nationalization. He even went so far as to maintain against Malthus that the only cause of population pressure is the rural exodus that floods the cities because of the land monopoly by a few. Mankind's bliss calls for everyone to earn one's means of subsistence by one's own farmstead as, according to him, was then the case in New Zealand and Utah (1898, 1899). In a mode that recalls Colin Clark's well-known exaggeration of the Earth's carrying capacity, Oppenheimer calculated that even though every family possessed enough subsistence land, there would still remain '*twothirds of the planet unoccupied*'.

Oppenheimer characterized himself as a 'liberal socialist', yet his perspective was a mixture of socialism and anarchism. Above all, no agrarian votary went as far as Oppenheimer to believe that in a thorough agrarian society there could be no population problem.

How tenacious was Oppenheimer's attachment to his ideological beliefs was demonstrated by the cooperative agrarian settlements established by him. He even came to live in one of them when poor health compelled him to retire from teaching. Devoted also to Zionism, he was for years the editor of *Palästina* and wrote several related essays.

Oppenheimer expressed his thoughts with stalwart conviction and great vigour. As J.A. Schumpeter judged, 'a man of mark', who did much to keep alive the interest in economic problems.

Selected Works

1896. *Die Siedlungsgenossenschaft: Versuch einer positiven Uberwindung des Kommunismus durch Losung des Genossenschafts Problems und der Agrarfrage*. Jena: Fischer, 1922.

1898. *Grossgrundeigentum und soziale Frage: Versuch einer neuen Grundlegung der Gesellschaftswissenschaft*. Jena: Fischer, 1922.

1899. Die Utopie als Tatsache. *Zeitschrift fur Sozial-Wissenschaft* 2.

1900. *Das Bevolkerungsgesetz des T.R. Malthus und der neueren Nationalokonomie*. Bern/ Leipzig: J. Edelheim.

1903. *Das Grundgesetz der Marxschen Gesellschaftslehre: Darstellung und Kritik*. Jena: Fischer, 1926.

1907. *The state: Its history and development viewed sociologically*. Trans. from the German, Indianapolis: Bobbs-Merrill, 1914.

1908. *Rodbertus' Angriff auf Ricardo's Rententheorie und der Lexis-Diehl'sche Rettungsversuch*. Berlin: Reimer.

1909. *David Ricardo's Grundrententheorie: Darstellung und Kritik*. Berlin: Reimer.

1910. *Theorie der reinen und politischen Okonomie*. Berlin: Reimer.

1912. *Die soziale Frage und der Sozialismus: Eine kritische Auseinandersetzung mit der Maxistischen Theorie*. Jena: Fischer, 1925.

1931. *Erlebtes, Erstrebtes, Erreichtes: Lebenserinnerungen* (Preface by Ludwig Erhard). Dusseldorf: Melzer, 1964.

1932. *Weder Kapitalismus noch Kommunismus*. 2nd edn. Jena: Fischer.
1941. Wages and trades unions. *American Journal of Economics and Sociology* 1(1): 45–77.

References

Aaron, R. 1936. *German Sociology*. Trans. M. and T. Bottomore. Glencoe: Free Press, 1957.
Gerth, H. 1968. Oppenheimer, Franz. In *International encyclopedia of the social sciences*, vol. II. New York: Macmillan.
Heinman, E. 1944. Franz Oppenheimer's economic ideas. *Social Research* 11: 27–39.
Honingsheim, P. 1948. The sociological doctrines of Franz Oppenheimer: An agrarian philosophy of history and social reform. In *An introduction to the history of sociology*, ed. H.E. Barnes. Chicago: University of Chicago Press.
Schultz, B. 1948. *Die Grundgedanken des Systems der theoretischen Volkswirtschaftslehre von Franz Oppenheimer*. Jena: Fischer.

Opportunity Cost

James M. Buchanan

Keywords
Choice; Opportunity cost; Scarcity

JEL Classifications
D0

The concept of *opportunity cost* (or alternative cost) expresses the basic relationship between scarcity and choice. If no object or activity that is valued by anyone is scarce, all demands for all persons and in all periods can be satisfied. There is no need to choose among separately valued options; there is no need for social coordination processes that will effectively determine which demands have priority. In this fantasized setting without scarcity, there are no opportunities or alternatives that are missed, forgone, or sacrificed.

Once scarcity is introduced, all demands cannot be met. Unless there are 'natural' constraints that predetermine the allocation of end-objects possessing value (for example, sunshine in Scotland in February), scarcity introduces the necessity of choice, either directly among alternative end-objects or indirectly among institutions or procedural arrangements for social interaction that will, in turn, generate a selection of ultimate end-objects.

Choice implies rejected as well as selected alternatives. *Opportunity cost is the evaluation placed on the most highly valued of the rejected alternatives or opportunities*. It is that value that is given up or sacrificed in order to secure the higher value that selection of the chosen object embodies.

Opportunity Cost and Choice

Opportunity cost is the anticipated value of 'that which might be' if choice were made differently. Note that it is not the value of 'that which might have been' without the qualifying reference to choice. In the absence of choice, it may be sometimes meaningful to discuss values of events that might have occurred but did not. It is not meaningful to define these values as opportunity costs, since the alternative scenario does not represent a lost or sacrificed opportunity. Once this basic relationship between choice and opportunity cost is acknowledged, several implications follow.

First, if choice is made among separately valued options, someone must do the choosing. That is to say, a chooser is required, a person who decides. From this the second implication emerges. The value placed on the option that is not chosen, the opportunity cost, must be that value that exists in the mind of the individual who chooses. It can find no other location. Hence, cost must be borne exclusively by the chooser; it can be shifted to no one else. A third necessary consequence is that opportunity cost must be subjective. It is within the mind of the chooser, and it cannot be objectified or measured by anyone external to the chooser. It cannot be readily translated into a resource, commodity, or

money dimension. Fourth, opportunity cost exists only at the moment of decision when choice is made. It vanishes immediately thereafter. From this it follows that cost can never be realized; that which is rejected can never be enjoyed.

The most important consequence of the relationship between choice and opportunity cost is the *ex ante* or forward-looking property that cost must carry in this setting. Opportunity cost, the value placed on the rejected option by the chooser, is the obstacle to choice; it is that which must be considered, evaluated, and ultimately rejected before the preferred option is chosen. Opportunity cost in any particular choice is, of course, influenced by prior choices that have been made, but, with respect to this choice itself, opportunity cost is *choice-influencing rather than choice-influenced*.

Other Notions of Cost

The distinction between opportunity cost and other conceptions or notions of cost is best explained in this choice-influencing and choice-influenced classification. Once a choice is made, consequences follow, and these consequences may, indeed, involve utility losses, either to the person who has made initial choice or to others. In a certain sense it may seem useful to refer to these losses, whether anticipated or realized, as costs, but it must be recognized that these choice-determined costs, as such, cannot, by definition, influence choice itself.

A single example may clarify this point. A person chooses to purchase an automobile through an instalment loan payment plan, extending over a three-year period. The opportunity cost that informs and influences the choice is the value that the purchaser places on the rejected alternative, in that case the anticipated value of the objects which might be purchased with the payments required under the loan. Having considered the potential value of this alternative, and chosen to proceed with the purchase, the consequences of meeting the loan schedule follow. Monthly payments must be made, and it is common language usage to refer to these payments as 'costs' of the automobile. The individual will clearly suffer a sense of utility loss as the payments come due and must be paid. As choice-influencing elements, however, these 'costs' are irrelevant. The fact that, in a utility dimension, post-choice consequences can never be capitalized is a source of major confusion.

Economists recognize the distinction being made here in one sense. With the familiar statement that 'sunk costs are irrelevant', economists acknowledge that the consequences of choices cannot influence choice itself. On the other hand, by their formalized constructions of cost schedules and cost functions, which necessarily imply measurability and objectifiability of costs, economists divorce cost from the choice process.

Essentially the same results hold for accountants, who normally measure estimated costs strictly in the *ex post* or choice-influenced sense. Those 'costs' estimated by accountants can never accurately reflect the value of lost or sacrificed opportunities. Numerical estimates could be introduced in working plans for alternative courses of action prior to decision, but such estimates of opportunity costs would be the accountant's measure of the values for projects not undertaken rather than the value of commitments made under the project chosen.

As suggested, choice-influencing opportunity costs exist only for the person who makes choice. By definition, opportunity costs cannot 'spill over' to others. There may, of course, be consequences of a person's choice that impose utility losses on other persons, and it is sometime useful to refer to these losses as 'external costs'. The point to be emphasized is that these external costs are obstacles to choice, and hence a measure of forgone opportunities, only if the individual who chooses takes them into account and places his own anticipated utility evaluation on them.

Opportunity Cost and Welfare Norms

The source of greatest confusion in the analysis and application of opportunity cost theory lies in the attempted extension of the results of idealized

market interaction processes to the definition of rules or norms for decision makers in non-market settings. In full market equilibrium, the separate choices made by many buyers and sellers generate results that may be formally described in terms of relationships between prices and costs. Under certain specified conditions, prices are brought into equality with marginal costs through the working of the competitive process. Further, the general equilibrium states described by these equalities are shown to meet certain efficiency norms.

Prices may be observed; they are objectively measurable. A condition for market equilibrium is equalization of prices over all relevant exchanges for all units of a commodity of service. From this equalization it may seem to follow that marginal costs, which must be brought into equality with price as a condition for the equilibrium of each trader, are also objectively measurable. From this the inference is drawn that, if marginal costs are then measured, 'efficiency' in resource use can be established independently of the competitive process itself through the device of forcing decision makers to bring prices into equality with marginal costs.

The whole logic is a tissue of confusion based on a misunderstanding of opportunity cost. The equalization of marginal opportunity cost with price for each trader is brought about by the adjustments made by each trader along the relevant quantity dimension. The fact that the marginal opportunity costs for all traders are all brought into equalization with the relevant uniform price implies only that traders retain the ability to adjust quantities of goods until this condition is met. There is no implication to the effect that marginal opportunity costs are equalized in some objectively meaningful sense independently of the quantity adjustment to price.

Consider an idealized market for a good that is observed to be trading at a uniform price of $1 per unit. The numeraire value of the anticipated lost opportunity is $1 for each trader. But it is only as quantity is adjusted that the trader can bring the numeraire value of his subjectively experienced and anticipated utility sacrifice into equality with the objectively set price that he confronts. The anticipated value of that which is given up in taking a course of action is no more objectifiable and measurable than the anticipated value of the course of action itself. The two sides of choice are equivalent in all respects.

Independently of market choice, there is no means through which marginal opportunity costs can be brought into equality with prices. Hence, any 'rule' that directs 'managers' in non-market settings to use cost as the basis for setting price is and must remain without content. There is, however, a second equally important criticism of the welfare rule that opportunity cost reasoning identifies, quite apart from the measurability question. Even if the first criticism is ignored, and it is assumed that marginal opportunity cost can, in some fashion, be measured, instructions to 'managers' to use cost to set price must rely on 'managers' to behave, personally, as robots rather than rational utility-maximizing individuals. Why should a 'manager' be expected to follow the rule? Would he not be expected to behave so that marginal cost, that which he faces personally, be brought into equality with the anticipated value of the benefit side of choice? The fact that the 'manager' remains in a non-market setting insures that he cannot be the responsible bearer of the utility gains and losses that his choices generate. His own, privately sensed, gains and losses, evaluated either prior to or after choice, must be categorically different from those anticipated for principals before choice and enjoyed and/or suffered by principals after choice.

Opportunity Cost and the Choice Among Institutions

As noted earlier, in the absence of 'natural' constraints that predetermine allocation, the introduction of scarcity introduces the necessity of choice, either directly among ultimate 'goods' or indirectly among rules, institutions, and procedures that will operate so as to make final allocative determinations. Opportunity cost in the second of these choice-settings remains to be examined. In a sense, the use of institutionalized procedures to generate allocations of scarce resources may

eliminate 'choice' in the familiar meaning used above and is akin in this respect to the 'natural' constraints noted. Results may emerge from the operation of some institutional process without any person or group of persons 'choosing' among endstate alternatives, and, hence, without any subjectively-experienced opportunity cost. Despite the absence of this important bridge between cost and choice in the ordinary sense, however, values may be placed on the 'might have beens' that would have emerged under differing allocations. The patterns of these estimated value losses, over a sequence of institution-determined allocations, may enter, importantly, in a rational choice calculus involving the higher-level choice among alternative institutional procedures for allocation. In this higher-level choice, opportunity cost again appears as the negative side of choice even if 'choice' in the standard usage of the term is not involved in the making of allocations, taken singly.

Consider the following extreme example. There are two mutually exclusive thermostat settings for a building, *High* and *Low*. An institution is in being that uses an unbiased coin to 'choose' between these two settings each day. It is meaningful for an individual to discuss the potential value to be anticipated if the setting is *High* rather than *Low*, even if the individual does not make the selection, individually or as a member of a collective. The setting that is 'chosen' by the coin flip has consequences for individual utility and these consequences may be anticipated in advance of the actual 'choice'. So long as the institutional procedure remains in effect, however, with respect to a single day's selection, the anticipated value lost by one setting of the thermostat rather than the other cannot represent opportunity cost.

Suppose, now, that instead of the unbiased and equally weighted device, the institution in being is one that allows all persons in the building to vote, each morning, on the thermostat setting with the majority option 'chosen' for the day. Assume, further, that the group of voters is large, so that the influence of a single person on the expected majoritarian outcome is quite small. It is important to emphasize that, in this procedure, as with the coin toss, no person really 'chooses' among the alternative end-states. Each voter confronts the quite different, intrainstitutional choice between 'voting for High' and 'voting for Low', with the knowledge that any individual has relatively little influence on the outcome. In the choice that he confronts, the voter cannot rationally take into account the anticipated losses from the ultimate alternatives, either for himself or for others, in any full-value sense of the term. The loss anticipated from, say, a Low thermostat setting may be estimated to be valued at $1,000 for the individual. Yet if he considers himself to have an influence on the outcome of the voting choice only in one case out of a thousand, the expected utility value of the anticipated loss will be only $1 in terms of the numeraire. This $1 will then represent the numeraire value of the *opportunity cost* involved in voting for High.

Since these same results hold, with possibly differing values, for all voters, no one 'chooses' in accordance with fully evaluated gains and losses. 'Choices' emerge from the institutional procedure without full benefit – cost considerations being made by anyone, taken singly or in aggregation. In the relevant opportunity-cost sense, effective choice is shifted to that among alternative institutions. The results of the 'choices' made within an institution over a whole sequence of periods (over many days in our thermostat example) may, of course, become data for the choice comparison among institutions themselves. And, to the extent that the individual, when confronted with a choice among institutions, knows that he is individually responsible for the selection, the whole opportunity cost logic then becomes relevant at the level of institutional or constitutional choice. This result is accomplished, however, only if each person in the relevant community does, in fact, become the chooser among institutional rules. Only if, at some ultimate level of institutional-constitutional choice the Wicksellian unanimity rule becomes operative, hence giving any person potential choice authority, can the opportunity cost of alternatives for choice be expected to enter and to inform individual decisions.

Summary

Opportunity cost is a basic concept in economic theory. In its rudimentary definition as the value of opportunities forgone as a result of choice in the presence of scarcity, the concept is simple, straightforward, and widely understood. In the analysis of choices made by buyers and sellers in the marketplace, the complexities that emerge only in rigorous definition of the concept remain relatively unimportant. But when attempts are made to extend opportunity cost logic to non-market settings, either in the derivation of norms to guide decisions or in application to choice within and among institutions, the observed ambiguity and confusion suggest that even so basic a concept requires analytical clarification.

Bibliography

Alchian, A. 1968. Cost. In *Encyclopedia of the social sciences*, vol. 3. New York: Macmillan.

Buchanan, J.M. 1969. *Cost and choice*. Chicago: Markham, Republished as Midway Reprint, Chicago: University of Chicago Press, 1977.

Buchanan, J.M., and G.F. Thirlby (eds.). 1973. *LSE essays on cost*. London: Weidenfeld and Nicholson. Reissued by New York University Press, 1981.

Coase, R.H. 1960. The problem of social cost. *Journal of Law and Economics* 3: 1–44.

Optimal Control and Economic Dynamics

W. A. Brock

Optimal control methods and the related methods of dynamic programming and the calculus of variations are ubiquitous in the analysis of dynamic economic systems. This is so because the serious modeller of dynamic economic phenomena in positive economics or in welfare economics, in capitalistic economies or in socialist economies is forced to do four things (i) model the restraints that absence of intertemporal arbitrage opportunities places upon the evolution of the economy over time, (ii) relate expectations of future prices to actual past prices and present prices in a useful notion of equilibrium, (iii) model the learning by the economy's participants of relevant parameters in an evolving economy (iv) design the models so they lead naturally to the implementation of received methods of econometrics in order to confront their predictions with data.

For the positive economist the objective is to achieve an analytically tractable framework to explain and organize data.

For the normative economist the objective is to achieve an analytically tractable framework to analyse the following issues detailed below which are central to economics. In order that the welfare conclusions carry conviction with scientists as well as with philosophers, this framework should be compatible with that designed by the positive economist who is disciplined by confrontation with data. Some issues are: (i) Is capitalism inherently unstable or inherently stable? What forces determine the speed of adjustment to (or divergence from) steady state evolution? (ii) Is it possible to decentralize a planned economy with prices or with some other signals? Is decentralization possible with the micro agents needing to know only a *finite* number of prices or other signals at each point in time? (iii) Does speculation serve any socially useful purpose?

Section "The Framework" of this entry exposits an optimal control framework to deal with these issues. The section develops notions of stability that are used in economic dynamics, while section "The Case δ Near Zero" develops the proposition that if agents do not discount the future very much then a centrally planned multi-sector economy is asymptotically stable under general conditions, that is any two trajectories come together rather than diverge as time progresses. The notions of bliss and overtaking

criteria are exposited in these two sections. These notions play a key role in asymptotic stability theory of optimal control.

Section "Some Economic Applications of the Theory" contains a brief exposition of the modern theory of speculative bubbles, manias, and hyperinflations. This theory uses the *necessity* of the transversality condition of optimal control to investigate possible market forces that may temper the inherent instability displayed by the equations for the myopic perfect foresight asset market equations.

Section "Equilibrium Dynamics" reviews an approach to adjustment dynamics and Samuelson's correspondence principle inspired by optimal control methods. The basic idea is to use optimal control and rational expectations to endogenize the adjustment dynamics with respect to (wrt) which the hypothesis of stability is used to place restrictions on comparative statics. In this way one can push the correspondence principle further than the original version, where the dynamics were ad hoc. This is so because endogenized dynamics contain more restrictions linked to tastes and technology than ad hoc dynamics. Finally section "A Summing Up" presents a brief summing up.

The Framework

In continuous time the general optimal control problem is stated thus:

$$V(y,t_0) \equiv \max \int_{t_0}^{T} v(x,u,s)\mathrm{d}s + B[x(T),T], \tag{1.1}$$

$$\text{s.t. } \dot{x} = f(x,u,t), x(t_0) = y. \tag{1.2}$$

where V: $R^n \times R \to R$; f: $R^n \times R^m \times R \to R^n$; v: $R^n \times R^m \times R \to R$; B: $R^n \times R \to R$. Here V is the state valuation function, also called the indirect utility function, starting at state y at time t_0; υ is the instantaneous utility or payoff when the system is in state $x=x(s) \in R^n$ at time s, and control $u=u(s) \in R^m$ is applied at date s; B is a bequest or scrap value function giving the value of the state x (T) at date T; and $\dot{x} \equiv \mathrm{d}x/\mathrm{d}t = f(x,u,t)$ gives the law of motion of the state. The discrete time version of step size h of (1.1) and (1.2) is analogous, with $\dot{x}$ replaced by $(x(t+h) - x(t))/h$, $\int$ replaced by Σ. Under modest regularity conditions the solution to the discrete time problem converges to the solution to the continuous time problem as $h \to 0$. The horizon T may be finite or infinite.

Under regularity assumptions, by dynamic programming the value function V satisfies the Hamilton–Jacobi–Bellman (HJB) equation; furthermore the co-state–state necessary conditions must be satisfied with $p \equiv V_x$:

$$-V_t = \max_u H^*(p,x,u,t) \equiv H^{*0}(p,x,t), \text{(HJB equation)} \tag{1.3}$$

$$H^*(p,x,u,t) \equiv v + pf, \text{(Hamiltonian definition)} \tag{1.4}$$

$$\dot{p} = -H_x^{*0}, \dot{x} = H_p^{*0}, \quad x(t_0) = y, \text{(co - state equations)} \tag{1.5}$$

$$V(x,T) = B(x,T), \quad p(T) = B_x(x,T), \text{(transversality conditions)} \tag{1.6}$$

The variable p is called the costate variable, adjoint variable, or dual variable; and the function H^* is called the Hamiltonian. These variables are introduced for the same reasons and have the same interpretation that Lagrange–-Kuhn–Tucker multipliers are introduced in nonlinear programming. The terminal conditions (1.6) are sometimes called transversality conditions.

Equations (1.3)–(1.6) are the workhorses of optimal control theory. We briefly explain their derivation and meaning here.

Equation (1.1) may be written:

O

$$\begin{aligned} V(y,t_0) &= \max\left\{\int_{t_0}^{t_0+h} \upsilon(x,u,s)\mathrm{d}s + \int_{t_0+h}^{T} \upsilon(x,u,s)\mathrm{d}s + B[x(T),T]\right\} \\ &= \max\left[\int_{t_0}^{t_0+h} \upsilon(x,u,s)\mathrm{d}s + \max\left\{\int_{t_0+h}^{T} \upsilon(x.u,s)\mathrm{d}s + B[x(T),T]\right\}\right] \\ &= \max\left\{\int_{t_0}^{t_0+h} \upsilon(x,u,s)\mathrm{d}s + V[x(t_0+h),t_0+h]\right\} \\ &= \max\{\upsilon(y,u,t_0)h + V(y,t_0) + V_x(y,t_0)\Delta x + V_t(y,t_0)h + o(h)\} \\ &= \max\{\upsilon h + V(y,t_0) + V_x f h + V_t h + o(h)\}. \end{aligned} \tag{1.7}$$

The first equation is obvious; the second follows from the following principle called the 'principle of optimality': to maximize a total sum of payoffs from x(t0) = y over [t0, T] you must maximize the subtotal of the sum of payoffs from x(t0 + h) over [t0 + h, T]; the third follows from the definition of the state valuation function; the fourth follows from the integral mean value theorem and expansion of V (x (t0 + h), t0 + h) in a Taylor series about x (t0) = y, t0; and the fifth follows from Δx ≡ x (t0 + h) − x(t0) = fh + o(h). Here o(h) is any function of h that satisfies

$$\lim_{h\to 0} o(h)/h = 0.$$

Subtract $V(y, t_0)$ from the LHS and the extreme RHS of the above equation; divide by h and take limits to get (1.3). So (1.3) is nothing but the principle of optimality in differential form. That is all there is to the HJB equation.

Equation (1.4) is just a definition. To motivate this definition rewrite equation (1.7), thus putting $p \equiv V_x$.

$$-V_t = \max\{v(y,u,t_0) + pf(y,u,t_0) + o(h)/h\}. \tag{1.8}$$

The function H^*, called the Hamiltonian function, just collects the terms that contain the control u. The control u must be chosen to maximize H^* along an optimum path. This follows directly from equation (1.7).

The principle that the optimal control u^0 must maximize H^* is important. It is called the *maximum principle*. This principle squares with commmon sense: you should choose the control to maximize the sum of current instantaneous payoff $u(y,u,t_0)$ and future instantaneous value p $\dot{x} = pf(y,u,t_0), p \equiv V_x$. The quantity p, called the *costate* variable, is the marginal value of the state variable. It measures the incremental sum of payoffs from an extra unit of state variable. Equations (1.5) are easy to derive. The relation $\dot{x} = H_p^{*0}$ follows from $\dot{x} = f(x,u^0,t)$ and the envelope theorem. The relation $\dot{p} = -H_x^{*0}$ follows from substitution of the derivative of (1.3) wrt x into the expression for $\mathrm{d}p/\mathrm{d}t = (\mathrm{d}/\mathrm{d}t)V_x$.

Finally (1.6) is obvious. If there is an inequality constraint $x(t) \geqq 0$ for all t, but $B \equiv 0$, then, the transversality condition, $p(T) = B_x(x, T)$ takes the form $p(T)x(T) = 0$. The condition $p(T)x(T) = 0$ means that nothing of value is left over at the terminal date T. When T is infinite, for a large class of problems the condition takes the form

$$\lim_{T\to\infty} p(T)x(T) = 0 \tag{1.9}$$

and is called the *transversality conditions at infinity*. Benveniste and Scheinkman (1982), Araujo and Scheinkman (1983), and Weitzman (1973) show that (1.9) is necessary and sufficient for optimality for a large class of problems.

Let me give a very rough heuristic argument to motivate why (1.9) might be necessary for optimality. For any date T with terminal date in (1.1) set equal to infinity, assume the state valuation function $V(y, T)$ is concave in y. (Note that 't_0' is replaced with 'T' and 'T' is replaced by '∞' in (1.1) here.) Use concavity and $p(T) \equiv V_x(x(T), T)$ to get the bound

$$\begin{aligned} V(x(T),T) - V(x(T)/2,T) &\geqq V_x(x(T),T)x(T)/2 \\ &= p(T)x(T)/2 \end{aligned} \tag{1.10}$$

Now suppose that the distant future is insignificant in the sense that $V(z(T), T) \to 0$, $T \infty \infty$ for any state path **z**. Then it is plausible to expect that the LHS of (1.10) will go to 0 as $T \infty \infty$. If $x(T) \geqq 0$ and $p(T) \geqq 0$ (more x is better than less) then

$$\lim_{T\to\infty} p(T)x(T) = 0$$

which is (1.9).

Examples exist where (1.9) is not necessary for optimality. The idea is that if the distant future is 'significant' then there is no reason to expect the value of 'leftovers' $p(T)x(T)$ to be forced to zero along an optimum path. See Benveniste and Scheinkman (1982), and Araujo and Scheinkman (1983) for the details and references.

In the same manner and for the same reasons as a time series analyst transforms his time series to render it time stationary the dynamic economic modeller searches for a change of units so that (abusing notation to economize on clutter) problem (1.1) may be written in the time stationary form

$$V(y, t_0) = \int_{t_0}^{T} e^{-\delta t} v(x, u) ds + e^{-\delta t} B[x(T)] \quad (1.11)$$

$$\dot{x} = f(x, u), x(t_0) = y. \quad (1.12)$$

By the change of units $W(y, t_0) = e^{\delta t} V(y, t_0)$, $q = e^{\delta t} p$, $H = e^{\delta t} H^*$ and we may write the optimality conditions (1.3)–(1.6) in the form:

$$\delta W - W_t = \max_u H(q, x, u) = H^0(q, x) \quad (1.13)$$

$$H(q, x, u) \equiv v(x, u) + qf \quad (1.14)$$

$$\dot{q} = \delta q - H^0_x, \dot{x} = H^0_q, x(t_0) = y \quad (1.15)$$

$$W(x, T) = B(x), q(T) = B_x(x). \quad (1.16)$$

When the horizon $T = \infty$, W becomes independent of T so that $W_t = 0$; the transversality condition becomes (cf. Benveniste and Scheinkman 1982)

$$\lim_{t\to\infty} e^{-\delta t} q(t) x(t) = 0, \quad (1.17)$$

and (1.17) is *necessary* as well as sufficient, for a solution of (1.15) to be optimal. The condition (1.17) determines q_0.

Equipped with the framework (1.11) and (1.12) together with the optimality conditions (1.13)–(1.17) we are now ready to discuss the economic questions mentioned in the introduction.

Stability

We now have a framework in which to discuss stability of an ideal centrally planned economy. After we do that we will show that the same framework can be used to study related issues in an ideal capitalist economy.

There are five basic notions of stability:

(i) stability of the optimum path with respect to small changes in the horizon and target stocks;
(ii) stability of the optimum path with respect to small changes in υ, f;
(iii) existence of an optimum steady state $(\bar{x}, \bar{u})$ and asymptotic stability of optimum paths wrt $(\bar{x}, \bar{u})$;
(iv) asymptotic stability of $(x(t, u(t))$ wrt $(\bar{x}(t), \bar{u}(t))$ for any two optimum paths $(x(t), u(t))$, $(\bar{x}(t), \bar{u}(t))$;
(v) asymptotic stability of optimal paths $x(t)$ towards a general attractor set Λ.

First, there is an extensive literature (e.g., Mitra 1979, 1983; Majumdar and Zilcha (1987), and their references) that studies the conditions that one must impose upon $\upsilon\ \upsilon, f$ in order that

$$\lim_{T\to\infty} x(t, x_0, T) = x(t, x_0, \infty) \quad (2.1)$$

where $x(t, x_0, T)$, $x(t, x_0, \infty)$ denote solutions to problem (1.1) with T finite and infinite respectively. Here $x(t_0, x_0, T) = x(t_0, x_0, \infty) = x_0$. Sufficient conditions on υ, f needed to obtain the insensitivity result (2.1) are very weak. The result (2.1) is important because it shows that the choice

O

of the terminal time T is unimportant for the initial segment of an optimal plan provided that T is large. We do not have space here to discuss the 'insensitivity' literature any further.

The second notion of stability requires that optimal solutions do not change much when the functions υ, f do not change much. We shall not treat this type of stability in this entry. It is a standard topic in the mathematical theory of optimal control and can be found in many textbooks on the subject. In many economic applications the conditions sufficient for this type of stability are automatically imposed. This kind of stability is a minimal requirement to impose on a problem in order that it be 'well posed'.

The third notion of stability is ubiquitous in economic analysis. The basic notions are easy to explain.

Definitions The pair of vectors $(\overline{q},\overline{x}) \in R^{2n}$ is an *optimal steady state* (OSS) if $(\overline{q},\overline{x})$ solves (1.15) while $\dot{q} = 0$, $\dot{x} = 0$. The optimal steady state $\overline{x}$ is said to be *locally* (globally) *asymptotically stable* if the solution $x(t, y)$ of the optimal dynamic system

$$\dot{x} = H_q^0(q,x) = H_q^0(W_x(x),x) \equiv h(x), \quad x(t_0) = y$$

converges to $\overline{x}$ as $t \to \infty$ for initial conditions y near $\overline{x}$ (for all initial conditions y).

The Case δ Near Zero

We will show in this case that a centrally planned multisector economy is asymptotically stable under modest concavity assumptions. The case $\delta = 0$ is the case where the central planner does not discount the future. F. P. Ramsey's famous paper (1928) on one sector optimal growth introduced the notion of *bliss* in order to deal with the possibly non-convergent integral in (1.7) for the infinite horizon case. That is to say Ramsey put B equal to the maximum obtainable rate of utility or enjoyment and minimized $\int_0^\infty (B - \upsilon)\mathrm{d}t \equiv R(x_0)$ and his famous rule: $B - \upsilon = \dot{x}u'$ follows directly from the HJB equation for R.

The desire to treat utility functions that did not satiate, to treat multiple sectors, and to treat classes of problems where Ramsey's integral $\int_0^\infty (B - \upsilon)\mathrm{d}t$ was not well defined led later investigators (von Weiszäcker 1965; Gale 1967; Brock 1970) to replace

$$B \text{ by } \overline{\upsilon} \equiv \max_{x,u} \upsilon(x,u) \text{ s.t.} f(x,u) \geq 0,$$

and to introduce the overtaking ordering (von Weiszäcker 1965) in various guises. We explain two common versions of overtaking type orderings and their corresponding notions of optimality here. McKenzie's (1976), (1981), syntax is used.

Definitions let $\mathbf{Z} \equiv (\mathbf{x},\mathbf{u}), \mathbf{Z} \equiv (\mathbf{x}',\mathbf{u}')$ be two paths. We say that $\mathbf{Z}$ catches up to $\mathbf{Z}'$ if

$$\overline{\lim_{T\to\infty}} \int_0^T [\upsilon(Z') - \upsilon(Z)]\mathrm{d}t \leq 0. \tag{3.1}$$

Here $\overline{\lim}\, a_T$ denotes the larges cluster point (i.e., the limit superior) of the sequence a_T as $T \to \infty$. Inequality (3.1) states that the accrued utility along $\mathbf{Z}$ eventually exceeds the accured utility along $\mathbf{Z}'$. as $T \to \infty$. This defines a partial ordering of paths $\mathbf{Z}$, $\mathbf{Z}'$. An *optimal* path (Gale 1967) catches up to every other path that starts from the same initial in conditions x_0. We say that $\mathbf{Z}'$ *overtakes* $\mathbf{Z}$ if there is $\varepsilon > 0$ such that

$$\underline{\lim_{T\to\infty}} \int_0^T [\upsilon(Z') - \upsilon(Z)\mathrm{d}t \geq \varepsilon. \tag{3.2}$$

A *weakly maximal* path (Brock 1970) is not overtaken by any other path that starts from the same initial condition x_o. an optimal path beats every other path. A weakly maximal path is not beaten by any other path.

Under the assumption of strict concavity of the payoff and convexity of the constraint set Gale (1967) proved for a discrete time model that a unique optimal path existed and the unique optimal steady state was globally asymptotically stable. For the same model Brock (1970) replaced Gale's strict concavity assumption on the payoff with the weaker

assumptions of concavity of the payoff, uniqueness of the optimal steady state, and convexity of the technology, and, under these weaker assumptions, shortened the proof of Gale's existence theorem, proved existence of weakly maximal programmes, gave an example where the optimal steady state failed to be optimal in the class of all paths starting from it, and proved that time averages of weakly maximal paths converged to the optimal steady state even though the paths themselves may not converge. Continuous time versions of these theorems are in Brock and Haurie (1976). The assumptions needed in the continuous time case basically amount to concavity of $H^0(q, x)$ in x.

Theorems of this type are useful for the stability question because they show the truth of the following proposition.

Proposition If you do not discount the future and you make the usual concavity and convexity assumptions of diminishing marginal rates of substitution and nonincreasing returns on utility and technology then all optimal paths converge to a unique optimal steady state.

This is a strong result. It is independent of the number of sectors. A similar result holds for δ near zero (Scheinkman 1976). These results may be motivated as follows. Linearize (1.15) about the optimal steady state $(\overline{q},\overline{x})$ to obtain, putting

$$\Delta z = \begin{bmatrix} \Delta q \\ \Delta x \end{bmatrix}, \Delta \dot{z} = J\Delta z, \Delta x(0) = x_0 - \overline{x}, \quad (3.3)$$

where J is defined by

$$J = \begin{bmatrix} \delta - H^0_{xq} & -H^0_{xx} \\ H_{qq} & H^0_{qx} \end{bmatrix}. \quad (3.4)$$

It is known (see Levhari and Leviatan 1972, for the discrete time analogue) that if λ is an eigenvalue of J so is $-\lambda + \delta$.

In the case $\delta = 0$ we see that eigenvalues of J came in pairs $-\lambda$, λ so that, except for hairline cases, exactly n of the eigenvalues have negative real parts and exactly half of the eigenvalues have positive real parts. Hence, except for hairline cases, the stable manifold LW_s of (3.3), which is called the local stable manifold of (1.15) (i.e., the set of $(\Delta q(0), \Delta x(0))$ such that the solution of (3.3) starting from $(\Delta q(0), \Delta x(0))$ converges to (0, 0)) is an n-dimensional vector space embedded in R^{2n} whose projection on x-space is n-dimensional. In the 'nondegenerate case the space LW_s is the linear vector space in R^{2n} that is spanned by the n eigenvectors corresponding to the n eigenvalues with negative real parts. To put it another way, except for hairline cases, to each $\Delta x(0)$ there is a unique $\Delta q(0)$ such that $(\Delta x(0), \Delta q(0)) \in LW_s$. Unstable manifolds are defined the same way by reversing the flow of time.

Now the stable manifold W_s of (1.15) at (q, x), which is defined by $W_s \equiv \{(q_0,x_0)|$ the solution of (1.15) starting from (q_0,x_0) converges to $(\overline{q},\overline{x})$ as $t \to \infty\}$ is tangent to LW_s at $(\overline{q},\overline{x})$. The existence and stability theorems for $\delta = 0$ show that the initial costate q_0 must be chosen so that $(q_0,x_0) \in W_s$ for each initial state x_0.

Scheinkman's result (1976) may be interpreted intuitively as continuity of W_s in δ at $\delta = 0$, so global asymptotic stability of an optimal steady state holds provided that δ is near zero. That is to say, in nondegenerate cases, the manifold W_s does not change much when δ does not change much. There is another way to see the role a small δ plays in ensuring stability of a multisector economy.

Differentiate the function

$$V = \dot{q}^T\dot{x} = \dot{x}^T W''\dot{x} \leq 0 \quad (3.5)$$

along solutions of (1.15) that satisfy the transversality condition (1.17) [which by Benveniste–Scheinkman (1982) is *necessary* for optimum] to obtain

$$\dot{V} = \dot{z}^T Q\dot{z} \quad (3.6)$$

where

$$Q = \begin{bmatrix} \delta/2I_n & -H^0_{xx} \\ H^0_{qq} & \delta/2I_n \end{bmatrix} \quad (3.7)$$

Equation (3.6) is easy to derive. Differentiate (1.15) wrt t and substitute the results into $\dot{V} = \ddot{q}^T\dot{x} + \dot{q}^T\ddot{x}$. Let α, β denote the smallest

O

eigenvalue of $-H^0_{xx}, H^0_{qq}$ respectively. Brock and Scheinkman (1976) show that

$$4\alpha\beta > \delta^2 \tag{3.8}$$

implies Q is positive definite so V increases and, hence, global asymptotic stability (G.A.S.) holds. This is so because V is always negative (cf. (3.5)) and is zero only at $\bar{x}$ where $\dot{x} = 0$. It can be shown that (3.8) implies that the optimal steadly state $\bar{x}$ is unique. Hence V increasing in time forces convergence of $x(t)$ to $\bar{x}$ as $t \to \infty$. Since, except for hairline cases, $-H^0_{xx}, H^0_{qq}$ are positive definite for problems with H^0 concave in the state x, therefore G.A.S. holds provided that δ is small enough.

Finally there is yet one more way to see why a small δ forces global asymptotic stability of optimal paths. Put $\delta = 0$ and look at the objective.

$$\text{`max'} \int_0^\infty [v(x,u) - v(\bar{x},\bar{u})]\mathrm{d}t, \ \text{s.t.}\ \dot{x} = f(x,u), \tag{3.9}$$

Here 'max' means weak maximality. Now under strict concavity of v, f in (x, u) and natural monotonicity usually assumed in economic applications $(\bar{x}, \bar{u})$ is the unique solution to the nonlinear programming problem.

$$\max v(x,u) \text{s.t.} f(x,u) \geq 0. \tag{3.10}$$

Hence, intuitively $(x(t), u(t))$ must converge to $(\bar{x}, \bar{u})$ otherwise (3.9) would blow up since the future is not discounted. See Brock and Haurie (1976) for the details. So if δ is close to zero, by continuity of W_s in δ, global asymptotic stability to a unique steady state is preserved. McKenzie (1974) treats the case where $(\bar{x}, \bar{u})$ depends on t.

We have focused on asymptotic *stability* in the foregoing. It is natural to ask what economic forces cause *instability* in a centrally planned economy. Intuitively, instability is present when the underlying dynamics $\dot{x} = f(x, u)$ are unstable when no control u is applied, when control is ineffective ($\partial f/\partial u$ is 'small' in 'absolute value'), when control is expensive, when it is not costly to be out of equilibrium in the state, and when the discount δ, on the future is large. This seems clear. Why spend a lot of resources now in ineffective expensive control to push an economy back into state equilibrium when it currently costs little to be out of equilibrium and benefits arrive in the future which is deeply discounted? A discussion on instability and alternative sufficient conditions for asymptotic stability to those presented here is in Brock (1977). We have no more space to discuss it here. In any event the notions of 'overtaking' and 'bliss' were introduced mainly to resolve issues of existence of optimum paths (Magill 1981) and to investigate asymptotic stability of optimum paths when the future is not discounted.

It is possible for trajectories of centrally planned economies to converge to a limit set Λ that is not a steady state or even a limit cycle. There are more complicated limit sets called 'strange' attractors: they have the property that each pair of nearby trajectories starting in Λ locally diverge at an exponential rate and each trajectory in Λ moves in an apparently 'random' manner. But as we have seen above such 'unstable' phenomena cannot appear when future payoffs are worth almost as much as present payoffs. See Grandmont (1986) for literature on strange attractors in economics as well as literature on empirically testing economic time series for the presence of strange attractors.

Since, as we shall see in section "Some Economic Applications of the Theory" below, each model of a centrally planned economy has a rational expectations market model analogue; therefore the stability literature discussed above applies directly to market models. The strategy of turning optimal growth models into market models and borrowing results from optimal growth theory is at the heart of much of modern macroeconomics and real theories of the business cycle (Kydland and Prescott 1982; Long and Plosser 1983). This kind of application has made the analytical techniques discussed above an essential element of the modern economist's tool-box. We turn now to some of the applications mentioned in the introduction.

Some Economic Applications of the Theory

Are Asset Markets Inherently Unstable?

Rewrite equations (1.15) as

$$\dot{q}_t/q_i + H^0_{x_i}/q_i = \delta, \tag{4.1}$$

$$\dot{x}_i = H^0_{q_i}, i = 1, 2, \ldots, n, x_0 \text{ given}, \tag{4.2}$$

and interpret (4.1) as 'capital gains on asset i plus net yield on asset i = a common rate of return δ', and (4.2) as 'demand for investment in i = supply of investment in i'. The system (4.1), (4.2) has similar mathematical structure to the system of equations describing a market for n assets under myopic perfect foresight analysed by F. Hahn (1966). One may view Hahn's paper as an attempt to formalize the idea held by many people that asset markets are inherently unstable. Indeed Hahn noticed that the linearization of a set of equations much like (4.1), (4.2) around a steady state $(\overline{q},\overline{x})$ displayed a saddle point structure, so that unless q_0 was chosen 'just right' (i.e., on the stable manifold at $(\overline{q},\overline{x})$), then solutions of (4.1), (4.2) starting at (q_0,x_0) would diverge.

The knife-edge problem noticed by Hahn is ubiquitous in models of intertemporal equilibrium in asset markets. See, for example, Gray (1984). Obstfeld and Rogoff (1983, 1986) and references, However, market participants might be expected, knowing the structure of the system (4.1), (4.2), to attempt to forecast the future evolution of earnings of each asset along the solution of the system starting from (q_0, x_0). If capitalized earnings were less than q_0 one would expect traders to bid down q_0, if greater to bid up q_0. Only when q_0 is equal to the present value of anticipated earnings of the asset would one expect no pressure for change of q_0 in the market. Dechert (1978) solves the dynamic integrability problem of when intertemporal equilibrium equations solve some optimal control problem.

The intuitive solution to the knife-edge instability problem given above can be made rigorous for rational expections asset pricing models. See Benveniste and Scheinkman (1982) and references for the deterministic case and Brock (1982, p. 17) for the stochastic case.

To exposit how this line of argument goes, look at the neoclassical one-sector optimal growth model.

$$W(x_0) \equiv \max \int_0^\infty \mathrm{e}^{-\delta t} u(c)\mathrm{d}t, \text{s.t.} c + \dot{x} = f(x) \tag{4.3}$$

where $u' > 0$, $u'(0) = +\infty$, $u'(\infty) = 0$, $u'' < 0$, $f(0) = 0$, $f'(0) = +\infty$, $f' < 0$, $\delta > 0$ are the maintained assumptions on utility u and production frunction f. Make an asset pricing model out of this by introducing a representative consumer who faces a, r, π parametrically and solves.

$$\max \int_0^\infty \mathrm{e}^{-\delta t} u(c)\mathrm{d}t, \text{s.t.} c + a\dot{z} + \dot{x} = rx + \pi x + \pi z, z(0) = 1, \quad x(0) = x_0 \tag{4.4}$$

and a representative firm who leases capital from consumers at rate r to solve.

$$\pi \equiv \max_x [f(x) - rx] \tag{4.5}$$

Here a, r, π, z, c, x denote asset price, interest or rental rate, profits, quantity of asset, consumption, and quantity of capital respectively. There is one perfectly divisible share of the asset available at each point in time. General multisector control planning models may be turned into market models in the same way as the single sector model treated here. For example, such a multisector market model is fitted to data and used to explain business cycles in Long and Plosser (1983).

The collection a, r, π, z, c, x an *equilibrium* if facing a, r, π the solutions of (4.4) and (4.5) agree and $z = 1$ so that all markets clear at all points in time. The necessary conditions of optimality of c, z, x from (4.4) are

$$\delta - \dot{u}'/u' = r = \dot{a}/a + \pi/a \text{ (simple control theory)} \tag{4.6}$$

O

$$\begin{aligned}\lim e^{-\delta t}u'x &= \lim_{t\to\infty} e^{-\delta t}u'az \\ &= 0\,(\text{Benveniste} - \text{Scheinkman}, 1982)\end{aligned} \tag{4.7}$$

Equations (4.7) state that the present value of capital and asset stocks must go to zero as $t \to \infty$. Since (4.5) implies $r = f'$ and $z = 1$ in equilibrium we must have setting $q = u'$, $c(q) \equiv u'^{-1}$,

$$\dot{q} = \delta q = qf', \ \dot{x} = f(x) - c(q), x(0) = x_0. \tag{4.8}$$

The system (4.8) which is the dynamics of the standard neoclassical one sector optimal growth model dramatically displays the knife-edge instability discussed by Hahn (1966) when phase diagrammed. We come to the main substantive point of this section:

Proposition: The necessity of the transversality condition at infinity for the consumer's problem determines the initial value of q_0 and a_0 To put it another way equilibrium c, x are characterized by the solution to (4.3). Furthermore for each t the equilibrium asset price is given by

$$a(t) = \int_0^\infty \exp\left[-\int_t^s (\delta - \dot{u}'/u')\mathrm{d}t\right][f(x) - f'(x)x]\mathrm{d}s$$

evaluated along the solution to (4.3).

A detailed discussion of this kind of result for the case of uncertainty is in Brock (1982).

At an abstract theoretical level this proposition is a resolution of the classical knife-edge instability problem of capital asset markets but how relevant is such a resolution in practice? The assumption of the absence of arbitrage profits and correct expectations over the short period embodied in (4.6) probably captures a central tendency in well developed asset markets like stock exchanges. It is the long term fundamentalist rationality embodied in (4.7) that is more problematic. A more thorough discussion of the economic plausibility of (4.7) is contained in Gray (1984), Obstfeld and Rogoff (1986), and references. Furthermore there is no allowance for short-term or long-term learning and forecasting in the framework.

The study of learning and disequilibrium adjustment mechanisms in capital asset markets is still in its infancy. The literature has not progressed much beyond the work discussed by Blume et al. (1982).

Nevertheless optimal control theoretic intertemporal general equilibrium models much like the one articulated here have had a large impact on the scientific study of asset market bubbles and speculative manias both theoretical (e.g., Gray (1984), Obstfeld and Rogoff (1983), (1986), and empirical (e.g., Flood and Garber 1980; Meese 1986). Indeed, one might say that such methods launched the modern empirical study of bubbles, hyperinflations and speculative manias.

The 'theoretical resolution' of the short-run instability of myopic perfect foresight asset markets has a family resemblance to the problem of decentralization of an infinitely lived economy with the microagents using only a finite number of prices or other signals at each point in time. For example, in the model discussed above, the presence of a stock market forced Pareto optimality of all equilibria. This conclusion is also true in many cases for models where individuals have finite lives (Tirole 1985). Hence, in a sense, decentralizability can be achieved by a finite number of markets at each point in time even though the economy is infinitely lived. To put it another way, in Samuelson–Diamond overlapping generations models where competitive equilibria may be inefficient the mere addition of a stock market eliminates the inefficient equilibria. See Tirole's (1985) discussion of unpublished work by J. Scheinkman for the argument.

Equilibrium Dynamics

We have seen how the notion of transversality condition at infinity contributed to the theoretical and empirical investigation of instability and bubbles in markets for speculative assets. Turn now to a contribution of the asymptotic stability theory of optimal control to the modelling of adjustment dynamics.

Critical articles such as Gordon and Hines (1970) and Lucas (1976) have made many

economists wary of 'ad hoc' dynamic models such as the Walrasian tatonnement ent $\dot{p} = E(p)$ where p is price and E is excess demand, as well as techniques such as Samuelson's Corespondence Principle that rule out 'unstable' equilibria wrt such ad hoc dynamics. We exposit here a framework, using optimal control, that gets around the objection that the dynamics are 'ad hoc' under adjustment costs.

Suppose that a vector x of goods is produced with convex cost function $B(x)$. Suppose that demand is integrable in the sense that there is a social benefit function $B(x)$ such that $Bx = D(x) \equiv p$. Then intertemporal competitive equilibrium is characterized by the solution to the surplus maximization problem.

$$\max \int_0^\infty \mathrm{e}^{-\delta t}[B(x) - C(x,\dot{x})]\mathrm{d}t \equiv W(x_0) \quad (5.1)$$

which yields the necessary conditions

$$\begin{aligned} \dot{q} &= rq - H^0_x, \dot{x} = H^0_q, x(0) = x_0, \\ H^0(q,x) &\equiv \max[B(x) - C(x,\dot{x}) + q\dot{x}]. \end{aligned} \quad (5.2)$$

This is easy to see. For let a representative firm face **p** parametrically and solve

$$\max \int_0^\infty \mathrm{e}^{-rt}[px - C(x,\dot{x})\mathrm{d}t] \quad (5.3)$$

to yield necessary conditions

$$\begin{aligned} \dot{\lambda} &= r\lambda - G^0_x, \dot{x} = G^0_q, \quad x(0) \\ &= x_0, G^0(\lambda, x) \\ &\equiv \max_x [px - C(x,\dot{x}) + \lambda\dot{x}]. \end{aligned} \quad (5.4)$$

Equilibrium requires

$$p = D(x). \quad (5.5)$$

Note that $H^0_x = B_x - C_x = D - C_x \equiv p - C_x = G^0_x$. Identify λ with q and use Benveniste-Scheinkman's (1982) theorem on the necessity of the transversality condition at infinity to finish the proof.

Does $\dot{p}$ in the 'new' framework where the dynamics are endogenous relate naturally to any notion of 'excess demand' as in the traditional but ad hoc Walrasian tatonnement? Differentiate (5.5) along the solution of (5.1) to obtain, denoting the optimal value of $\dot{x}$ by $h(x) \equiv H^0_q(W_x(x), x)$,

$$\dot{p} = D_x\dot{x} = D_x h(x) \equiv K(x) = K\left(D^{-1}(p)\right) \equiv L(p). \quad (5.6)$$

Notice that in the one good case, p moves opposite to x if $D_x<0$. But there is little relationship between the function $L(p)$ and any obvious notion of 'excess demand'. This is as it should be, because the optimal dynamics $h(x)$ embodies future information whereas static excess demand depends only upon current information (or, in distributed lag models, past information).

The optimal control framework laid out here can be used to make four points.

First, although the issue of learning is begged, this framework suggests what actors in the model should be learning *about* in a useful model. That is they should be modelled as learning about the function $h(x)$. See Blume et al. (1982) and their references for literature on learning.

Second, this framework gets around the Gordon–Hines–Lucas objection to 'ad hoc' dynamic modelling like the Walrasian tatonnement. No agent in the model, knowing $h(x)$, can make money on this knowledge. Hence the 'equilibrium' adjustment dynamics $\dot{x} = h(x)$ are 'stable' against profit-seeking behaviour. This shows that it is logically possible to write down models of adjustment dynamics that are immune to the famous 'Lucas Critique' (Lucas 1976).

Third, this framework suggests a reformulation of the Samuelson correspondence principle (Brock 1976) that gets around two fundamental objections to Samuelson's original version: (i) the dynamics were ad hoc and not linked to self-interested purposive behaviour by agents in the model, (ii) the principle had no content because any continuous function can be an excess demand function (the Sonnenschein–Mantel–Debreu Theorem; Debreu 1974). Dynamics (5.6) are equilibrium rational expectations dynamics so objection (i) is met.

O

Objection (ii) is that the original correspondence principle was contentless since excess demand functions are arbitrary. Although when r is small (5.1) imposes many restrictions on $\dot{x} = h(x)$, it can be shown that there are few restrictions on h provided that r is large enough (Grandmont 1986). Nevertheless the structure of (5.1) has been used to formulate versions of the correspondence principle that exhibit restrictions on comparative statics imposed by global asymptotic stability of $\dot{x} = h(x)$. Perhaps the most important thing to realize is that the results of section "The Case δ Near Zero" imply that the adjustment dynamics $\dot{x} = h(x)$ possess a unique steady state which is globally asymptotically stable when the real interest rate, r, is close enough to 0. This is a very strong restriction on the dynamics $\dot{x} = h(x)$ for the empirically relevant case of small real interest rate. See Brock (1976), Magill and Scheinkman (1979), and McKenzie (1981) for results along this line.

Fourth, quadratic versions of (5.1) with the addition of uncertainty generate a large class of empirically useful and econometrically tractable models. See Sargent (1981) for this development.

A Summing Up

In the applications section of this entry we have shown how optimal control methods have contributed to the investigation of basic economic questions such as inherent stability or instability of capitalism, and in centrally planned economies determination of the strength of forces for and against stability, and decentralizability of economies that last foreover. For an example, myopic perfect-foresight asset market equations display a similar saddle point knifeedge instability to that found in the costate–state equations of optimal control (which are necessary for optimum). The corrective force in optimal control theory is the transversality condition at infinity, which motivates search for market forces that are analogous to it; the modern literature on speculative manias emerged from this search.

Bibliography

Araujo, A., and J. Scheinkman. 1983. Maximum principle and transversality condition for concave infinite horizon economic models. *Journal of Economic Theory* 30: 1–16.

Benveniste, L., and J. Scheinkman. 1982. Duality theory of dynamic optimization models of economics: The continuous time case. *Journal of Economic Theory* 27: 1–19.

Blume, L., M. Bray, and D. Easley. 1982. Introduction to the stability of rational expectations equilibrium. *Journal of Economic Theory* 26(2): 313–317.

Brock, W. 1970. On existence of weakly maximal programmes in growth models. *Review of Economic Studies* 37: 275–280.

Brock, W.A. 1976. A revised version of Samuelson's correspondence principle: Applications of recent results on the asymptotic stability of optimal control to the problem of comparing long run equilibria. In *Models of economic dynamics*, Lecture notes in economics and mathematical systems, ed. H. Sonnenschein, Vol. 264. New York: Springer , 1986.

Brock, W. 1977. The global asymptotic stability of optimal control: A survey of recent results. In *Frontiers of quantitative economics*, ed. M.D. Intriligator, Vol. 3A, 297–338. Amsterdam: North-Holland.

Brock, W.A. 1982. Asset prices in a production economy. In *The economics of information and uncertainty*, ed. J.J. McCall. Chicago: University of Chicago Press.

Brock, W.A., and A. Haurie. 1976. On existence of overtaking optimal trajectories over an infinite time horizon. *Mathematics of Operations Research* 1(4): 337–346.

Brock, W., and J. Scheinkman. 1976. The global asymptotic stability of optimal control systems with applications to the theory of economic growth. *Journal of Economic Theory* 12: 164–190.

Burmeister, E. 1980. *Capital theory and dynamics*. New York: Cambridge University Press.

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

Dechert, W. 1978. Optimal control problems from second order difference equations. *Journal of Economic Theory* 19: 50–63.

Flood, R., and P. Garber. 1980. Market fundamentals versus price-level bubbles: The first tests. *Journal of Political Economy* 88: 745–770.

Gale, D. 1967. On optimal development in a multi-sector economy. *Review of Economic Studies* 34: 1–18.

Gordon, D., and A. Hines. 1970. On the theory of price dynamics. In *Microeconomic foundations of inflation and employment theory*, ed. E.S. Phelps, 369–393. New York: W.W. Norton.

Grandmont, J.M. (ed.) 1986. *Nonlinear Dynamics: Journal of Economic Theory Proceedings Volume*.

Gray, J. 1984. Dynamic instability in rational expectations models: An attempt to clarify. *International Economic Review* 25(1): 93–122.

Hahn, F.H. 1966. Equilibrium dynamics with heterogeneous capital goods. *Quarterly Journal of Economics* 80: 633–646.
Kydland, F., and E. Prescott. 1982. Time to build and aggregate fluctuations. *Econometrica* 50: 1345–1370.
Levhari, D., and N. Leviatan. 1972. On stability in the saddle point sense. *Journal of Economic Theory* 4: 88–93.
Long, J., and C. Plosser. 1983. Real business cycles. *Journal of Political Economy* 91: 39–69.
Lucas, R.E. 1976. Econometric policy evaluation: A critique. In *The phillips curve and labor markets*, Carnegie-Rochester conference series on public policy, ed. K. Brunner and A. Meltzer, Vol. 1. Amsterdam: North-Holland.
Magill, M.J.P. 1981. Infinite horizon programs. *Econometrica* 49(3): 679–711.
Magill, M., and J. Scheinkman. 1979. Stability of regular equilibira and the correspondence principle for symmetric variational problems. *International Economic Review* 20(2): 297–315.
Majumdar, M. and Zilcha, I. 1987. Optimal growth in a stochastic environment. *Journal of Economic Theory.*
McKenzie, L. 1974. Turnpike theorems with technology and welfare function variable. In *Mathematical models in economics*, ed. J. Los and M.W. Los. New York: American Elsevier.
McKenzie, L. 1976. Turnpike theory. *Econometrica* 44(5): 841–866.
McKenzie, L. 1981. Optimal growth and turnpike theorems. In *Handbook of mathematical economics*, ed. K. Arrow and M. Intriligator. New York: North-Holland.
Meese, R. 1986. Testing for bubbles in exchange markets: A case of sparkling rates? *Journal of Political Economy* 94: 353–373.
Mitra, T. 1979. On optimal economic growth with variable discount rates: Existence and stability results. *International Economic Review* 20: 133–145.
Mitra, T. 1983. Sensitivity of optimal programs with respect to changes in target stocks: The case of irreversible investment. *Journal of Economic Theory* 29(1): 172–184.
Obstfeld, M., and K. Rogoff. 1983. Speculative hyperinflations in maximizing models: Can we rule them out? *Journal of Political Economy* 91: 675–705.
Obstfeld, M., and K. Rogoff. 1986. Ruling out divergent speculative bubbles. *Journal of Monetary Economics* 17(May): 349–362.
Ramsey, F.P. 1928. A mathematical theory of saving. *Economic Journal* 38(December): 543–559.
Sargent, T. 1981. Interpreting economic time-series. *Journal of Political Economy* 89(2): 213–248.
Scheinkman, J. 1976. On optimal steady states of n-sector growth models when utility is discounted. *Journal of Economic Theory* 12: 11–30.
Tirole, J. 1985. Asset bubbles and overlapping generations. *Econometrica* 53: 1071–1100.
von Weizsäcker, C.C. 1965. Accumulation for an infinite time horizon. *Review of Economic Studies* 32: 85–104.
Weitzman, M. 1973. Duality theory of convex programming for infinite horizon economic models. *Management Science* 19: 783–789.

Optimal Fiscal and Monetary Policy (with Commitment)

Mikhail Golosov and Aleh Tsyvinski

Abstract

'Optimal fiscal and monetary policy with commitment' is a policy of choosing taxes and transfers or monetary instruments to maximize social welfare. 'Commitment' refers to ability of a policymaker to make binding policy choices.

Keywords

Bonds; Commitment; Disability insurance; Friedman rule; Homotheticity; Incentive compatibility; Inflation targeting; Inflation tax; Intermediate good; Linear taxation; Lump sum taxes; Mirrlees, J.; Money; Nominal interest rates; Optimal fiscal and monetary policy with commitment; Optimal interest rate; Optimal monetary policy; Optimal taxation; Output gap; Partial equilibrium; Ramsey taxation; Separability; Sticky prices; Tax smoothing; Taxation of capital income; Taxation of consumption; Taxation of income; Taylor rule; Time preference; Uniform commodity taxation

JEL Classifications

D4; D10

The Ramsey Approach to the Optimal Taxation

'Ramsey approach to optimal taxation' is the solution to the problem of choosing optimal taxes and transfers given that only distortionary tax instruments are available.

A starting point of a Ramsey problem is postulating tax instruments. Usually, it is assumed that only linear taxes are allowed. Importantly, lump sum taxation is prohibited. Another assumption crucial to this approach is that all activities of agents are observable.

Given the set taxes, a social planner (government) maximizes its objective function given that agents (firms and consumers) are in a competitive equilibrium. Usually, it is assumed that government's objective is to finance an exogenously given level of expenditures. It is important to note that if the lump sum taxes were allowed than the first welfare theorem would hold, and the unconstrained optimum would be achieved.

There are two common approaches to solving Ramsey problems. The first is the *primal* approach, which characterizes a set of allocations that can be implemented as a competitive equilibrium with taxes. By 'implementation we mean' the following: for a set of taxes find a set of (consumption and labour) allocations and equilibrium prices such that these allocations are a competitive equilibrium given taxes. Conversely, a set of (consumption and labour) allocations is implementable if it is possible to find taxes and equilibrium prices such that these allocations are a competitive equilibrium given these prices and taxes. Implementation often makes it possible to simplify a Ramsey problem by reformulating a problem of finding optimal taxes as the problem of finding implementable allocations. This reformulation is referred to as the *primal approach* to Ramsey taxation.

Main Lessons of Ramsey Taxation: Uniform Commodity Taxation, Zero Capital Tax in the Long Run, and Tax Smoothing

One of the central results of the literature on Ramsey taxation is *uniform commodity taxation* (Atkinson and Stiglitz 1972). Consider a model with a finite set of consumption goods that can be allocated between government and private consumption. All of these goods are produced with labour. Assume that each consumption good can be taxed at a linear rate. Then, under certain separability and homotheticity assumptions, commodity taxation is uniform, that is, the optimal taxes are equated across consumption goods.

Ramsey taxation provides a compelling argument against taxing capital income in the long run in a model of infinitely lived households. The *Chamley–Judd result* (Chamley 1986; Judd 1985) states that in a steady state there should be no wedge between the intertemporal rate of substitution and the marginal rate of transformation, or, alternatively, that the optimal tax on capital is zero. The intuition for the result is that even a small intertemporal distortion implies increasing taxation of goods in future periods in contrast to the prescription of the uniform commodity taxation. Therefore, distorting the intertemporal margin is very costly for the planner. Jones et al. (1997) extend the applicability of the Chamley–Judd result by showing that the return to human capital should not be taxed in the long run. Chari et al. (1994) provide the state-of-the art numerical treatment for optimal Ramsey taxation over the business cycle and conclude that the *ex ante* capital tax rate is approximately zero.

There has been a long debate on the optimal composition of taxation and borrowing to finance government expenditures. Barro (1979) considers a partial equilibrium economy and argues that it is optimal to smooth distortions from taxation over time, a policy referred as *tax smoothing*. The implication of this analysis is that optimal taxes should follow a random walk. Lucas and Stokey (1983) consider an optimal policy in a general equilibrium economy without capital, and show that, if government has access to state-contingent bonds, optimal taxes inherit the stochastic process of the shocks to government purchases. Chari et al. (1994) extend this analysis to an economy with capital and show the Lucas and Stokey results remain valid in that set-up with or without state contingent debt, as long as the government can use taxes on capital to effectively vary the *ex post* after-tax rate of return on bonds. Finally, Aiyagari et al. (2002) show that, if *ex post* taxation of returns is impossible, the optimal taxes follow a process similar to a random walk. They also show the conditions under which the tax smoothing hypothesis is valid.

The Mirrlees Approach to Optimal Taxation

The Mirrlees approach to optimal taxation is built on a different foundation from Ramsey taxation. Rather than stating an ad hoc restricted set of tax instruments as in Ramsey taxation, Mirrlees (1971) assumed that an informational friction endogenously restricted the set of taxes that implement the optimal allocation. This set-up allows arbitrary nonlinear taxes, including lump-sum taxes.

The informational friction posed in those models is unobservability of agents' skills: only labour income of agents can be observed. Therefore, from a given level of labour income it cannot be determined whether a high-skill agent provides a low amount of labour or effort, or whether a low-skill agent works a prescribed amount. The objective of the social planner (government) is to maximize *ex ante,* before the realization of the shocks, utility of an agent. This objective can be interpreted as either insurance against adverse shocks or as *ex post* redistribution across agents of various skills. An informational friction imposes *incentive compatibility* constraints on the planner's problem: allocations of consumption and effective labour must be selected such that an agent chooses not to misrepresent its type.

In summary, the objective of the Mirrlees approach is to find the optimal incentive–insurance trade-off: how to provide the best insurance against adverse events (low realizations of skills) while providing incentives for the agents to reveal their types (provide high amount of labour).

Main Lessons of the Mirrlees Approach in a Static Framework

Theoretical results providing general characterization of the optimal taxes in the static Mirrlees environment are limited. The central result is that the consumption–leisure margin of an agent with the highest skill is undistorted, implying that the marginal income tax at the top of the distribution should be optimally set equal to zero. Saez (2001) is a state-of-the art treatment of the static Mirrlees model in which he derives a link between the optimal tax formulas and elasticities of income. Mirrlees (1971) was also able to establish broad conditions that would ensure that the optimal marginal tax rate on labour income was between zero and 100 per cent.

Main Lessons of Dynamic Mirrlees Literature: Distorted Intertemporal Margin

Recent literature starting with Golosov et al. (2003) and Werning (2001) extends the static Mirrlees (1971) framework to dynamic settings. Golosov et al. (2003) consider an environment with general dynamic stochastically evolving skills. An example of a large unobservable skill shock is disability that is often difficult to observe (classical example is back pain or mental illness). Golosov et al. (2003) show for arbitrary evolution of skills that, as long as the probability of agent's skill changing is positive, any optimal allocation includes a positive intertemporal wedge: a marginal rate of substitution across periods is lower than marginal rate of transformation. The reason for this is that this wedge improves the intertemporal provision of incentives by implicitly discouraging savings. This result holds even away from the steady state and sharply contrasts with the Chamley–Judd result that stems from the exogenous restriction on tax instruments. Golosov et al. (2003) and Werning (2001) show that in a case of constant types a version of uniform commodity taxation holds and the intertemporal margin is not distorted.

Implementation of dynamic Mirrlees models is more complicated than implementation of either static Mirrlees models, which are implemented with an income tax, or Ramsey models of linear taxation. By 'implementation' we mean finding tax instruments such that the optimal allocation is a competitive equilibrium with taxes. One possible implementation is a direct mechanism that mandates consumption and labour menus for each date. However, such a mechanism can

include taxes and transfers never used in practice. Three types of implementations have been proposed. In Albanesi and Sleet (2006), wealth summarizes agents' past histories of shocks that are assumed to be i.i.d. and allows us to define a recursive tax system that depends only on current wealth and effective labour. Golosov and Tsyvinski (2006) implement an optimal disability insurance system with asset-tested transfers that are paid to agents with wealth below a certain limit. Kocherlakota (2005) allows for a general process for skill shocks and derives an implementation with linear taxes on wealth and arbitrarily nonlinear taxes on the history of effective labour.

Optimal Monetary Policy

The theory of the optimal monetary policy is closely related to the theory of optimal taxation. Phelps (1973) argues that the inflation tax is similar to any other tax, and therefore should be used to finance government expenditures. Although intuitively appealing, this argument is misleading. Chari et al. (1996) extend the Ramsey approach to analyse optimal fiscal and monetary policy jointly in several monetary models, and find that typically it is optimal to set the nominal interest rate to be equal to zero. Such a policy is called a 'Friedman rule', after Milton Friedman, who was one of the first proponents of zero nominal interest rates (Friedman 1969). To understand intuition for the optimality of Friedman rule, it is useful to think about the distinctive the features that distinguish money from other goods and assets. In most models money plays a special role of providing liquidity services to households that cannot be obtained by using other assets such as bonds. Inefficiency arises if the rates of return on bonds and money are different, since by holding money balances households lose the interest rate. When a nominal interest rate is equal to zero, which in a deterministic economy implies that inflation is negative, with nominal prices declining with the rate of households' time preferences, the real rates of return on money and bonds are equalized and this inefficiency is eliminated.

The optimality of the Friedman rule stands in a direct contrast with Phelps' arguments for use of the inflationary tax together with other distortionary taxes such as taxes on consumption or labour income. The reason for this is that money, unlike consumption or leisure, is not valued by households directly but only indirectly, as long as it facilitates transactions and provides liquidity. Therefore, it is more appropriate to think of money as an intermediate good in acquiring final goods consumed by households. Diamond and Mirrlees (1971) established very general results about the undesirability of distortion of the intermediate goods sector, which in monetary models implies that the inflationary tax should not be used despite the distortions caused by taxes on the final goods and services.

The intuition developed above is valid under the assumption that nominal prices are fully flexible, and firms adjust to them immediately in response to changes in market conditions. However, even casual observation suggests that many prices remain unchanged over long periods of time, and Bils and Klenow (2004) document inflexibility of prices for a wide variety of goods. Inflexible or *sticky prices* lead to additional inefficiencies in the economy that could be mitigated by monetary policy. For example, an economy-wide shock, such as an aggregate productivity shock or change in government spending, may call for readjustment of real prices. If adjustment of nominal prices is sluggish, the central bank can increase welfare by adjusting nominal interest rates and affecting real prices.

It is important to recognize that the government is also able to affect real (aftertax) prices using fiscal instruments instead. In fact, Correia et al. (2002) show that, if fiscal policy is sufficiently flexible and can respond to aggregate shocks quickly, then the Friedman rule continues to be optimal even with sticky prices, with fiscal instruments being preferred to monetary ones. In current practice, however, it appears that it takes a long time to enact changes in tax rates, while monetary policy can be adjusted quickly. Schmitt-Grohe and Uribe (2004) show that, as long as tax levels are fixed or the government is not able to levy some of the taxes on goods or firms' profits, then the optimal interest rate is positive and variable.

Most of the applied literature on the monetary policy is based on the joint assumption of sticky prices and inflexible fiscal policy. Woodford (2003) provides a comprehensive study of the optimal policy in such settings. This analysis examines how central bank response should depend on the type of the shock affecting the economy, the degree of additional imperfections in the economy, and the choice of policies that would rule out indeterminacy of equilibria. Two common policy recommendations for central banks share many of the features of the optimal policy responses in this analysis. One of such recommendations – a *Taylor rule* (see Taylor 1993) – calls for the interest rates to be increased in response to an increase in the output gap (the difference between actual and a target level of GDP) or inflation. Another recommendation, *inflation forecast targeting*, requires that the central bank commits to adjust interest rate to ensure that the projected future path of inflation or other target variables does not deviate from the pre-specified targets.

In addition to the analysis set out above, several new, conceptually different approaches to the analysis of monetary policy have emerged in the recent years. For example, da Costa and Werning (2005) re-examine optimal monetary policy with flexible prices in Mirrleesian settings and confirm the optimality of the Friedman rule there. Seminal work by Kiyotaki and Wright (1989) has given rise to a large search-theoretic literature seeking to understand the fundamental reasons that money differs from other goods and assets in the economy. Lagos and Wright (2005) provide a framework for the analysis of optimal monetary policy in such settings.

See Also

- Income Taxation and Optimal Policies
- Monetary and Fiscal Policy Overview
- New Keynesian Macroeconomics
- Optimal Fiscal and Monetary Policy (Without Commitment)
- Optimal Taxation
- Taylor Rules

Bibliography

Aiyagari, S., A. Marcet, T. Sargent, and J. Seppala. 2002. Optimal taxation without state-contingent debt. *Journal of Political Economy* 110: 1220–1254.

Albanesi, S., and C. Sleet. 2006. Dynamic optimal taxation with private information. *Review of Economic Studies* 73: 1–30.

Atkinson, A., and J. Stiglitz. 1972. The structure of indirect taxation and economic efficiency. *Journal of Public Economics* 1: 97–119.

Barro, R. 1979. On the determination of the public debt. *Journal of Political Economy* 87: 940–971.

Bils, M., and P. Klenow. 2004. Some evidence on the importance of sticky prices. *Journal of Political Economy* 112: 947–985.

Chamley, C. 1986. Optimal taxation of capital income in general equilibrium with infinite lives. *Econometrica* 54: 607–622.

Chari, V., L. Christiano, and P. Kehoe. 1994. Optimal fiscal policy in a business cycle model. *Journal of Political Economy* 102: 617–652.

Chari, V., L. Christiano, and P. Kehoe. 1996. Optimality of the Friedman rule in economies with distorting taxes. *Journal of Monetary Economics* 37: 203–223.

Correia, I., J.-P. Nicolini, and P. Teles. 2002. Optimal fiscal and monetary policy: Equivalence results. Working Paper No. WP-02-16, Federal Reserve Bank of Chicago.

da Costa, C., and I. Werning. 2005. On the optimality of the Friedman rule with heterogeneous agents and non-linear income taxation. Working Paper, MIT.

Diamond, P., and J. Mirrlees. 1971. Optimal taxation and public production I: Production efficiency. *American Economic Review* 61: 8–27.

Friedman, M. 1969. The optimum quantity of money. In *The optimum quantity of money and other essays*. Chicago: Aldine.

Golosov, M., and A. Tsyvinski. 2006. Designing optimal disability insurance: A case for asset testing. *Journal of Political Economy* 114: 257–279.

Golosov, M., N. Kocherlakota, and A. Tsyvinski. 2003. Optimal indirect and capital taxation. *Review of Economic Studies* 70: 569–587.

Jones, L., R. Manuelli, and P. Rossi. 1997. On the optimal taxation of capital income. *Journal of Economic Theory* 73: 93–117.

Judd, Kenneth L. 1985. Redistributive taxation in a simple perfect foresight model. *Journal of Public Economics* 28: 59–83.

Kiyotaki, N., and R. Wright. 1989. On money as a medium of exchange. *Journal of Political Economy* 97: 927–954.

Kocherlakota, N. 2005. Zero expected wealth taxes: A Mirrlees approach to dynamic optimal taxation. *Econometrica* 73: 1587–1622.

Lagos, R., and R. Wright. 2005. A unified framework for monetary theory and policy analysis. *Journal of Political Economy* 113: 463–484.

Lucas, R., and N. Stokey. 1983. Optimal fiscal and monetary policy in an economy without capital. *Journal of Monetary Economics* 12: 55–93.

Mirrlees, J. 1971. An exploration in the theory of optimum income taxation. *Review of Economic Studies* 38: 175–208.
Phelps, E. 1973. Inflation in the theory of public finance. *Swedish Journal of Economics* 75: 67–82.
Saez, E. 2001. Using elasticities to derive optimal income tax rates. *Review of Economic Studies* 68: 205–229.
Schmitt-Grohe, S., and M. Uribe. 2004. Optimal fiscal and monetary policy under sticky prices. *Journal of Economic Theory* 114: 183–209.
Taylor, J. 1993. Discretion versus policy rules in practice. *Carnegie-Rochester Conference Series on Public Policy* 39: 195–214.
Werning, I. 2001. Optimal unemployment insurance with hidden savings. Mimeo, University of Chicago.
Woodford, M. 2003. *Interest and prices*. Princeton: Princeton University Press.

Optimal Fiscal and Monetary Policy (Without Commitment)

Mikhail Golosov and Aleh Tsyvinski

Abstract

'Optimal fiscal and monetary policy' is a policy of choosing taxes and transfers or monetary instruments to maximize social welfare. 'Absence of commitment' refers to inability of a policymaker to make binding policy choices.

Keywords

Bonds; Central bank; Commitment; Dimensionality; Friedman rule; Infinite-horizon models; Inflation; Inflationary bias; Inflationary cap; Markov-perfect equilibria; Nominal interest rates; Optimal fiscal policy without commitment; Optimal monetary policy without commitment; Output gap; Ramsey taxation; Rational expectations; Sustainable equilibrium; Taxation of capital; Taylor rule; Time consistency of monetary and fiscal policy

JEL Classifications

D4; D10

Most of the results of optimal taxation literature in the Ramsey framework are derived under the assumption of commitment. Commitment is usually defined as ability of a government to bind future policy choices. This assumption is restrictive. A government, even a benevolent one, may choose to change its policies from those promised at an earlier date. The first formalization of the notion of time inconsistency is due to Kydland and Prescott (1977), who showed how timing of government policy may change economic outcomes. Furthermore, equilibrium without commitment can lead to lower welfare for society than when a government can bind its future choices.

An example that clarifies the notion of time inconsistency in fiscal policy is taxation of capital. A classical result due to Chamley (1986) and Judd (1985) states that capital should be taxed at zero in the long run. One of the main assumptions underlying this result is that a government can commit to a sequence of capital taxes. However, a benevolent government will choose to deviate from the prescribed sequence of taxes. The reason is that, once capital is accumulated, it is sunk, and taxing capital is no longer distortionary. A benevolent government would choose high capital taxes once capital is accumulated.

The reasoning above leads to the necessity of the analysis of time inconsistent policy as a game between a policymaker (government) and a continuum of economic agents (consumers). A formalization of such a game and an equilibrium concept is due to Chari and Kehoe (1990). They formulate a general equilibrium infinite-horizon model in which private agents are competitive, and the government maximizes the welfare of the agents. They define an equilibrium concept – sustainable equilibrium – which is a sequence of history-contingent policies that satisfy certain optimality criteria for the government and private agents.

Recent developments in solving for the set of sustainable government policies use the techniques of the analysis of repeated games due to Abreu (1986) and Abreu et al. 1990). Phelan and Stachetti (2001) extend these methods to analyse the equilibria of the Ramsey model of capital taxation. Their contribution is to provide a method

in which the behaviour of consumers is summarized as a solution to the competitive equilibrium, thus significantly reducing the dimensionality of the problem. They provide a characterization of the whole set of sustainable equilibria of the game. Their methods are especially relevant for the environments in which the punishment to the deviator is difficult to characterize analytically.

Benhabib and Rusticchini (1997) and Marcet and Marimon (1994) provide an alternative method to solve policy games without commitment. They use the techniques of optimal control in which they explicitly impose additional constraints on the standard optimal tax problem such that a government does not deviate from the prescribed sequence of taxes. Their methods, while easier to use than those of Abreu (1986), Abreu et al. (1990) and Phelan and Stachetti (2001), are efficient only if the worst punishment to the deviating government can be easily determined.

Klein et al. (2004) numerically solve for equilibria where reputational mechanisms are not operative and characterize Markov-perfect equilibria of the dynamic game between successive governments in the context of optimal Ramsey taxation. For a calibrated economy, they find that the government still refrains from taxing at confiscatory rates.

Optimal Monetary Policy Without Commitment

The problem of time consistency also arises in monetary economics. Kydland and Prescott (1977) and Barro and Gordon (1983) analyse a reduced form economy with a trade-off between inflation and unemployment. Consider an economy where the growth rate of nominal wages is being set one period in advance. The government can decrease unemployment by having setting the inflation rate higher than the wage rate, thus reducing the real wage; but inflation is socially costly. Suppose that a monetary authority chooses the inflation rate after nominal wages were set in the economy to maximize social welfare. Such a rate would equalize the marginal benefits of reducing unemployment and the marginal costs of increasing inflation. But now consider wage determination in a rational-expectations equilibrium. In anticipation of the government's policy, agents will choose a positive growth rate of wages to avoid losses from inflation. Therefore, in equilibrium the monetary authority is not able to affect unemployment, but there is a positive rate of inflation. This outcome is inefficient since by committing not to inflate ex ante the monetary authority could achieve the same level of unemployment but with zero inflation. Therefore, the lack of commitment by the monetary authority will lead to inflationary bias, or an inefficiently high level of inflation.

Similar effects are present in many other monetary models. For example, Calvo (1978) shows time inconsistency of the optimal policy in a general equilibrium model. Chang (1998) considers a version of Calvo's model to find the optimal monetary policy without commitment. Similar to Phelan and Stacchetti (2001), he uses tools of repeated game theory to describe the best equilibrium in the game between the central bank and a large group of agents.

A substantial amount of work has been done in finding the ways to overcome time consistency problems. One of the first practical proposals is Rogoff's (1985) suggestion to appoint a 'conservative' central banker, whose private valuation of the costs of inflation is higher than the social valuation. Such a banker has less temptation to inflate, and the inflationary bias will be reduced.

Pre-specifying the rules of conduct for monetary policy reduces the discretionary actions a central bank can undertake and improves time consistency. For example, the commonly advocated Taylor rule prescribes that the central bank sets nominal interest rates as a linear function of inflation and the output gap with fixed coefficients (see, for example, Woodford 2003). On the other hand, it may be desirable to leave some discretion to the central bank, particularly if it has access to information about economic conditions which is impossible or impractical to incorporate into predetermined rules. Athey et al. (2005) consider an example of such an economy where the central bank has private information about the state of the economy, which is unavailable to others. They show that the optimal policy in such settings is

an inflationary cap that allows discretion to the central bank as long as the inflation rate is below a certain bound.

Following Lucas and Stokey's (1983) analysis, substantial work has been done in determining conditions under which the government can eliminate the time consistency problem by optimally choosing debt of various maturities. Lucas and Stokey themselves point out the fundamental difficulty with this approach in monetary economies since, as long as the government holds a positive amount of nominal debt, it is tempted to inflate in order to reduce its real value. Two recent papers describe some of the conditions under which this problem can be overcome. Alvarez et al. (2004) consider several monetary models and show that if it is optimal to set nominal interest rates at zero (that is, the optimal monetary policy with commitment is to follow the Friedman rule), then the time consistency problem can be solved. By issuing a mixture of nominal and real (indexed) bonds in such a way that the present value of the nominal claims is zero, the temptation for inflation can be removed. Persson et al. (2006) consider a model where the Friedman rule is not optimal, but they still are able to characterize the optimal maturity structure of nominal and indexed bonds that achieve the social optimum with commitment even with time-inconsistent government.

See Also

- ▶ Monetary and Fiscal Policy Overview
- ▶ Optimal Fiscal and Monetary Policy (with Commitment)
- ▶ Optimal Taxation
- ▶ Repeated Games

Bibliography

Abreu, D. 1986. Extremal equilibria of oligopolistic supergames. *Journal of Economic Theory* 39: 191–225.

Abreu, D., D. Pearce, and E. Stacchetti. 1990. Toward a theory of discounted repeated games with imperfect monitoring. *Econometrica* 58: 1041–1063.

Alvarez, F., P. Kehoe, and P. Neumeyer. 2004. The time consistence of optimal monetary and fiscal policies. *Econometrica* 72: 541–567.

Athey, S., A. Atkeson, and P. Kehoe. 2005. The optimal degree of monetary policy discretion. *Econometrica* 73: 1431–1476.

Barro, R., and D. Gordon. 1983. A positive theory of monetary policy in a natural rate model. *Journal of Political Economy* 91: 589–610.

Benhabib, J., and A. Rustichini. 1997. Optimal taxes without commitment. *Journal of Economic Theory* 77: 231–259.

Calvo, G. 1978. On the time consistence of optimal policy in a monetary economy. *Econometrica* 46: 1411–1428.

Chamley, C. 1986. Optimal taxation of capital income in general equilibrium with infinite lives. *Econometrica* 54: 607–622.

Chang, R. 1998. Credible monetary policy in an infinite horizon model: Recursive approach. *Journal of Economic Theory* 81: 431–461.

Chari, V., and P. Kehoe. 1990. Sustainable plans. *Journal of Political Economy* 98: 783–802.

Judd, K. 1985. Redistributive taxation in a simple perfect foresight model. *Journal of Public Economics* 28: 59–83.

Klein, P., Krusell, P., and Ríos-Rull, J.-V. 2004. Time consistent public expenditures, Discussion paper no. 4582. London: CEPR.

Kydland, F., and E. Prescott. 1977. Rules rather than discretion: The inconsistency of optimal plans. *Journal of Political Economy* 85: 473–492.

Lucas, R., and N. Stokey. 1983. Optimal fiscal and monetary policy in an economy without capital. *Journal of Monetary Economics* 12: 55–93.

Marcet, A., and Marimon, R. 1994. Recursive contracts, Working paper no. 337. Department of Economics and Business, Universitat Pompeu Fabra.

Persson, M., T. Persson, and L. Svensson. 2006. Time consistency of fiscal and monetary policy: A solution. *Econometrica* 74: 193–212.

Phelan, C., and E. Stacchetti. 2001. Sequential equilibria in a Ramsey tax model. *Econometrica* 69: 1491–1518.

Rogoff, K. 1985. The optimal degree of commitment to an intermediate monetary target. *Quarterly Journal of Economics* 100: 1169–1190.

Woodford, M. 2003. *Interest and prices*. Princeton: Princeton University Press.

Optimal Savings

Sukhamoy Chakravarty

How much should a nation save or, to put it differently, what is the optimal rate of growth? This question is at the heart of the extensive literature on 'optimum savings' which developed as a

complement to the literature on descriptive growth models in the 1950s and 1960s. Let it be noted that the reasonableness of the question presupposes a utilitarian welfare-theoretic outlook, which locates a source of 'market failure' in the intertemporal context stemming from what A.C. Pigou (1928) had described as a defective telescopic faculty. While Böhm-Bawerk, Fisher and other economists had noted the fact the individuals show a preference for advancing the timing of future satisfaction, they refrained from making any normative statement. Instead they constructed theories of interest which utilized this crucial behavioural characteristic on the part of individual economic agents. Pigou, however, read into the fact that individuals discount future satisfaction at a positive rate, that is display impatience, 'a far reaching economic disharmony' (1928, p. 26). This made him seriously question the 'optimality' of the rate of savings thrown up by an otherwise fully competitive market even under conditions of full employment. Pigou's ideas on this question received support from the Cambridge philosopher-mathematician Frank P. Ramsey, who took the next most important step of determining a rule for determining the optimum rate of savings based on the logic of intertemporal utility maximization, one of the early exercises in economics using the technique of classical calculus of variations. Ramsey was relatively precise in laying down the normative postulates underlying his enquiry, ingenious in deriving the characteristics of the optimal path and not so much concerned with demonstrating that an optimal solution will always exist even on the premises laid out by him.

Ramsey's paper was much appreciated by John Maynard Keynes who, in his obituary note on Ramsey's untimely death, which appeared in the *Economic Journal*, called it 'one of the most remarkable contributions to mathematical economics ever made' (1930).

Despite Keynes, Ramsey's paper received very little attention for nearly three decades, partly, because of the 'Great Depression' where 'excessive savings' in the sense of too high a propensity to save appeared to many economists including Keynes himself, to be the problem, and the emergence of a new welfare economics, which found the cardinal approach towards utility embedded in Ramsey's formulation of the problem extremely questionable, if not unacceptable.

During the late 1950s, however, attention was redirected to the question which Ramsey had posed, especially by those who were particularly concerned with problems of development planning in relation to low income countries. Experience of sustained full employment in the advanced capitalist countries, obvious inadequacy of the stock of capital in the poorer countries from the point of view of generating employment at an adequate level of remuneration, and a back-door entry of 'cardinal utility' via the von Neumann–Morgenstern axioms, although applicable only to risky prospects, made the intellectual environment more receptive to the class of issues that Ramsey had dealt with in his 1928 paper.

Discussion was initiated by Tinbergen (1960), Goodwin (1961) and Chakravarty (1962a) from a development theoretic point of view the motivation was to help planners arrive at optimal growth paths for labour surplus economies based on explicit parametric forms of utility and production functions.

These exercises showed that even in relatively simple cases, optimal paths do not always exist for an open-ended future. The special nature of the assumptions made by Ramsey became more evident through extensive investigations initiated by Koopmans (1960) on the axiomatization of intertemporal utility functions which were in some suitable sense continuous. While Koopmans was concerned with complete and continuous preference orderings, a different approach was taken by Von Weizsäcker (1965) and dealt with a partial order on the programme space defined by the principle of 'overtaking'. A consumption path C_t is said to overtake an alternative path C_t if there exists a time T^* such that $\int_0^T u(c_t)\mathrm{d}t > \int_0^T U\left(c_t^*\right)\mathrm{d}t$ for all $T \geqslant T^*$. The overtaking criterion, being a partial order, allows for non-comparable paths but as subsequent discussion showed this may not matter under certain economically relevant conditions, thereby

providing an extension of the Ramsey criterion which deals with improper integrals of the form $\int_0^\infty u(c(t))\,\mathrm{d}t$ with concave $U(c(t))$ functions.

While Ramsey dealt with a stationary population, during the 1960s characterization of 'optimal growth paths' in the 'overtaking sense' was extended to situations involving exogenously growing population. A reasonably complete analysis was given by Cass (1965) and Koopmans (1965) for the one good case with continuous time and twice continuously differentiable production and utility functions.

A multisectoral generalization of the original Ramsey model was carried out by Samuelson and Solow (1956) in the mid-1950s. During the 1960s, Gale (1967) and others derived multisectoral generalizations for situations involving growing populations, using once again the 'overtaking' criterion.

Aggregative models involving exogenous technical change were carried out by Mirrlees (1967), Inagaki (1970) and several others. These authors used explicit 'time discounting' and obtained for certain special case lower bounds which a constant rate of time discounting must obey.

Most recently, Magill (1981) has provided a very thorough analysis of the existence question for optimal infinite horizon programmes involving complete orderings, and a variety of technologies. Welfare maximization over time involving exhaustible resources was first studied by Hotelling (1931), more or less contemporaneously with Ramsey. This literature has proliferated in recent years and has been exhaustively dealt with by Dasgupta and Heal (1979). A recent paper by de Grandville (1980) combines capital accumulation along with depletion of stocks and derives optimal growth paths, following the Samuelson–Solow paper.

A. Ramsey characterized the social welfare function or more accurately, the welfare function over time as the integral of deviations of current utility levels from a postulated finite upper bound on instantaneous utility levels, denoting it by 'Bliss' 'B', assumed zero time discounting concave utility functions, a stationary population and no technical progress. He distinguished between two types of 'bliss', one due to capital saturation and the other due to utility saturation. In compact mathematical notation. Ramsey's problem was to minimize an expression $\int_0^\infty (B - U(c(t))\,\mathrm{d}t$ subject to $c(t) + k'(t) = f(l,k)$ where $c(t)$ stands for consumption at time t, $k(t)$ denoted the stock of capital at 't' and 'l' for a given labour force. $k'(t \equiv dk/dt)$ represents the rate of capital formation, measured on a 'net' basis. This is a standard problem in the calculus of variations excepting for the choice of an infinite time horizon, as can be seen through substitution. The integral is of the form $\int_0^\infty (k, k')\,\mathrm{d}t$. Using the Euler necessary condition for a minimum value of the functional one can write down implicitly the optimal path for savings over time, provided it exists. In general, the path will not belong to a class of paths characterized by a constant saving ratio over time. Concavity of utility function $u(c)$ and diminishing returns to capital assure that the second order conditions are also satisfied. Ramsey, however, succeeded in deducing through an elegant transformation of the independent variable, namely, time, a very remarkable rule which optimal paths must necessarily satisfy. Keynes provided an intuitive explanation for the same rule. The 'Keynes–Ramsey' rule states that the optimal rate of capital accumulation at any given instant of time multiplied by the marginal utility of optimal consumption at that point of time must equal the excess of the bliss level of utility over the utility of the current optimal level of consumption. The remarkable thing about the Keynes–Ramsey rule is that it is 'altogether independent of the production function except in so far as this determines bliss, the maximum rate of utility obtainable' (Ramsey 1928).

In the presence of time discounting, the integrand becomes $F(k, k', t)$. With this modification,

the Euler differential equation which in general constitutes a second order nonlinear differential equation does not necessarily possess a first integral and hence, the optimum growth path does not lend itself to a simple characterization in terms of decision rule which is formally independent of time ('t').

Ramsey assumes a stationary population, although he allowed for utility maximizing choice on the part of current labour. In the context of the discussion that took place in the early 1960s, population was generally assumed to be growing at an exogenously given rate. Thus $L(t)$ was put at $L_0 e^{nt}$. With this modification, the Ramsey concept of 'bliss' has to be altered.

Assuming a constant returns to scale production function and expressing all relevant variables on a per worker basis, one can derive the relationship $c(t) + k'(t) = f(k(t)) - nk$. Assuming that we are considering only steady growth paths, we have $k'(t) = 0$ and a time independent expression $c = f(k) - nk\ c$. The expression c is maximized for k such that $f'(k) = n$. Under some mild restriction on $f(k)$ this expression can be solved for a finite value of 'k' and the corresponding consumption level $\hat{c}$. $\hat{c}$ can be interpreted as the highest level of sustainable consumption per worker over time, that is the best among the steady states for a given technology. Instead of Ramsey's expression B, we can now write $\int_0^\infty (u(\hat{c}) - u(c))\, dt$ as the integral, and minimize this modified functional subject to the production conditions given earlier. $u(\hat{c})$ is generally referred to in the literature as the utility of consumption attached to the 'golden rule of accumulation'.

Koopmans (1965) and Cass (1965) demonstrated that if $k(0) < \hat{k}$ the optimal paths will approach $\hat{k}$ from below over time whereas if $k(0) < \hat{k}$ it will approach it from above.

Cass included time discounting ($p > 0$), and obtained the 'modified golden rule' for which $f'(k) = n + p$ and deduced the optimal growth path.

For the case of $p = 0$, the Keynes–Ramsey rule is restored again for the same reason as noted in the Ramsey case. However, when population is growing, it is not clear whether one should use the instantaneous utility function $u(c)$ where 'c' is consumption per worker (or per capita, if the participation rate is constant) or a different social welfare functional altogether. Thus, Arrow and Kurz (1970) have argued in favour of maximizing an expression $\int_0^\infty e^{-Pt} u[c(t)]P(t), dt$ where $P(t)$ stands for population at time 't' on the ground that 'if more people benefit, so much the better' (Arrow and Kurz 1970, p. 12). It is clear that in this case, a $P > 0$ is essential if an optimal solution is to exist at all.

The extension of the model to many sector cases was first attempted by Samuelson and Solow (1956). They showed that the Fisher arbitrage rule regarding prices over time could be extended to an n-good case as a necessary property of all optimal paths, no matter which specific utility function is used as it depends only on the question of intertemporal efficiency. An analogue of a 'golden rule' was obtained in situations involving no joint production, a single consumer good and relevant convexity condition.

Linear analysis applied in the neighbourhood of the 'golden rule' solution displays a 'catenary type' behaviour in the one good case, a phenomenon noticed first by Samuelson in a multisectoral context and subsequently proved in the context of closed consumptionless systems by Radner and others. However, any general treatment of n-dimensional cases involving discounted utility functions can throw up pathologies which are not present in simpler cases, especially if joint production is allowed (Samuelson and Liviatan 1969).

B. Revival of discussion on the optimum rate of savings in the early 1960s was motivated by policy considerations. Dissatisfaction with a politically determined rate of savings or with the market solution, especially when the capital market was considered to be subject to considerable imperfection, led economists to look more closely into the character of growth paths based on an ethically explicit criterion function over time. As time is open-ended, the discussion veered towards problems posed by

an infinite planning horizon. With the discovery that optimal paths may not exist with otherwise well behaved production and utility functions, economists devoted a great deal of attention to possible modifications of the Ramsey–Pigou valuation premises to get around the non-existence problem. Koopmans, in particular, felt the need for introducing an assumption relating to time discounting to get over the problem of non-existence and so did Arrow and Kurz.

Some authors tried to explore the sensitivity of finite horizon optimal growth paths to terminal conditions, which in the nature of the case, has to be arbitrary. The idea behind these exercises was to offer an alternative to the procedure of discounting which equally violated the postulate of ethical neutrality between generations (Chakravarty 1962b), the aim being to examine whether optimal paths will prove insensitive at least in their initial phase to terminal conditions, provided the horizon was sufficiently long. Brock (1971) subsequently generalized this type of analysis quite considerably.

Based on these discussions, Hammond and Mirrlees (1973) proposed a category of growth profiles which seemed to avoid the Scylla of 'time discounting' and the Charybdis of a given 'terminal capital stock', by suggesting a category of paths called 'agreeable paths' with the property that if an optimum path exists over an infinite time horizon, it is agreeable.

Furthermore, 'an agreeable path exists if and only if a perpetually feasible) locally optimal path exists'. It is then the maximal locally optimal path' (Hammond and Mirrlees 1973). Hammond subsequently extended the analysis to a multi-sectoral context (1976).

Agreeable paths possess an operational appeal to planners and therefore need to be pursued in greater depth. Among areas of current interest, one can also mention models which relax the assumption of additive separability, which does not seem to be sufficiently strongly grounded in ethical intuition, as well as the assumption of 'stationarity' in the sense defined first by Koopmans (1960).

Despite the existence of several unsolved problems, literature on 'optimal savings' has been of interest to economic theorists for having explored with considerable thoroughness the 'open-endedness' of the future from a national decision-theoretic point of view and for providing a convenient parametric method of generating optimal growth paths in a precise sense of the term with associated dual prices, which can be used for social benefit–cost analysis. It has also posed a philosophical issue of broader interest as to whether one can adopt ethical principles that are independent of environmental consideration in the broad sense of the term (i.e., population growth and/or technological progress).

Bibliography

Arrow, K.J., and M. Kurz. 1970. *Public investment, the rate of return, and optimal fiscal policy*. Baltimore: Johns Hopkins Press.

Brock, W.A. 1971. Sensitivity of optimal paths with respect to a change in target stocks. *Zeitschrift für Nationalökonomie*, Supplement 1.

Cass, D. 1965. Optimal growth in an aggregative model of capital accumulation. *Review of Economic Studies* 32: 233–240.

Chakravarty, S. 1962a. The existence of an optimum savings program. *Econometrica* 30: 178–187.

Chakravarty, S. 1962b. Optimal savings with finite planning horizon. *International Economic Review* 3: 338–355.

Dasgupta, P.S., and G. Heal. 1979. *The economic theory of exhaustible resources*. London: Nisbet.

de Grandville, O. 1980. Capital theory, optimal growth and efficiency conditions with exhaustible resources. *Econometrica* 48(7): 1763–1776.

Gale, D. 1967. On optimal development in a multi-sector economy. *Review of Economic Studies* 34: 1–18.

Goodwin, R.M. 1961. The optimal growth path for an underdeveloped economy. *Economic Journal* 71: 756–774.

Hammond, P.J., and J.A. Mirrlees. 1973. Agreeable plans. In *Models of economic growth*, ed. J.A. Mirrlees and N. Stern. London: Macmillan.

Hotelling, H. 1931. The economics of exhaustible resources. *Journal of Political Economy* 39: 137–175.

Inagaki, M. 1970. *Optimal economic growth; Finite shifting vs infinite time horizon*. Amsterdam: North-Holland.

Keynes, J.M. 1930. Ramsey, F.P.: An obituary. *Economic Journal* 40: 153–154.

Koopmans, T.C. 1960. Stationary ordinal utility and impatience. *Econometrica* 28: 287–309.

Koopmans, T.C. 1965. On the concept of optimal economic growth. In *The econometric approach to development planning.* Amsterdam/Chicago: North-Holland/Rand MacNally. (A reissue of *Pontificiae Academiae Scientarium Scripta Varia*, vol. XXVIII, 1965).
Koopmans, T.C. 1967. Objectives, constraints, and outcomes in optimal growth models. *Econometrica* 35: 1–15.
Magill, M.J.P. 1981. Infinite horizon programs. *Econometrica* 49(3): 679–711.
Mirrlees, J.A. 1967. Optimum growth when technology is changing. *Review of Economic Studies* 34: 95–124.
Pigou, A.C. 1952. *The economics of welfare*. London: Macmillan.
Ramsey, F.P. 1928. A mathematical theory of saving. *Economic Journal* 38: 543–559.
Samuelson, P.A., and N. Liviatan. 1969. Notes on 'turnpikes', stable and unstable. *Journal of Economic Theory* 1(4): 454–475.
Samuelson, P.A., and R.M. Solow. 1956. A complete capital model involving heterogeneous capital goods. *Quarterly Journal of Economics* 70: 537–562.
Tinbergen, J. 1960. Optimum savings and utility maximization over time. *Econometrica* 28: 481–489.
von Weizsäcker, C.C. 1965. The existence of optimal programmes of accumulation for an infinite time horizon. *Review of Economic Studies* 32: 85–104.

Optimal Tariffs

Nuno Limão

Abstract

Optimal tariffs allow a country to exploit its market power in international trade. A country can improve its terms of trade by unilaterally restricting its exports if it faces a downward-sloping demand for them or restricting its imports if it faces an upward-sloping foreign export supply. This argument against unilateral free trade is over 150 years old but it remains central to modern theories that explain trade agreements and their rules. This, along with recent evidence that prior to such agreements countries exploit their market power in trade, shows that optimal tariffs may be an important positive theory of protection.

Keywords

Corn Laws; Cross-elasticities; Free trade; Imperfect competition; Marginal rates of transformation; Market power; Marketing boards; Monopolistic competition; Monopsony pricing; Optimal tariffs; Optimal taxation; Smoot Hawley Tariff Act of 1930 (USA); Tariffs; Terms of trade; Torrens, R.; Trade agreements; Trade policy, political economy of

JEL Classifications

F1

A country that faces a downward-sloping demand for its exports has market power and therefore, as a monopolist, can benefit from restricting its export supply. When a country's exporters are perfectly competitive, the government can coordinate this restriction via an export tax, which increases the world price for its exports and so improves its terms of trade. Analogously, a country facing an upward-sloping export supply has market power in imports and can benefit from restricting them via a tariff. Generally, the optimal tariff is defined as the rate that unilaterally maximizes a country's welfare and is given by the inverse elasticity of foreign export supply, as determined by optimal monopsony pricing.

The terms-of-trade argument against unilateral free trade is over 150 years old, yet it remains one of the hardest to refute theoretically. The reason is simple. A country's atomistic consumers impose an externality on each other since, by increasing import demand, they raise the equilibrium price for all. The optimal instrument to correct an externality must target it at the source (Bhagwati and Ramaswami 1963), and the optimal tariff does this by reducing import demand. This quantity reduction entails a cost but, for a sufficiently small tariff, it is more than offset by the improved terms of trade. This is one of the only cases when, in the absence of retaliation, the tariff is a first-best instrument. However, if a tariff improves a country's terms of trade, it worsens those of its trading partner, who is therefore likely to retaliate. The typical trade war outcome is to leave

both worse off relative to free trade, which explains many economists' opposition to optimal tariffs as a normative theory. The trade war outcome points to the benefits from reciprocal tariff reductions and as such the terms-of-trade argument remains central to modern theories of trade agreements and their rules. This, along with recent evidence that prior to such agreements countries exploit their market power in trade, shows that optimal tariffs may be an important positive theory of protection rather than an irrelevant normative one.

Informal Derivation and Applications

The standard derivation of the optimal tariff focuses on a standard neoclassical economy with no domestic externalities and available lump-sum transfers to address any resulting redistribution issues (see Graaf 1949–50). A Pareto optimum for the closed economy requires the domestic marginal rate of substitution between any two goods i and j to equal their marginal rates of transformation, which in a competitive economy is done via the domestic relative prices, that is, $MRS_{ij} = p_i/p_j = MRT_{ij}$. An open economy can exchange goods at the prevailing world prices, which can be thought of as having access to a new technology or foreign rate of transformation. Now efficient production requires the domestic MRT_{ij} to equal the marginal foreign rate of transformation (MFRT). Optimal ad valorem tariffs, t_i, imposed on the world prices, π_i, ensure that this additional condition for efficiency is met, by introducing a wedge such that the relative price faced by domestic producers is equal to $MFRT_{ij}$, that is, $p_i/p_j = \pi_i(1 + t_i)/\pi_j(1 + t_j)$. The final step is to determine the $MFRT_{ij}$, which simply reflects the relative marginal cost of these goods in the world market. The marginal cost for the importer is $\pi_i + a_i$, the price paid for the unit plus the marginal change in price(S) it causes, $a_i = \Sigma_k m_k \partial p_k/\partial m_i$. Therefore the optimal *ad valorem* tariff rates are determined by

$$\frac{\pi_i(1+t_i)}{\pi_j(1+t_j)} = \frac{\pi_i(1+a_i/\pi_i)}{\pi_j(1+a_j/\pi_j)} \quad \text{for all } i \text{ and } j, \tag{1}$$

which is satisfied by any tax structure such that

$$t_i = \frac{a_i}{\pi_i} \quad \text{for all } i. \tag{2}$$

When all cross-price elasticities are zero, a_i/π_i is simply the inverse of the foreign export supply for i, $1/\varepsilon_i$, and we obtain the standard formula, $t_i = 1/\varepsilon_i$. Otherwise t_i also includes the cross-elasticities, as a_i captures the weighted sum of marginal world price changes in all goods due to the increase in demand for i.

Since the cross-effects in a_i can be negative the optimal tariff may be zero or even a subsidy on any or all goods. However, that can't be the case with only one import and one export good, $i = m$ and e, as is easily shown if their cross-elasticity is zero. To see this, note that with two goods we can attain the same outcome with either a tariff or an export tax (Lerner 1936), which simultaneously accounts for market power in imports and exports. Solving (1) with $t_e = 0$ we have the import tariff rate

$$t_m = \frac{1/\varepsilon_m + 1/\varepsilon_e}{1 - 1/\varepsilon_e}, \tag{3}$$

which is positive given the positively defined elasticities of foreign export supply, ε_m, and foreign import demand, ε_e, and $1/\varepsilon_e < 1$.

The result extends in several ways under perfect competition settings. If a domestic distortion exists, and is addressed by a first-best instrument, then the rate in (2) is generally still optimal. Graaf (1949–50) shows this for external (dis)economies in production, for example. The equivalence of tariffs and quantity restrictions, that is, quotas, under certainty implies that quotas can be used to the same effect provided that their rents accrue to the country that imposes them (but the welfare and trade volume outcomes of tariff and quota wars differ, as shown by Rodriguez 1974). Kemp (1966) and Jones (1967) derive the optimal tax structure when capital is mobile and a country has market power in goods and factors trade. Similarly to (2) the optimal tariffs on goods take into account their effect on the price of capital and vice versa.

Tariffs can also affect a country's terms of trade under imperfect competition. However, there are fewer general results in these settings, even in the simpler cases of zero cross-price elasticities. Nonetheless, a few points are worth noting. First, if a country has a monopoly importer or exporter (for example, an agricultural marketing board or a cartel of oil exporters such as OPEC) then there is no first best role for a trade tax – a monopolist would already optimally restrict quantities. A tariff may still be necessary to internalize effects from any cross-elasticities, as Gros (1987) shows for a monopolistic competition model. Second, under imperfect competition a tariff can affect a country's terms of trade even if it has an infinitesimally small share of the world's expenditure. For example, if imposing a small *ad valorem* tariff reduces the import demand elasticity, then it also improves the welfare of a small importer facing a monopoly exporter (Katrak 1977; Brander and Spencer 1984). However, when the country is small the tariff is generally not the first best instrument.

Early Contributions and Current Relevance as a Positive Theory

There have been four important waves in the development and application of the optimal tariffs idea. They were each about 40 to 50 years apart, and at least three appear to be linked to important policy events.

Several early advances in trade theory arose during the debate over the repeal of British import duties imposed by the Corn Laws of 1815. Robert Torrens, a famous classical economist, initially supported the repeal but eventually turned against unilateral free trade as he understood that countries may gain from tariffs through an improvement in their terms of trade. This basic idea and the intuition for it are found in Torrens (1833, 1844) and Mill (1844). However, a country will actually gain only if the terms-of-trade benefit offsets the cost from lower import volume; in a second phase of development, Edgeworth (1894) shows that this is the case unless the foreign country's offer curve is perfectly elastic, while Bickerdike (1906, 1907) develops the first optimal tariff formula, similar to (3).

Renewed interest in the topic came after the Smoot Hawley Tariff Act of 1930, which raised US average tariffs to about 50 per cent and triggered a cycle of tariff retaliation. The key contributions by Kaldor (1940), Scitovsky (1942) and Johnson (1953–4) focus on the outcomes when countries retaliate. Johnson (1953–4) shows the outcome of a tariff war using tariff reaction curves that summarize a country's best response. He confirms that two *symmetric* countries prefer free trade to a trade war; but otherwise one of them may be better off under a trade war.

The latest developments in the topic also came in the wake of important economic events. Mayer (1981) examines the possible tariff outcomes under the tariff cutting formulas used in the 1973–9 multilateral trade negotiations under the General Agreement on Tariffs and Trade (GATT). Since then numerous authors have relied on the tariff war equilibrium as the threat point for the theoretical analysis of multilateral and bilateral trade agreements. Notably, Bagwell and Staiger (1999, 2002) argue that the purpose of the GATT and its successor, the World Trade Organization, is to allow countries to reciprocally lower protection in a way that eliminates the terms-of-trade component of tariffs, and show that such an economic theory of GATT can explain several of its key rules.

Despite the success of the terms-of-trade motive for tariffs in explaining important features of trade agreements, its power as a positive theory of trade protection is often questioned for two reasons. First, governments do not set tariffs to maximize social welfare. Although governments often set tariffs to redistribute income across interest groups, this does not imply that tariffs will not reflect market power. For example, Johnson (1950) derives the revenue-maximizing tariff rate, which does not maximize welfare but is nonetheless increasing in a country's market power since a given tariff rate yields higher revenue, under a less elastic export supply. Moreover, recent micro-founded political economy models predict that a large country's unilateral tariff reflects its market power, even if the government

places no weight on social welfare (Grossman and Helpman 1995). Thus, even if the primary objective of the government is a political economy one, its tariffs can reflect market power since this allows it to achieve that objective at a lower cost as it captures some income from its trading partners via improved terms of trade.

The second critique is the argument that most countries cannot affect their terms of trade and so it is not an important determinant of protection. Critics concede that certain commodity exporters do have some market power and have at times exerted it (for example, OPEC's oil restrictions or export taxes by marketing boards). But evidence of market power in exports appears to go beyond these obvious cases since aggregate estimates of ε_e in (3) are often found to be low, sometimes close to unity. Nonetheless, there are considerable difficulties in estimating and interpreting such aggregate elasticities, which are often estimated only for countries already setting their tariffs cooperatively. Therefore cross-country comparisons of average tariffs and these aggregate elasticities cannot provide much insight into the empirical importance of the terms-of-trade motive.

There is also growing evidence of market power in imports since when countries change their exchange rates or tariffs part of the effect is absorbed by the foreign exporters (cf. Kreinin 1961). Broda et al. (2006) provide compelling evidence that countries have and exploit their market power. They estimate inverse foreign export supply elasticities by good and country, and find that even small countries have some market power, which is increasing in country size and degree of good differentiation. They then examine tariffs for countries that are not setting them cooperatively and find that they are set higher in goods with higher inverse elasticities. They conclude that market power is an economically and statistically important determinant of tariffs.

In sum, optimal tariffs are evolving from a curious normative theory to a positive one. The broad applicability of the terms-of-trade motive for tariffs; its theoretical success in explaining important rules of trade agreements; and the recent evidence that countries exploit their market power, all indicate this will remain a key concept in economics.

See Also

- ▶ Bickerdike, Charles Frederick (1876–1961)
- ▶ Corn Laws, Free Trade and Protectionism
- ▶ Johnson, Harry Gordon (1923–1977)
- ▶ Tariffs
- ▶ Torrens, Robert (1780–1864)

Bibliography

Bagwell, K., and R.W. Staiger. 1999. An economic theory of GATT. *American Economic Review* 89: 215–248.

Bagwell, K., and R.W. Staiger. 2002. *The economics of the world trading system*. Cambridge, MA: MIT Press.

Bhagwati, J., and V.K. Ramaswami. 1963. Domestic distortions, tariffs and the theory of optimum subsidy. *Journal of Political Economy* 71: 44–50.

Bickerdike, C.F. 1906. The theory of incipient taxes. *Economic Journal* 16: 529–535.

Bickerdike, C.F. 1907. Review of A.C. Pigou's protective and preferential import duties. *Economic Journal* 17: 98–108.

Brander, J.A., and B.J. Spencer. 1984. Tariff protection and imperfect competition. In *Monopolistic competition and international trade*, ed. H. Kicrzkowski. Oxford: Oxford University Press.

Broda, C., N. Limão, and D. Weinstein. 2006. *Optimal tariffs: The evidence*. Working Paper No. 12033. Cambridge, MA: NBER.

Edgeworth, F.Y. 1894. The theory of international values. *Economic Journal* 4: 35–50.

Graaf, J.V. 1949–50. On optimum tariff structures. Review of Economic Studies 17, 47–59.

Gros, D. 1987. A note on the optimal tariff, retaliation and the welfare loss from tariff wars in a framework with intra-industry trade. *Journal of International Economics* 23: 357–367.

Grossman, G., and E. Helpman. 1995. Trade wars and trade talks. *Journal of Political Economy* 103: 675–708.

Johnson, H.G. 1950. Optimum welfare and maximum revenue tariffs. *Review of Economic Studies* 19: 28–35.

Johnson, H.G. 1953–4. Optimum tariffs and retaliation. Review of Economic Studies 21, 142–153.

Jones, R.W. 1967. International capital movements and the theory of tariffs and trade. *Quarterly Journal of Economies* 81: 1–38.

Kaldor, N. 1940. A note on tariffs and the terms of trade. *Economica* 7: 377–380.

Katrak, H. 1977. Multi-national monopolies and commercial policy. *Oxford Economic Papers* 29: 283–291.

Kemp, M.C. 1966. The gains from international trade and investment: A neo-Heckscher–Ohlin approach. *American Economic Review* 65: 788–809.
Kreinin, M.E. 1961. Effect of tariff changes on the prices and volume of imports. *American Economic Review* 51: 310–324.
Lerner, A.P. 1936. The symmetry between import and export taxes. *Economica* 3: 306–313.
Mayer, W. 1981. Theoretical considerations on negotiated tariff adjustments. *Oxford Economic Papers* 33: 135–153.
Mill, J.S. 1844. *Essays on some unsettled questions of political economy.* London: Parker.
Rodriguez, C. 1974. The non-equivalence of tariffs and quotas under retaliation. *Journal of International Economics* 4: 295–298.
Scitovsky, T. 1942. A reconsideration of the theory of tariffs. *Review of Economic Studies* 9: 89–110.
Torrens, R. 1833. *Letters on commercial policy.* London: Longman.
Torrens, R. 1844. *The budget: On commercial and colonial policy.* London: Smith-Elder.

Optimal Taxation

Louis Kaplow

Abstract

Optimal taxation concerns how various forms of taxation should be designed to maximize social welfare. The task requires an integrated consideration of the revenue-raising and distributive objectives of taxation. The central instrument in developed economies is the labour income tax, the analysis of which was pioneered by Mirrlees (*Review of Economic Studies* 68:175–208, 1971). Subsequently, Atkinson and Stiglitz (*Journal of Public Economics* 6:55–75, 1976) showed how commodity taxes should be set in the presence of an optimal income tax, the results differing qualitatively from, and in important respects displacing, the teachings derived from Ramsey's (*Economic Journal* 37:41–61, 1927) seminal analysis of the pure commodity tax problem.

Keywords

Ability; Commodity taxation; Externalities; Income taxation; Labour supply; Leisure; Linear income tax; Lump-sum taxes; Marginal cost pricing; Marginal tax rates; Marginal utility of consumption; Mirrlees, J.; Nonlinear income tax; Optimal government policy; Optimal tax systems; Optimal taxation; Pigouvian taxes; Public goods; Ramsey taxation; Redistribution; Revelation principle; Separable preferences; Social preferences; Social welfare function; Taxation of capital; Taxation of income; Transfer programmes; Uniform taxation; Value-added tax

JEL Classifications

H2

Optimal taxation concerns the question of how various forms of taxation should be designed in order to maximize a standard social welfare function subject to a revenue constraint. The task requires an integrated consideration of the revenue-raising and distributive objectives of taxation. The central instrument in developed economies is the labour income tax. Mirrlees (1971) pioneered the analysis of this challenging problem. Subsequently, Atkinson and Stiglitz (1976) showed how commodity taxes should be set in the presence of an optimal income tax. The results are qualitatively different from – and in important respects displace – prior teachings that originate in Ramsey's (1927) analysis of the pure commodity tax problem. In addition to setting particular taxes optimally, it is also necessary to choose optimally among tax systems.

Income Taxation

Model

The standard optimal income tax model involves a one-period setting in which individuals' only choice variable is their degree of labour effort l. There is a single composite consumption good c. An individual's utility is given by $u(c, l)$, where $u_c > 0$ and $u_l < 0$. An individual's consumption is given by

epsilon epsilon epsilon (1)

where w is the individual's wage rate and T is the tax-transfer function.

The motivation for redistributive taxation is that individuals differ, in particular in their wages, that is, their earning abilities. The distribution of abilities will be denoted $F(w)$, with density $f(w)$. Individuals' wage rates are taken to be exogenous. Their pre-tax earnings wl are the product of their wage rate and level of labour effort. More broadly, one can interpret labour effort as including not only hours of work but also intensity and not only productive effort but also investments in human capital.

Taxes and transfers, $T(wl)$, at any income level may be positive or negative. The (uniform) level of the transfer received by an individual earning no income, that is, $-T(0)$, is sometimes referred to as the grant g. Taxes may be interpreted broadly, to include sales taxes or value-added tax (VAT) payments in addition to income taxes. Transfers include those through the tax system in addition to welfare programmes. The inclusion of transfers is important both practically, since they are in fact significant, and conceptually, since otherwise redistribution would be limited to transfers between the rich and the middle class, once the poor were exempted from the tax system.

Taxes and transfers are taken to be a function of individuals' incomes, assumed to be observable, and it is this dependence of taxes on income that is the source of distortion. If taxes could instead depend directly on individuals' abilities, w, individualized lump-sum taxes would be feasible and redistribution could be accomplished without distorting labour supply. Ability, however, is assumed to be unobservable.

The government's problem is to choose $T(wl)$ to maximize social welfare, which can be stated as

$$\int W(u(c(w), l(w)))f(w)dw, \tag{2}$$

where c and l are each expressed as functions of w to refer to the levels of consumption achieved and labour effort chosen by an individual of ability w. If W is linear, the welfare function is utilitarian, whereas if W is strictly concave, additional weight is given to inequality in utility levels (not just levels of marginal utilities).

This maximization is subject to a revenue constraint and to constraints regarding individuals' behaviour. The former is

$$\int T(wl(w))f(w)dw = R, \tag{3}$$

where R is an exogenously given revenue requirement. Here, revenue is to be interpreted as expenditures on public goods that should be understood as implicit in individuals' utility functions; because these expenditures are taken to be fixed, they need not be modelled explicitly. Regarding the latter constraints, individuals are assumed to respond to the given tax schedule optimally, which determines the functions $c(w)$ and $l(w)$.

Mirrlees's (1971) original exposition has been followed by subsequent elaborations, much of which is synthesized and extended in Atkinson and Stiglitz (1980), Stiglitz (1987), Tuomala (1990), and Salanié (2003). Because the problem is formidable, the present discussion will be confined to stating basic results, such as are embodied in first-order conditions and produced by simulations.

Linear Income Tax

Substantial illumination with greatly reduced complexity is provided by first examining a linear income tax,

$$T(wl) = twl - g, \tag{4}$$

where t is the (constant, income-independent) marginal tax rate and g, as previously noted, is the uniform per-capita grant. Because of the presence of the grant g, a linear income tax can be highly redistributive (consider setting t at 100 per cent and g equal to mean income net of any per capita revenue requirement – in the absence of incentive constraints) or not at all redistributive (t may be 0 per cent and g equal to the negative of the per capita revenue requirement). Foreshadowing discussion of the nonlinear income tax, the degree of redistribution is more directly related to the levels of t and g than to the shape (deviation from linearity) of the tax schedule.

To derive the optimal linear income tax, the government's maximization problem can be written in Lagrangian form as choosing t and g to maximize

$$\int [W(u((1-t)wl(w)+g, l(w))) + \lambda(twl(w)-g-R)]f(w)\,dw, \tag{5}$$

where λ is the shadow price of revenue, referring to the constraint (3), and expression (4) is substituted into expression (1) so that consumption is expressed in terms of the specific linear tax system under consideration. Following Atkinson and Stiglitz (1980) and Stiglitz (1987), the first-order condition for the optimal tax rate can usefully be expressed as

$$\frac{t}{1-t} = -\frac{\mathrm{cov}(\alpha(w), y(w))}{\int y(w)\varepsilon(w)f(w)dw}, \tag{6}$$

where $y(w) = wl(w)$, income earned by individuals of ability w; $\varepsilon(w)$ is the compensated elasticity of labour effort of individuals of ability w; and $\alpha(w)$ is the net social marginal valuation of income, evaluated in dollars, of individuals of ability w:

$$\alpha(w) = \frac{W' u_c(w)}{\lambda} + tw\left(\frac{\partial l(w)}{\partial g}\right). \tag{7}$$

The numerator of the first term on the right side of expression (7) indicates how much additional (lump-sum) income to an individual of ability w contributes to social welfare (u_c indicates how much utility rises per dollar and W' indicates the extent to which social welfare increases per unit of utility) and this product is converted to a dollar value by dividing by the shadow price of government revenue. The second term takes into account the income effect, namely, that giving additional lump-sum income to an individual of ability w will reduce labour effort ($\partial l(w)/\partial g < 0$), which in turn reduces government tax collections by tw per unit reduction in $l(w)$.

Expression (6) indicates how various factors affect the optimal level of a linear income tax. Beginning with the numerator on the right side, a higher (in magnitude) covariance between α and y favours a higher tax rate. In the present setting, $\alpha(w)$ will (under assumptions ordinarily postulated) be falling with income. Note that a larger covariance does not involve a closer (negative) correlation but rather a higher dispersion (standard deviation) of α and of y. The dispersion of α will tend to be greater the more concave (egalitarian) is the welfare function W and the more concave is utility as a function of consumption (that is, the greater the rate at which marginal utility falls with income). Income, y, will have a higher dispersion (again, under standard assumptions) when the distribution of underlying abilities is more unequal. In sum, more egalitarian social preferences, more rapidly declining marginal utility of consumption, and higher underlying inequality each contribute to a higher optimal tax rate.

The denominator indicates that a higher compensated labour supply elasticity favours a lower tax rate. The other terms in the integrand indicate that, ceteris paribus, the labour supply elasticity matters more with regard to high-income individuals and at ability levels where there are more individuals (typically the middle of the income distribution) because of the greater sacrifice in revenue.

The foregoing exposition is incomplete in not emphasizing the various respects in which income effects are relevant (they influence α and also λ) and in ignoring that the values on the right side of expression (6) are endogenous. Especially for the latter reason, the literature has relied heavily on simulations.

The most-reported optimal linear income taxation simulations are those of Stern (1976). For his preferred case – an elasticity of substitution between consumption and labour of 0.4, a government revenue requirement of 20 per cent of national income, and a social marginal valuation of income that decreases roughly with the square of income – he finds that the optimal tax rate is 54 per cent and that individuals' lump-sum grant equals 34 per cent of average income. To illustrate the benefits of redistribution, he finds that a scheme that uses a lower tax rate, just high enough to finance government programmes (that is, with a grant of zero), produces a level of social welfare that is lower by an amount equivalent to

approximately 5 per cent of national income. If there is very little weight on equality, the optimal tax rate is only 25 per cent, whereas if there is extreme weight on equality, the optimal tax rate is 87 per cent. Returning to his central case, an extremely low labour supply elasticity implies an optimal tax rate of 79 per cent, and an elasticity as high as had been used in some earlier literature implies an optimal tax rate of 35 per cent. In the absence of the need to finance government expenditures, the optimal tax rate is 48 per cent, and if government expenditures are twice as high, the optimal tax rate is 60 per cent.

Nonlinear Income Tax

Mirrlees (1971) and subsequent investigators employ control-theoretic techniques to address the more general formulation of the optimal nonlinear income taxation problem, which requires choosing an entire tax schedule T (wl) rather than a single tax rate. In this maximization, the constraints regarding individuals' maximizing behaviour entail that no individual of any type w will prefer the choice specified for any other type w°. This approach is related to the use of the revelation principle in work on mechanism design, and in similar spirit many researchers following Stiglitz (1982) and others analyse a simpler, discrete variant of the problem, often involving two types, in which the binding incentive constraint is usually that the high-ability type not have an incentive to mimic the low-ability type in order to pay less tax.

The analysis of the continuous case can be summarized in a first-order condition for the optimal marginal income tax rate at any income level y^*, where w^* and l^* correspond to the ability level and degree of labour effort supplied by the type of individual who would earn y^*. Making the simplifying assumptions that utility is separable between consumption and labour effort and that marginal utility u_c is constant, the condition can be expressed as

$$\frac{T'(w^*l^*)}{1-T'(w^*l^*)} = \frac{1-F(w^*)}{\xi^* w^* f(w^*)} \int_{w^*}^{\infty} \left(1 - \frac{W'(u(w))u_c}{\lambda}\right) f(w)dw \Big/ 1-F(w^*), \tag{8}$$

where $\xi^* = 1/(1 + l^* u_{ll}/u_l)$ – which, when marginal utility is constant as assumed here, equals $\varepsilon/(1 + \varepsilon)$, where ε again is the elasticity of labour supply. For derivations of related expressions, see, for example, Auerbach and Hines (2002), Atkinson and Stiglitz (1980), Dahan and Strawczynski (2000), Diamond (1998), Saez (2001), and Stiglitz (1987). Note that this formulation (like those in recent literature) includes $1 - F(w^*)$ in both the numerator and the denominator on the right side. The motivation is that, in the first term, $(1 - F(w^*))/f(w^*)$ is purely a property of the distribution of w, and, in the second term, because the numerator is an integral from w^* to ∞, the term as a whole gives an average value for the expression in parentheses in the integrand. Both aspects aid intuition, as will be seen in the discussion to follow.

Expression (8), being a first-order condition, should be interpreted by reference to an adjustment that slightly raises the marginal tax rate at income level y^* (say, in a small interval from y^* to $y^* + \delta$), leaving all other marginal tax rates unaltered. There are two effects of such a change. First, individuals at that income level face a higher marginal rate, which will distort their labour effort, a cost. Second, all individuals above income level y^* will pay more tax, but these individuals face no new marginal distortion. That is, the higher marginal rate at y^* is inframarginal for them. Since those thus giving up income are an above-average-income slice of the population (it is the part of the population with income above y^*), there tends to be a redistributive gain.

The right side of expression (8) can readily be interpreted in terms of this perturbation (although it should be kept in mind that this interpretation omits, inter alia, income effects and the endogeneity of variables). Begin with the first term. Revenue is collected from all individuals with incomes above y^*, which is to say all ability types above w^*; hence the $1 - F(w^*)$ in the numerator. This factor favours marginal tax rates that fall with income. As there are fewer individuals who face the inframarginal tax, the core benefit of a higher marginal rate declines. In the extreme, if there is a highest known type in the

income distribution, the optimal marginal rate at the top would be zero because $1 - F$ would be zero: a higher rate collects no revenue but distorts the behaviour of the top individual. However, when there is no highest type, known with certainty in advance, this result is inapplicable. Furthermore, with a known highest type, simulations suggest that zero is not a good approximation of the optimal marginal tax rate even quite close to the top of the income distribution, so the zero-rate-at-the-top result is of little practical importance.

To continue with the first term, raising the marginal rate at a particular point distorts only the behaviour of the marginal type, which explains the $f(w^*)$ in the denominator. For standard distributions, this factor is rising initially and then falling, which favours falling marginal rates at the bottom of the income distribution and rising rates at the top. The denominator also contains weights of ξ^*, indicating the extent of the distortion, and w^*, indicating how much productivity is lost per unit of reduction in labour effort. The elasticity is often taken to be constant, although some empirical evidence on the elasticity of taxable income supports a rising elasticity due to the greater ability of higher-income individuals to avoid taxes. This consideration may favour marginal rates that fall with income. Finally, w^* is rising, which also favours falling marginal rates: The greater is the wage (ability level), the greater is the revenue loss from a given decline in labour effort.

The second term applies a social weighting to the revenue that is collected. The expression in parentheses in the integrand in the numerator is the difference between the marginal dollar that is raised and the dollar equivalent of the loss in welfare that occurs on account of individuals above w^* paying more tax. As in the interpretation of expression (7), u_c is the marginal utility of consumption to such individuals, W' indicates the impact of this change in utility on social welfare, and division by λ, the shadow price on the revenue constraint, converts this welfare measure into dollars. This integral is divided by $1 - F(w^*)$, which as noted makes the second term an average for the affected population.

This term tends to favour marginal rates that rise with income. The greater is w, the lower is W' (unless the welfare function is utilitarian, in which case this is constant) and the lower would be the marginal utility of income u_c (had we not abstracted from this effect in the simplifying assumptions). Hence, at a higher w^* the average value of the term subtracted in the integrand is smaller, making the entire term larger. Note further that, if social welfare or utility is reasonably concave, $W'u_c$ will approach zero at high levels of income, at which point this term will be nearly constant in w^*. That is, the term favours rising marginal tax rates when income is low or moderate, but has little effect on the pattern of marginal tax rates near the top of the income distribution.

Because of difficulties in determining the shape of the optimal income tax schedule by mere inspection of the first-order condition (8), analysts beginning with Mirrlees (1971) have used simulations to help join the theoretical analysis with empirical estimates of labour supply elasticities and of the distribution of skills or income in order to provide further illumination. Tuomala (1990) offers a useful survey and set of calculations. In all the cases he reports, marginal tax rates fall as income increases, except at very low levels of income. Mirrlees's (1971) original calculations had displayed a similar tendency, but subsequent researchers had questioned the extent to which this result may have depended on the social preferences he stipulated or the arguably high labour supply response he assumed. Later work, however, suggests that a greater social preference for equality or a lower labour supply response tends to increase the level of optimal marginal tax rates but does not generally result in a substantially different shape. This phenomenon is also illustrated by Slemrod et al. (1994), who examine the optimal two-bracket income tax. In all of their simulations, the optimal upper-bracket marginal rate is lower than the lower-bracket rate; indeed, this gap widens as the social preference for equality increases because of the additional value of raising the lower-bracket rate in generating funds to increase the grant, which is of greatest relative benefit to the lowest-income individuals.

Subsequent work further explores the circumstances in which optimal marginal tax rates might rise with income. Kanbur and Tuomala (1994) find that, when inequality in individuals' abilities (wages) is significantly greater than previously assumed (but at levels they suggest to be empirically plausible), optimal marginal tax rates do increase with income over a substantial range, although for upper-income individuals optimal marginal rates still fall with income. Diamond (1998) examines a Pareto distribution of skills (instead of the commonly used lognormal distribution), under which the $(1 - F)/f$ component of expression (8) rises more rapidly at the top of the distribution, and finds that optimal marginal tax rates are rising at the top. However, Dahan and Strawczynski's (2000) simulations indicate that Diamond's result was driven in large part by his additional assumption that preferences were quasi-linear, thus removing income effects. (Nevertheless, their diagrams do suggest that, consistent with Diamond's claim, moving from a lognormal to a Pareto distribution favours higher rates – still falling, but notably less rapidly – at the top of the income distribution.) Saez (2001), using income distribution data in the United States from 1992 and 1993, finds that the shape of the distribution of $(1 - F)/wf$ is such that optimal rates should fall substantially well into the middle of the income distribution, to an income of approximately $75,000, rise until approximately $200,000, and then be essentially flat thereafter.

An additional result from the simulations is that, at the optimum, a nontrivial fraction of the population does not work, and this fraction is larger when social preferences favour greater redistribution and when the labour supply elasticity is higher. This outcome should hardly be surprising because, as the analysis of expression (8) and the simulations suggest, high marginal rates tend to be optimal at the bottom of the income distribution, along with a sizable grant. Relatedly, little productivity and thus little tax revenue is sacrificed when those with very low abilities are induced not to work (whereas substantial revenue is raised from the rest of the population, for whom marginal tax rates on the first dollars of income are inframarginal).

Extensions

Given the central importance of income taxation to the revenue and distributive objectives of government, further exploration of various aspects of the problem should be a high research priority. A number of features have received some, although generally quite limited, attention. For broader discussions and further references, see Atkinson and Stiglitz (1980), Stiglitz (1987), Tuomala (1990), Salanié (2003), and Kaplow (2008).

A critical assumption in optimal income tax analysis is that earning ability is unobservable so that income, a signal of ability, is taxed instead, which is the source of distortion. Hence, it is worth considering the possibilities for basing taxation more directly on ability. To some degree, hours may be observable, and ability (wages) can thus be inferred. But in many occupations (notably, self-employment) hours are difficult to observe, and both hours and wages are manipulable, such as by extending reported hours and lowering the reported wage. Another approach would be to measure proxies of earning ability, such as through testing. Unfortunately, skills measurable by testing explain only some of the variance in earning ability, and, if taxes were to be based on test results or other ability measures, individuals would adjust their performance and thereby distort the measurement. A third technique – one sometimes employed – is to adjust taxes and transfers for observable personal attributes, such as physical disability, age or family composition.

In general, tax and transfer schedules could be made a function of various imperfect signals of ability (or of other pertinent differences, such as in utility functions). For each value of the signal, there would in essence be a different tax schedule, governed by the first-order condition (8); each of these tax schedules would, however, be linked in a common optimization by the shadow price A. One might view models like those of Akerlof (1978), in which he assumes that a subset of the lowest-ability group can be identified perfectly ('tagged'), and Stern (1982), in which he examines the usefulness of a noisy signal of ability in a two-type model, as special cases of this more general formulation.

There exist myriad additional complications. One is that income may be a noisy signal of ability, whether because of variations in occupations (for a given ability, one job may pay more to compensate for specific disamenities) or in preferences (an individual may earn more not because of greater ability but rather due to a higher marginal utility of consumption or a lower marginal disutility of labour effort). Another possibility is that individuals may have preferences concerning redistribution itself, perhaps due to altruism or envy. Other topics that have been explored include liquidity constraints, general equilibrium effects of taxation on the distribution of pre-tax wages, uncertainty, interactions with non-tax distortions, and human capital.

Commodity Taxation

Commodity Taxation with Income Taxation

To examine optimal commodity taxation with labour income taxation, the foregoing model can be modified as follows. In place of consumption c, individuals choose commodity vectors $\boldsymbol{x}$ and, as before, labour effort l to maximize the utility function $u(x, l)$. On the left side of individuals' budget constraints (1), c is replaced by $\boldsymbol{\rho x}$, where $\boldsymbol{\rho}$ is the consumer price vector equal to $\boldsymbol{p} + \boldsymbol{\tau}$: the sum of a producer price vector (taken to be constant and equal to production costs) and a vector of commodity taxes (which, if negative, are subsidies).

Atkinson and Stiglitz (1976) demonstrate that, when the income tax is set optimally, commodity taxes should be undifferentiated, that is, $\boldsymbol{\tau} = 0$, when utility is weakly separable in labour (on which more in a moment). Alternatively, other levels of $\boldsymbol{\tau}$ are similarly optimal as long as the ratio of any two consumer prices equals the ratio of producer prices, with the difference in consumer price level being offset by an adjustment to the income tax schedule. (For example, if all commodity taxes are ten per cent rather than zero, the income tax schedule may be reduced so that, at all levels of pre-tax income wl, disposable income is ten per cent higher.) Subsequent work extends this uniformity result to examine cases in which the income tax need not be optimal and to assess various partial reforms, one result being that any proportionate reduction in non-uniform commodity taxes can generate a Pareto improvement (see Kaplow 2006; also Konishi 1995; Laroque 2005).

The intuition behind the uniformity result is that, despite the second-best setting (due to the inherently distortionary character of a redistributive labour income tax), there is nothing to be gained – except distortion of consumption – by differentiating commodity taxes when the utility function is weakly separable in labour. When that assumption is relaxed, one has the qualification – due originally to Corlett and Hague (1953) in a Ramsey setting – that complements to leisure (labour) should be taxed (subsidized). For example, taxing beach attendance or the purchase of novels may make leisure less attractive, encouraging labour effort and thereby reducing the distortion due to the income tax. Other qualifications, including with regard to preferences that depend on ability, other preference heterogeneity, and administrative and enforcement concerns, are catalogued in Kaplow (2008).

The Ramsey Problem: Commodity Taxation Alone

The foregoing analysis is usefully contrasted with that of Ramsey (1927), who considered how to set commodity taxes on a population of identical individuals to meet a revenue requirement. The familiar result is that commodity taxes should be inversely proportional to the elasticity of demand, with refinements for demand interdependencies. Introducing nonidentical individuals leads to modifications reflecting distributive concerns that entail higher taxes than otherwise on luxuries and lower taxes on necessities. See generally Atkinson and Stiglitz (1976, 1980), Auerbach and Hines (2002), Salanié (2003), and Stiglitz (1987).

As initially emphasized in Atkinson and Stiglitz (1976) and elaborated in Stiglitz (1987) and Kaplow (2008), however, neither prescription is apt if there is also an income tax. In the original Ramsey model in which all individuals are identical and thus there are no distributive concerns,

the optimal tax obviously would be a uniform lump-sum extraction (a limiting case of an income tax), which, it should be noted, neither requires information about individuals' types nor is distributively objectionable in this setting. When differences in earning ability are admitted, the optimal tax is a nonlinear income tax, and in typical cases the lumpsum component involves a uniform lump-sum subsidy. Nevertheless, optimal commodity taxation still is not guided either by the familiar inverse-elasticity rule or by the general preference for harsher treatment of luxuries than of necessities. As noted, in the basic case optimal differentiation is nil regardless of the demand elasticity or how demand changes with income, and qualifications such as that favouring taxation (subsidization) of leisure complements (substitutes) are largely unrelated to the level of the own-elasticity of demand for a commodity or its income elasticity.

Applications

Optimal commodity taxation is, in an important sense, a building block for the analysis of many other important problems. For example, Atkinson and Stiglitz (1976) explain how the analysis of optimal capital taxation can be assimilated into the framework, for it involves nonuniform taxation of consumption in different time periods, which may be interpreted in terms of the model simply as differently indexed commodities. Hence, in the basic case, the optimal tax on capital is zero.

Furthermore, as discussed by Kaplow (2004, 2008), other types of government policy may be analysed in a similar fashion. Allowing for externalities, the no-differential-tax prescription may be interpreted as requiring that consumer price ratios equal not producer price ratios but instead ratios of full social costs; hence, first-best Pigouvian taxes and subsidies (that is, set equal to marginal external effects) are optimal despite second-best concerns about distortionary income taxation and distributive effects. For public goods, the analogy to differential taxation is a departure from the pure Samuelson rule, so in the basic case, that cost-benefit test also does not require modification on account of income tax distortions and distributive concerns. Likewise, deviations from marginal cost pricing of public production is counter-indicated.

By contrast, much prior and ongoing work examines these problems and others in a Ramsey-like setting. As Stiglitz (1987) observes, this course may be appropriate for developing economies in which income taxation is largely infeasible, but not for developed economies with an income tax.

Optimal Tax Systems

Most optimal taxation analysis simply assumes that certain tax instruments are available and others are not. Mirrlees (1971), Atkinson and Stiglitz (1980), and Slemrod (1990), however, emphasize the importance of motivating the presumed set of available instruments by administrative and enforcement concerns that indicate what actually is feasible. Ideally, these concerns would not be stipulated but rather would be made endogenous. Often, feasibility is a matter of degree, and one must choose among various imperfect systems, the quality of each being determined by policy choices regarding administration and enforcement and also by how the instrument is used.

To illustrate these trade-offs, note that a nonlinear income tax may be a more fine-tuned redistributive instrument than a linear income tax but is subject to additional types of manipulations that are costly to regulate. Likewise, if nonuniform commodity taxation is employed, there exist incentives to reclassify commodities. More comprehensive tax bases may avoid unnecessary distortions but be more costly to administer. The extent of evasion under any system may depend on the level of tax rates and on what other taxes are in place.

Greater attention to the choice among tax systems seems warranted. Whether or not to have a 20 per cent VAT, relying far less on income taxes, is probably a more important decision than how to set commodity tax differentials in light of subtle qualifications to the uniformity result. System choices are likely to be particularly important for developing countries, where fewer options are

feasible and the available instruments are changing over time and in ways that are influenced by other government policies.

See Also

- ▶ Income Taxation and Optimal Policies
- ▶ Redistribution of Income and Wealth
- ▶ Social Welfare Function
- ▶ Taxation of Income

Bibliography

Akerlof, G.A. 1978. The economics of 'tagging' as applied to the optimal income tax, welfare programs, and manpower planning. *American Economic Review* 68: 8–19.

Atkinson, A.B.., and J.E. Stiglitz. 1976. The design of tax structure: Direct versus indirect taxation. *Journal of Public Economics* 6: 55–75.

Atkinson, A.B.., and J.E. Stiglitz. 1980. *Lectures on public economics*. New York: McGraw-Hill.

Auerbach, A.J., and J.R. Hines. 2002. Taxation and economic efficiency. In *Handbook of public economics*, ed. A.J. Auerbach and M. Feldstein, vol. 3. Amsterdam: North-Holland.

Corlett, W.J., and D.C. Hague. 1953. Complementarity and the excess burden of taxation. *Review of Economic Studies* 21: 21–30.

Dahan, M., and M. Strawczynski. 2000. Optimal income taxation: An example with a U-shaped pattern of optimal marginal tax rates: Comment. *American Economic Review* 90: 681–686.

Diamond, P.A. 1998. Optimal income taxation: An example with a U-shaped pattern of optimal marginal tax rates. *American Economic Review* 88: 83–95.

Kanbur, R., and M. Tuomala. 1994. Inherent inequality and the optimal graduation of marginal tax rates. *Scandinavian Journal of Economics* 96: 275–282.

Kaplow, L. 2004. On the (ir)relevance of distribution and labor supply distortion to government policy. *Journal of Economic Perspectives* 18 (4): 59–75.

Kaplow, L. 2006. On the undesirability of commodity taxation even when income taxation is not optimal. *Journal of Public Economics* 90: 1235–1250.

Kaplow, L. 2008. *The theory of taxation and public economics*. Princeton: Princeton University Press.

Konishi, H. 1995. A Pareto-improving commodity tax reform under a smooth nonlinear income tax. *Journal of Public Economics* 56: 413–446.

Laroque, G. 2005. Indirect taxation is harmful under separability and taste homogeneity: A simple proof. *Economics Letters* 87: 141–144.

Mirrlees, J.A. 1971. An exploration in the theory of optimum income taxation. *Review of Economic Studies* 68: 175–208.

Ramsey, F.P. 1927. A contribution to the theory of taxation. *Economic Journal* 37: 41–61.

Saez, E. 2001. Using elasticities to derive optimal income tax rates. *Review of Economic Studies* 68: 205–229.

Salanié, B. 2003. *The economics of taxation*. Cambridge, MA: MIT Press.

Slemrod, J. 1990. Optimal taxation and optimal tax systems. *Journal of Economic Perspectives* 4 (1): 157–178.

Slemrod, J., S. Yitzhaki, J. Mayshar, and M. Lundholm. 1994. The optimal two- bracket linear income tax. *Journal of Public Economics* 53: 269–290.

Stern, N.H. 1976. On the specification of models of optimum income taxation. *Journal of Public Economics* 6: 123–162.

Stern, N.H. 1982. Optimum taxation with errors in administration. *Journal of Public Economics* 17: 181–211.

Stiglitz, J.E. 1982. Self-selection and Pareto efficient taxation. *Journal of Public Economics* 17: 213–240.

Stiglitz, J.E. 1987. Pareto efficient and optimal taxation and the new new welfare economics. In *Handbook of public economics*, ed. A.J. Auerbach and M. Feldstein, vol. 2. Amsterdam: North-Holland.

Tuomala, M. 1990. *Optimal income tax and redistribution*. Oxford: Clarendon Press.

Optimality and Efficiency

Peter Newman

Keywords

Allais, M.; Competitive equilibrium; Completness; Convexity; Core; Compensated core; Compensated equilibrium; Efficiency; Optimality and efficiency; Pareto efficiency; Quasi-equilibrium; Transitivity

JEL Classifications

D0

An exchange economy consists of a group of people, each of whom has preferences concerning what commodities he or she likes, and initial holdings of the various commodities available. Operating under whatever institutional rules permit freedom of contract, the society redistributes the initial holdings among itself so as to achieve a

distribution that is in some sense a *solution* to the exchange problem.

But in what sense? Over the years three common meanings of solution have emerged, each with ever greater clarity. In order of increasing structural content rather than historical origin they are: (*a*) optimality in the sense of Edgeworth (1881) and Pareto (1909), or for brevity *EP-optimality;* (*b*) core solutions, which originated wholly with Edgeworth (1881) but had to wait until the advent of game theory before they were properly understood; and (*c*) competitive equilibria, which owe most to Walras (1874). Diverse as they are these three concepts are linked by a common thread, that each agent's objective is to seek the greatest satisfaction possible within the constraints that bind him.

If the roles of objectives and constraints are interchanged in (*a*), (*b*) and (*c*) we obtain three new concepts of solution, which are in effect mirror images of the earlier ideas. Thus corresponding to EP-optimality there is $(a)'$ efficiency in the sense of Allais (1943, pp. 610–16, 637–44) and Scitovsky (1942), or in brief *AS-efficiency.* Corresponding to the core is the idea $(b)'$ of a *compensated core* (for which see below), and corresponding to competitive equilibria is the concept $(c)'$ of *compensated equilibria* due to Arrow and Hahn (1971, p. 108), although the closely related *quasi-equilibria* were defined earlier by Debreu (1962).

(Curiously enough, the passage in Scitovsky (1941–2) that gives his definition of $(a)'$ was omitted from the reprinted version in his collected essays (1964). Very clear accounts of his approach may however be found in Samuelson (1956) and Graaff (1957), while Allais has published many further elaborations of his ideas, for example in Allais (1978). Those ideas were clearly at work in the pioneering paper by Debreu (1951), where he used them to overcome the problem 'that no meaningful metrics exists in the satisfaction space, [that is, that utility is not cardinally measurable]' (1951, p. 273). Later, Debreu's proof of the Second Fundamental Theorem of Welfare Economics (the 'pricing-out' of EP-optima) in Chap. 6 of (1959) also depended quite explicitly on the use of AS-efficiency.)

The interrelations between competitive and compensated equilibria are well recognized (see for example cost minimization and utility maximization) and concern such matters as the existence of locally cheaper points, which in turn necessarily imply market valuations of commodity bundles. The analogous interrelations between EP-optimality and AS-efficiency, and between cores and compensated cores, do not involve market phenomena and perhaps as a consequence are not so well known.

Preliminaries

We need appropriate language and notation, and some general assumptions. The exchange economy consists of m agents, indexed by h, and n goods, indexed by i. Each agent has a preference relation $\precsim_k$ that is defined over some subset of the non-negative orthant of R^n and whose meaning is 'at least as good as'. It is assumed to be composed of two disjoint sub-relations, strict preferences $\succ_k$ and indifference $\sim_k$. Completeness and convexity of preferences are never assumed, and only partial transitivity is required, in the sense that the two cases $\left(z_k^1 \succ_k z_k^2 \text{ and } z_k^2 \succ_k z_k^3\right)$ and $\left(z_k^1 \succ_k z_k^2 \text{ and } z_k^2 \sim_k z_k^3\right)$ are each assumed to lead to the conclusion $z_k^1 \succ_k z_k^3$. In particular, $\sim_k$ need not be transitive.

A *distribution* (or *allocation*) Z is any $m \times n$ matrix of the individual holdings z_{hi}, so that Z^0 is the distribution of initial holdings z_{hi}^0 or the *endowment.* If two distributions Z^1 and Z^2 are such that $z_k^1 \succ_k z_k^2$ for every agent h, then we write $Z^1 \succ\succ Z^2$ and say that Z^i is *better* than Z^2. Similarly, $Z^1 \succsim Z^2$ means that $z_k^1 \succ_k z_k^2$ for every h; we say that Z^1 *meets* Z^2. If Z^1 meets Z^2 and the number of agents k for whom $z_k^1 \succ_k z_k^2$ is at least 1 and not m, then we write $Z^1 \succ Z^2$.

An agent's holdings are written $z_h = (z_{h1}, z_{h2}, \ldots, z_{hn})$. The symbolic expression ΣZ means the commodity vector $z = \Sigma z_h$ (*all* summations here are over the index h for agents). In particular, $\Sigma Z^0 = z^0$, the vector of total endowments of each good; $\mathbf{O}$ is the vector with zero amounts of every good. $Z^1 \prec\prec Z^2$ means $(z^2 - z^1) \prec\prec \mathbf{O}$, that is, ΣZ^1 is less in every component (good) than ΣZ^2; Z^1 is then

less than Z^2. Similarly, $Z^1 \leq Z^2$ means that ΣZ^1 is not greater than ΣZ^2 in any component. If $\Sigma Z^1 \leq \Sigma Z^2$ and the number of goods j for which $\sum z^1_{hj} = \sum z^2_{hj}$ is at least 1 and not n, we write $Z^1 < Z^2$.

The notation $Z^1 \succ\succ Z^2$ means $(z^1 - z^2) \prec\prec \mathbf{O}$. In an exchange economy it is natural to assume $z^0 \succ\succ \mathbf{O}$. A distribution Z is *feasible* if $\Sigma Z = z^0$, the use of $=$ here rather than $\leq$ implying that free disposal is not assumed; quantities have to be conserved during exchange.

The assumptions made in this section will be maintained throughout.

EP-Optimality and AS-Efficiency

Purely for simplicity the definition of EP-optimality given by Arrow–Hahn (1971, p. 91) is used here, generalized to allow for incompleteness of preferences but specialized to an exchange economy. For compactness, that exchange economy will always be denoted $E(\succsim_k, Z^0)$.

Definition 1 (D1) A distribution Z^1 is *EP-optimal* for $E(\succsim_k, Z^0)$ if: (*a*) it is feasible; and (*b*) there is no other feasible distribution Z that is better than Z^1.

Notice that D1 depends only on the totals z^0 and not on their distribution Z^0. Applying a weaker meaning of being better, namely: that $Z \succ Z^1$, produces a smaller set of allocations that can withstand such tests, the *strongly EP-optimal allocations*. Proofs of the interrelations between this more usual type of EP-optimality and the strong AS-efficiency defined below are similar to those given here, but with more complication, much as the theory of non-negative matrices is basically similar to but more complicated than the theory of positive matrices.

The following definition of AS-efficiency is implicit in the original works of Allais and Scitovsky.

Definition 2 (D2) A distribution Z^2 is *AS-efficient* for $E(\succsim_k, Z^0)$ if: (*c*) it is feasible; and (*d*) there is no other distribution Z which meets Z^2 and is less than Z^0.

Again, D2 depends only on z^0 and not on Z^0, since 'less than Z^0' actually involves only z^0. As before, a weaker meaning of being less, namely: that $Z \prec Z^2$, produces a smaller set of allocations that can withstand such tests, the *strongly AS-efficient allocations*.

The first special assumption asserts a kind of monotonicity of preference for the society considered as a whole.

Assumption 1 (A1) For any Z and any commodity vector $s \succ\succ \mathbf{O}$ there exists Z^s such that $\Sigma Z^s = \Sigma Z + s$ and $Z^s \succ\succ Z$.

Theorem 1 Assume A1. If Z^1 is EP-optimal then it is AS-efficient.

Proof This and all other proofs are by contraposition. If Z^1 is not AS-efficient, there exists Z such that $Z \succsim Z^1$ and $\Sigma Z \prec\prec Z^0$. So there is a vector of surpluses in every commodity, i.e., $s = (z^0 - \Sigma Z) \succ\succ \mathbf{O}$. Hence from A1 there is Z^s such that $\Sigma Z^s = \Sigma Z + s = z^0$ and $Z^s \succ\succ Z$. But then $Z^s \succ\succ Z \succsim Z^1$ implies $Z^s \succ\succ Z^1$, and Z^s is feasible. So Z^1 is not EP-optimal.

The second special assumption does not involve the topology of R^n but nevertheless plays the role of a continuity condition on preferences.

Assumption 2 (A2) For any agent h, $z^1_h \succ_n z^2_h$ implies the existence of $\mu_h \in (0, 1)$ such that $\lambda z^1_k \pm_k z^2_k$ for all $\lambda \in [\mu_h, 1)$.

Theorem 2 Assume A2. If Z^2 is AS-efficient then it is EP-optimal

Proof If not, there exists Z such that $\Sigma Z = z^0$ and $Z \succ\succ Z^2$. From A2 there exists for each z_h in Z some $\mu_h \in (0, 1)$ such that $\lambda z_k \succcurlyeq_k z^2_k$ for all $\lambda \in [\mu_h, 1)$. Put μ equal to the maximum of these μ_h, so that $\mu < 1$ and write μZ for the $m \times n$ matrix of the μz_h. Then $\mu Z \succsim Z^2$. But by construction and the fact that $z^0 \succ\succ \mathbf{O}$, $\Sigma \mu Z = \mu \Sigma Z \prec\prec \Sigma Z = z^0$. So Z^2 is not AS-efficient.

Cores and Compensated Cores

The language and notation of section "Preliminaries" need modification to cope with cores.

A *coalition* C is any non-empty subset of the m agents in the economy, and $|C|$ denotes its cardinality, so that $1 \leq |C| \leq m$. A *distribution over* C is the $|C| \times n$ matrix Z_c whose rows are the n-vectors z_k for $k \in C$. The notation ΣZ_c means the sum over the $|C|$ rows of Z_c, and for any Z^0 and any C we write $z_c^0 = \Sigma Z_c^0$, the total endowments available to the coalition C. Given any distribution Z for the whole economy and any coalition C, the C-section of Z is the distribution Z_c over C.

The notion and language of section "Preliminaries" for preferential and quantitative relations between distributions will be applied freely to C-sections. But rather than writing $Z_c^1 \succ\succ {}_cZ_c^3$ and $Z_0^2 \leq {}_cZ_c^4$ etc., the simpler notation $Z_c^1 \succ\succ Z_c^3$ and $Z_0^2 \leq Z_c^4$ will be used.

Just as in section "EP-Optimality and AS-Efficiency," stronger concepts of core and compensated core could be defined and corresponding results proved for them; but that is not done here.

Definition 3 (D3) A distribution Z^1 is in the *core* of $E(\succsim_k, Z^0)$ if: (i) it is feasible; and (ii) there is no coalition C and no distribution Z_c over C such that (*a*) $\sum Z_z = z_c^0$ and (*b*) Z_c is better than the C-section Z_c^1 of Z'.

D1 is the special case of D3 in which the only coalition allowed is the whole society, and similarly for D2 and D4, which is given next.

Definition 4 (D4) A distribution Z^2 is in the *compensated core* of Z_c if: (iii) it is feasible; and (iv) there is no coalition C and no distribution Z_c over C such that (c) Z_c meets Z_c^2 and (d) $\sum Z_c \prec\prec z_c^0$.

The rationale for D4 is clearly similar to that for D2. Equally clearly, it is the appropriate 'mirrored' version of the core, in which objectives and constraints are interchanged. For example, it is easy to show that any compensated equilibrium of $E(\succsim_k, Z^0)$ is in its compensated core. A much deeper result, for an exchange economy with a continuum of agents, is a ' compensated' version of the core equivalence theorem of Aumann (1964), namely: the set of the compensated equilibria is precisely the compensated core (Newman 1982). Moreover, the assumptions and proof needed for this result are significantly simpler than in the classic paper of Aumann; in particular, only a non-topological separating hyperplane theorem is needed.

Since a coalition can be of any size, from one agent to every agent, the monotonicity of preference for the society as a whole asserted by A1 is quite inadequate to prove interrelations between cores and compensated cores. Instead, we use a more standard monotonicity assumption:

Assumption 3 (A3) For any agent h, $z_h^1 \succ\succ z_h^2$ implies $z_h^1 \succ_h z_h^2$.

Theorem 3 Assume A3. If Z^1 is in the core of $E(\succsim_k, Z^0)$ then it is in its compensated core.

Proof If not there is a coalition C and a distribution Z_c over C such that $Z_c \succsim Z_c^1$ and $\sum Z_c \prec\prec z_c^0$. So for C there is a vector s_c of surpluses in every commodity, that is, $s_c = \left(z_c^0 - \sum Z_c^3\right) \succ\succ \mathbf{O}$.

Now form a new distribution Z_c^s over C by adding the vector $(|C|^{-1})\ s \succ\succ \mathbf{O}$ to each z_k for $k \in C$, and denote the result by z_k^s. Since $z_k^s \succ\succ z_k$, from A3 $z_k^s \succ {}_kZ_k$. Then $Z_c^s \succ\succ Z_c \succsim Z_c^1$ so that $Z_c^s \succ\succ Z_c^1$. Moreover, by construction, $\sum Z_c^s = z_c^0$. Hence Z^1 is not in the core.

Theorem 4 Assume A2. If Z^2 is in the compensated core of $E(\succsim_k, Z^0)$ then it is in its core.

Proof If not, there exists a coalition C and a distribution Z_c over C such that $\sum Z_c = z_c^0$ and $Z_c \succ\succ Z_c^2$ The proof then proceeds as in Theorem 2.

Conclusion

There is remarkable symmetry between the solution concepts (*a*) and (*b*) on the one hand, and (a)′ and (b)′ on the other. However, there is a major asymmetry. The concepts (*a*) and (*b*) implicitly give each member of the society a positive weight, that is, each person 'counts' for something. Hence, as Edgeworth first observed (1881, p. 23), it is easy to show that a distribution is

(strongly) EP-optimal if and only if it maximizes the satisfaction of any agent picked at random, *given* both the total endowments and the levels of satisfaction of the remaining (m − 1) agents.

The corresponding statement for strong AS-efficiency is not so obvious. Suppose (and this is Scitovsky's original argument) that we fix the levels of satisfaction of everyone in the society and the total amounts of all but one commodity chosen at random, say z_i. Then it is tempting to say that a distribution is AS-efficient if and only if it minimizes the usage of z_i. The trouble with this is that, in the situation prevailing z_i just might be a commodity that nobody wants. So it is not scarce, its shadow price is zero, and there is no point in trying to economize on its use. Unlike the case with persons, we cannot be sure that a commodity chosen at random will carry positive weight.

The obvious way of dealing with this point is to put sufficient structure on the problem to make z_i always desired. But then it ceases to be an arbitrary commodity, unless all commodities are always so desired; and that is a strong assumption indeed.

Exactly the same difficulty arises of course with efficient production programmes, if they are defined as allocations that maximize the output of an arbitrary product y_j given the supplies of all the factors and the quantities of all products other than y_j. This is really not surprising, since such 'Pareto-efficiency' is the analogue in a production economy of AS-efficiency in an exchange economy.

Bibliography

Allais, M. 1943. *A la recherche d'une discipline économique. Première partie, l'économie pure*. Paris: Ateliers Industria.

Allais, M. 1978. Theories of general economic equilibrium and maximum economic efficiency. In *Equilibrium and disequilibrium in economic theory*, ed. G. Schwodiauer. Dordrecht: D. Reidel.

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

Aumann, R.J. 1964. Markets with a continuum of traders. *Econometrica* 22: 39–50.

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

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

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

Edgeworth, F.Y. 1881. *Mathematical psychics*. London: Kegan Paul.

Graaff, J. de V. 1957. *Theoretical welfare economics*. Cambridge: Cambridge University Press.

Newman, P. 1982. Compensated cores and the equivalence theorem. Working papers in economics no. 112, Department of Political Economy, Johns Hopkins University.

Pareto, V. 1909. *Manuel d'économie politique*. Paris: Giard.

Samuelson, P.A. 1956. Social indifference curves. *Quarterly Journal of Economics* 70: 1–22.

Scitovsky, T. 1942. A reconsideration of the theory of tariffs. *Review of Economics Studies* 9: 89–110.

Scitovsky, T. 1964. *Papers on welfare and growth*. Stanford: Stanford University Press.

Walras, L. 1874. *Éléments d'économie politique pure*. Lausanne: Corbaz.

Optimism and Pessimism

F. C. Montague

The term optimism is difficult to define. Strictly it should signify the belief that everything which exists is the best possible. But as there is scarcely any pessimist who denies absolutely the existence of good, so there is scarcely any optimist who denies absolutely the existence of evil. Optimism therefore can describe only the belief that good greatly preponderates in the world, or that evil admits of being resolved ultimately into good. Such a belief may be the result either of temperament or of a process of logical inference. In so far as it is the result of a happy temperament, it cannot be communicated to those whose disposition is less cheerful. In so far as it is the result of logical inference it may take various forms. All who regard the universe as the work of reason, in other words, all theists, must be optimists in one sense or another. But among theists even within the bounds of the Christian church there may be wide differences in the nature of their optimism. Some may concentrate their minds on the

corruption of man and others upon the benevolence of his Creator. St Augustine or Calvin would hardly be termed optimists in the ordinary use of that word. Paley was an optimist in every sense. Now one of the characteristics of the period in which modern political economy took its rise, the period between the close of the Thirty Years' war and the outbreak of the French Revolution, was a general optimism. Religious wars and persecutions had impressed the most active minds with indifference or disgust for the theological views which came down from the middle ages, and which were permeated with distrust of human nature and aversion to the pursuits of the world. In contrast to these views the antique conception of nature kept alive by the Roman law again attracted philosophers and became the germ of new moral and political theories. Natural religion took the place of revelation, and natural goodness of asceticism. Natural instincts were again regarded as innocent and deserving of gratification. Much stress was laid on those amiable and social instincts which find their fulfilment in promoting the happiness of others. Providence, it was held, had so ordered the world that each man in seeking to satisfy his own desires contributed to the general welfare. Virtue was identified with the rational pursuit of happiness, and thus was made to appear easy and natural. From these first principles the inference in favour of freedom was irresistible. Restraint or compulsion was in itself an evil because it was painful, and in most cases restraint or compulsion was unnecessary, since human instincts harmonized by divine wisdom tended of themselves to bring about the good of mankind.

This form of optimism pervades the discussion of education, of legislation, and of economics by the most celebrated writers of the 18th century. It is very noticeable in the writings of the physiocrats and of Adam Smith. Adam Smith cannot indeed be charged with taking too exalted a view of human nature. He assumes that men are generally employed in promoting their own interests, and he objects to any regulation that can be dispensed with, because he thinks that it is likely to be inspired by selfishness. Adam Smith's optimism lies rather in overrating the ability of the individual to perceive his interest, and in assuming a providential harmony between the self-interest of various individuals if placed in a state of legal freedom and equality. It is only after a prolonged discipline that the ordinary civilized man has attained even to his present imperfect knowledge of what is good for him, and even now the pursuit of his own welfare by each individual constantly brings him into conflict with others.

Since Adam Smith wrote upon morals and economics, optimism has been discouraged by several causes. In the first place, the French Revolution showed that the glorification of natural impulses might end in crimes and disorders as great as had ever been produced by fanaticism. In the next place, the struggle of nation with nation, and of class with class, for the last hundred years, has compelled us to see that there is no pre-established harmony between the appetites of different human beings. In the third place, the rise in the standard of comfort has produced an all but universal discontent. Mankind are probably more comfortable than in any former age, yet the difference between that which they enjoy and that to which they think themselves entitled is more noticeable than ever. Lastly, the progress of science has disturbed the cheery, old-fashioned view of nature. Malthus showed that nature has not provided an abundant subsistence for an indefinite number of persons. Darwin showed the evolution of life to have been a process of almost infinite length involving wholesale waste and destruction. Those who have adopted a formal and philosophical pessimism are few, but those who maintain the easy optimism of the 18th century are fewer. There are many who propose to make mankind happy by political or economical changes, but as a rule they propose to do this by subjecting the individual to the community. For with the old optimism the old belief in liberty has also declined in strength.

Like the term Optimism, the term pessimism is used in a variety of senses. Properly it denotes the doctrine that, in the world as a whole, evil necessarily predominates over good. But it is often used loosely to describe the mood of those who are more alive to the evil than to the good of existence. Quite apart from any philosophic theory,

differences of temperament and of circumstances will cause men to differ very widely in their estimate of life. Individual feeling admits of infinite gradations which defy classification. Pessimism and optimism in this popular use are terms of merely relative import. Pessimism as a principle has manifested itself in religious forms, notably in Buddhism, and in philosophical forms, the most modern of which are associated with the names of Schopenhauer and Hartmann. A critical examination of pessimist theories would altogether transcend the limits of this article. They have their origin in the undeniable and awful contrast between human aspiration and human attainment. No form of philosophic pessimism has at present exerted much influence on political economy. The classical economists lived in an age of optimism and were in full sympathy with their age. They had a hearty faith in the unfettered energies of mankind. It is true that the theories of certain eminent economists, as Malthus and Ricardo, have been used to demonstrate that under existing conditions the state of the mass of mankind must steadily grow worse. The inference commonly drawn, however, was not that mankind were doomed by fate to suffer, but that the actual economic system must be modified. Those who do not expect well-being to result from individual effort are confident that it can be produced by the action of the community.

The rising generation of economists may probably be less optimistic in tone. The very diffusion and intensity of the desire for comfort tend to produce a formidable discontent which may at first discharge itself upon obnoxious institutions or classes, but must finally break against the unalterable facts of nature. Certain characteristics of modern civilization, notably the resulting prolongation of the lives of the weak, both in mind and body, and the heavy burthens imposed on the capable members of society, seem likely to retard progress as hitherto understood. The limits to the physical resources of our globe are becoming more apparent. Nearly the whole of its surface has been explored; the area which civilized man can occupy has been pretty well ascertained; the great forests are disappearing, the virgin soils are losing their spontaneous fertility, and mines are worked upon a scale which in many cases threatens exhaustion in no distant future. The assumption that mankind are destined to a practically infinite economic development is thus shaken. The economists of a past age were chiefly concerned with the advantages which would follow the destruction of artificial barriers; but the stringency of natural limitations which cannot be removed will probably attract more attention from the economists of the approaching time.

The change in the tone of economic literature can be realized by comparing Smith's *Wealth of Nations* with J.S. Mill's *Principles of Political Economy*. Leslie Stephen, *English Thought in the Eighteenth Century*; Bonar, *Philosophy and Political Economy*; Ritchie, *Natural Law*, may be consulted for information respecting the philosophical optimism of the 18th century.

Bibliography

Bonar, J. 1893. *Philosophy and political economy*. London: S. Sonnenschein & Co.
Mill, J.S. 1848. *Principles of political economy*. London: J.W. Parker.
Ritchie, D.G. 1890. *Natural law*. London.
Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*. London: W. Strahan & T. Cadell.
Stephen, L. 1876. *History of English thought in the eighteenth century*. London: Smith, Elder.

Optimum Currency Areas

Masahiro Kawai

An optimum currency area refers to the 'optimum' geographical domain having as a general means of payments either a single common currency or several currencies whose exchange values are immutably pegged to one another with unlimited convertibility for both current and capital transactions, but whose exchange rates fluctuate in unison against the rest of the world. 'Optimum' is defined in terms of the macroeconomic goal of maintaining internal and external balance. Internal

balance is achieved at the optimal tradeoff point between inflation and unemployment (if such a tradeoff really exists), and external balance involves both intra-area and inter-area balance of payments equilibrium.

The concept of optimum currency areas was developed in a context of the debate over the relative merits of fixed versus flexible exchange rates. Proponents of flexible exchange rates, such as Milton Friedman (1953), had argued that a country afflicted with price and wage rigidities should adopt flexible exchange rates in order to maintain both internal and external balance. Under fixed exchange rates with price and wage rigidities, any policy effort to correct international payments imbalances would produce unemployment or inflation, whereas under flexible exchange rates the induced changes in the terms of trade and real wages would eliminate payments imbalances without much of the burden of real adjustments. Such an argument in favour of flexible exchange rates left the general impression that any country must adopt flexible exchange rates irrespectively of its economic characteristics. However, countries differ in many ways. The theory of optimum currency areas claims that if a country is highly integrated with the outside world in financial transactions, factor mobility or commodity trading, fixed exchange rates may reconcile internal and external balance more efficiently than flexible exchange rates.

The pioneering work by Mundell (1961) and McKinnon (1963) (in addition to Ingram 1962), attempted to single out the most crucial economic properties to define an 'optimum' currency area. The subsequent work by Grubel (1970), Corden (1972), Ishiyama (1975) and Tower and Willet (1976) turned their attention to evaluating the benefits and costs of participating in a currency area. Hamada (1985) studied the welfare implications of individuals countries' participation decisions.

Properties of an Optimum Currency Area

Price and Wage Flexibility

Price and wage flexibility, or lack thereof, was the central issue in the debate over fixed versus flexible exchange rates. Indeed, the assumed price–wage inflexibility was the basis for Friedman's argument in favour of flexible exchange rates and the later development of the optimum currency area literature. (It is appropriate to point out, however, that Friedman did not entirely dismiss the idea that a group of countries, such as the sterling area, may fix their exchange rates with one another and let the rates fluctuate jointly against the rest of the world; Friedman 1953, p. 193.)

Consider an area which is made up of a group of regions (or countries), however they may be defined. Then it can be postulated that, if prices and (real) wages are flexible throughout the area in response to the changed conditions of demand and supply, the regions in the area should be tied together by fixed exchange rates. Complete flexibility of prices and wages would achieve market clearance everywhere and facilitate instantaneous real adjustments to disturbances affecting inter-regional payments without causing unemployment. The ultimate, real adjustment consists of 'a change in the allocation of productive resources and in the composition of the goods available for consumption and investment' (Friedman 1953, p. 182). The required changes in relative prices and real wages accomplish such adjustment, so that inter-regional (i.e., intra-area) exchange rate flexibility becomes unnecessary. Connecting the regions by fixed exchange rates is beneficial to the area as a whole, because it enhances the usefulness of money (see section "Benefits and Costs of Currency Area Participation"). External payments balance is maintained by the joint floating of the area's currencies against the outside world as well as by internal price–wage flexibility.

When prices and real wages are inflexible, however, the transition towards ultimate adjustment may be associated with unemployment in one region and/or inflation in another. In such an economy, exchange rate flexibility among the regions, as well as its substitutes, may partially assume the role of price–wage flexibility in the process of real adjustments to disturbances. The following measures of internal market integration have been proposed as substitutes for exchange rate flexibility so as to warrant the establishment of a currency area.

Financial Market Integration

Ingram (1962) noted the smooth way in which a high degree of internal financial integration financed inter-regional payments imbalances and eased the adjustment process within the United States, or as between the United States and Puerto Rico. This suggests that a successful currency area must be tightly integrated in financial trading.

When an inter-regional payments deficit is caused by a temporary, reversible disturbance, capital flows can be a cushion to make the real adjustment smaller or even unnecessary. When the deficit is caused by a persistent and irreversible disturbance, though financial capital flows (apart from those induced by differentials in long-run real rates of return) cannot sustain the deficit indefinitely, real adjustment is allowed to be spread out over a longer period of time. The cost of adjustment is reduced by the additional help from price–wage flexibility and internal factor mobility both of which tend to be higher in the longer run. Also financial transactions strengthen the long-term adjustment process through a different channel, i.e., wealth effects. The surplus region accumulating net claims raises expenditures and the deficit region decumulating net claims lowers them, thereby contributing to real adjustment.

Thus, financial market integration lessens the need for inter-regional (i.e., intra-area) terms-of-trade changes via exchange rate fluctuations, at least in the short run. Considering the undesirable effects that exchange rate flexibility and the associated exchange risk may have, i.e., drawing a sharp line of demarcation between 'local' and 'generalized' financial claims (Ingram 1962, p. 118) and thus separating regional financial markets, fixed exchange rates are preferred within the financially integrated area.

Factor Market Integration

Mundell (1961) argued that an optimum currency area is defined by internal factor mobility (including both inter-regional and inter-industry mobility) and external factor immobility. Internal mobility of factors of production can moderate the pressure to alter real factor prices in response to disturbances affecting demand and supply; hence the need for exchange rate variations as an instrument of real factor price change is mitigated. In this sense factor mobility is a partial substitute for price-wage flexibility, partial because factor mobility is usually low in the short run. Therefore, it is more effective in easing the cost of long-run real adjustment to persistent payments imbalances than short-run adjustment to temporary imbalances, which is minimized by financial capital mobility.

Thus, factor market integration enables the fixed exchange rate system not to interfere with the maintenance of inter-regional payments balance, while increasing the usefulness of money inside the currency area. Internal balance (the optimum inflation–unemployment tradeoff) can be secured by monetary and fiscal policy, and external balance relative to the rest of the world is achieved by the joint floating of the exchange rates.

Goods Market Integration

The apparent relative smoothness of longer-run inter-regional adjustment within the United States is often attributed to its internal openness. This suggests that a successful currency area must have a high degree of internal openness, i.e., extensive trading of products inside the area. 'Openness' for a given area is measured by such indicators as the ratio of tradable to nontradable goods in production or consumption, the ratio of exports plus imports to gross output, and the marginal propensity to import.

McKinnon (1963) raised the question whether an area with a certain degree of external openness should choose flexible exchange rates against other areas or join them to belong to a larger currency area. First, suppose the area is externally highly open so that tradables represent a large share of the goods produced and consumed. Then exchange rate flexibility vis à vis other areas is not effective in rectifying payments imbalances, because any exchange rate variation would be offset by price changes without significant impacts on the terms of trade and real wages. That is, the area is too small and open for expenditure-switching instruments to be potent, though wealth effects operate in the direction of

restoring payments equilibrium. The by-product is an unstable general price level. Instead, the area would find it beneficial to assign expenditure-reducing policy to external balance and fixed exchange rates to price stability, provided the tradable goods prices are stable in terms of the outside currency. Second, when the area is relatively closed against the rest of the world, it should peg its currency to the body of nontradable goods so as to stabilize the liquidity value of money, and assign exchange rate flexibility to external balance. Exchange rate flexibility is effective because it brings about the desired changes in the relative price of tradable goods and real wages.

Thus, the optimal monetary arrangements of an internally open, externally relatively closed economy would be to peg its currency (or currencies jointly) to the body of internally traded goods – which are viewed as nontradables from the standpoint of the outside world – for price stability, and adopt externally flexible exchange rates for external balance. Splitting such an economy into smaller regions with independently floating currencies is not desirable, nor is attaching itself to the outside world to become part of a larger currency area.

Political Integration

The analysis above demonstrates the case for a currency area when a given economy has a high degree of internal market integration for financial assets, productive resources or outputs. (Other properties such as product diversity (Kenen 1969) and similarities in tastes for inflationunemployment tradeoffs have also been proposed as 'criteria' for optimum currency areas.) It is obvious that the smooth functioning of a currency area system rests on absolute confidence in the permanent fixity of exchange rates and unlimited convertibility of member currencies inside the area. This will require close coordination of national monetary authorities and perhaps even the creation of a supranational central bank. Surrendering the national sovereignty over the conduct of monetary policy to a supranational authority involves not only an economic but political process as well. The recent experience of the European Monetary System indicates that, without commitment to reaching some form of political integration, managing a currency area as loose as EMS would not be easy. (EMS is a loose currency area or a 'pseudo-exchange-rate union' (Corden 1972) because occasional currency realignments are allowed.)

Benefits and Costs of Currency Area Participation

For a complete welfare analysis of optimum currency areas, one would, ideally, like to examine how the entire world economy should be divided into independent currency areas to maximize global welfare. But constructing a general analytical framework for such a task is almost impossible. Thus, cost-benefit analysts such as Ishiyama (1975) and Tower and Willet (1976) focused on the more restricted question whether particular countries should join with one another to form a currency area. Each country is assumed to evaluate the benefits and costs of currency area participation from a purely nationalistic point of view. The price of such a restricted approach is that a 'nationally' optimum currency area thus determined may not coincide with the 'globally' optimum currency area.

Benefits

The single most important benefit a country may derive from currency area participation is that the usefulness of money is enhanced (Mundell 1961; McKinnon 1963: Kindleberger 1972; Tower and Willet 1976). Money is a social contrivance which simplifies economic calculation and accounting, economizes on acquiring and using information for transactions, and promotes the integration of markets. The use of a single common currency (or currencies rigidly pegged to one another with full convertibility) would eliminate the risk of future exchange rate fluctuations, maximize the gains from trade and specialization and, thus, enhance allocative efficiency. The usefulness of money generally rises with the size of the domain over which it is used. Money is inherently a public good.

Related to the above benefit is the fact that externalities are provided in several forms. First, currency area participation means that the participating country pegs its currency to the class of representative goods in the area. Hence, a financially unstable country can enjoy a high liquidity value of money by joining in a more financially prudent currency area. Secondly, a financially well-integrated currency area offers the domain of risk-sharing. An inter-regional payments imbalance is immediately accommodated by a flow of financial transactions, which enable the deficit country to draw on the resources of the surplus country until the adjustment cost is efficiently spread out over time. (There are other benefits arising from currency area participation, such as the reduction of official reserves and the elimination of speculative capital flows.)

Costs

The system of flexible exchange rates, in principle, allows each country to retain monetary independence. However, the system of fixed exchange rates requires unified or closely coordinated monetary policy, constraining the participating countries' freedom to pursue independent monetary policy. This loss of monetary independence is considered the major cost of currency area participation, since it may force the member countries to depart from internal balance for the sake of external balance. The cost is deemed large if the country has a low tolerance for unemployment and is subject to strong price and wage pressures from monopolistic industries, labour unions and long-term contracts. On the other hand the cost may be small if it faces a relatively vertical Phillips curve (as in the case of a small, highly open economy), because in such a case the country would not have much freedom to choose the best inflation unemployment tradeoff in the first place.

Calculus of Participation

Currency area formation is a dynamic process. In the process towards more complete monetary integration, public confidence in the system will grow, some new benefits may emerge, the existing benefits may rise, and the costs may diminish. Thus, intertemporal balancing of the benefits against the costs is necessary. It can be postulated, therefore, that an individual country will decide to participate in a currency area if the expected (discounted value of future) benefit exceeds the expected (discounted value of future) cost.

Two remarks must be made in this calculus of participation. First, the country is assumed to compare two extreme exchange rate regimes, i.e., irrevocably fixed exchange rates and freely flexible exchange rates. However, from the viewpoints of maximizing national welfare (namely benefits minus costs), there will almost always be an optimal exchange market intervention strategy that allows some exchange rate flexibility and some changes in external reserves, and the polar cases of fixed and flexible exchange rates are unlikely to be optimal – see for example Boyer (1978), Roper and Turnovsky (1980) and Aizenman and Frenkel (1985).

Second, each country chooses the best exchange rate arrangement on the assumption that its choice and policy would not affect the rest of the world, though it may condition its actions on the policies pursued by other countries. As a result, the 'optimum' currency area thus determined may not be 'globally' optimum. As is emphasized by Hamada (1985), when the important benefits of currency area formation exhibit public-good characters and externalities and the costs are borne by individual countries, the rational theory of collective action (e.g., Buchanan 1969) suggests that individual countries' participation decisions tend to produce a currency area that is smaller than is 'socially' optimum. (However, if the public-bad character of the costs dominates the public-good character of the benefits, the resulting currency area based on individual calculations may well be larger than is globally optimum.) The proposed calculus of participation obviously neglects the possible strategic interactions among countries; there is no leader–follower relationship and no cooperation. The game-theoretic approach to optimal exchange rate arrangements has recently attracted economists' attention – see Hamada (1985), Canzoneri and Gray (1985) and papers in Buiter and Marston (1985).

What Have We Learned?

Several issues have been made clear in the course of the development of the optimum currency area literature and its cost–benefit application.

First, the choice of a flexible or fixed exchange rate regime is understood as one of secondbest solutions to friction-ridden economies (Komiya 1971). If the markets for outputs, factors of production and financial assets were completely integrated on a worldwide scale, relative prices and real wages were perfectly flexible, and economic nationalism (which attempts to insulate a national economy from the rest of the world by way of artificial impediments to trade, capital flows and foreign exchange transactions) were absent, then the optimum currency area would be the whole world. In such a case, the real adjustment to payments imbalances would be extremely smooth, factor resources would be always fully employed, and the usefulness of money would be maximized. However, to the extent that the payments adjustment mechanism is impaired by market fragmentation and price–wage rigidities, a country may adopt flexible exchange rates as a secondbest policy to attain internal and external balance. The optimum currency area literature has shown that measures of market integration (for financial assets, factor resources and goods) may partially, and more effectively, substitute the required role of price–wage flexibility than does exchange rate flexibility.

Second, the cost–benefit approach to optimum currency areas based on purely national interest is limited in the analysis of designing an optimum international monetary system. Given the degree of spillover effects and economic interdependence among closely integrated countries, the strategic behaviour on the part of national policy-makers must be explicitly incorporated in order to deepen our understanding of the nature of 'globally' optimum currency areas and optimal international monetary arrangements.

As a final note it is interesting to observe that the two economists who advanced the theory of optimum currency areas, Mundell and McKinnon, now support fixed exchange rates. Mundell has been advocating a worldwide gold standard system and McKinnon (1984) a fixing of the exchange rates among three major industrialized countries (USA, West Germany and Japan). Thus they regard the world as a whole or the industrial core of western society as capable of establishing a currency area.

See Also

► International Finance

Bibliography

Aizenman, J., and J. Frenkel. 1985. Optimal wage indexation, foreign exchange intervention, and monetary policy. *American Economic Review* 75(3): 402–423.

Boyer, R.S. 1978. Optimal foreign exchange market intervention. *Journal of Political Economy* 86: 1045–1055.

Buchanan, J.M. 1969. *Cost and choice*. Chicago: Markham.

Buiter, W.H., and R.C. Marston (eds.). 1985. *International economic policy coordination*. Cambridge: Cambridge University Press.

Canzoneri, M.B., and J. Gray. 1985. Monetary policy games and the consequences of non-cooperative behavior. *International Economic Review* 36(3): 547–564.

Corden, W.M. 1972. *Monetary integration*. Essays in International Finance No. 93, April, Princeton: International Finance Section, Princeton University.

Friedman, M. 1953. The case for flexible exchange rates. In *Essays in positive economics*, ed. M. Friedman. Chicago: University of Chicago Press.

Grubel, H.G. 1970. The theory of optimum currency areas. *Canadian Journal of Economics* 3: 318–324.

Hamada, K. 1985. *The political economy of international monetary interdependence*. Cambridge, MA: MIT Press.

Ingram, J.C. 1962. *Regional payments mechanisms: The case of Puerto Rico*. Chapel Hill: University of North Carolina Press.

Ishiyama, Y. 1975. The theory of optimum currency areas: A survey. *IMF Staff Papers* 22: 344–383.

Kenen, P.B. 1969. The theory of optimum currency areas: An eclectic view. In *Monetary problems of the international economy*, ed. R.A. Mundell and A.K. Swoboda. Chicago: University of Chicago Press.

Kindleberger, C.P. 1972. The benefits of international money. *Journal of International Economics* 2: 425–442.

Komiya, R. 1971. Saitekitsukachiiki no riron (Theory of optimum currency areas). In *Gendaikeizaigaku no Tenkai (The development of contemporary economics)*, ed. M. Kaji and Y. Murakami. Tokyo: Keisoshobo.

McKinnon, R.I. 1963. Optimum currency areas. *American Economic Review* 53: 717–725.

McKinnon, R.I. 1984. *An international standard for monetary stabilization*. Policy Analyses in International Economics 8, March. Washington, DC: Institute for International Economics.
Mundell, R.A. 1961. A theory of optimum currency areas. *American Economic Review* 51: 657–665.
Roper, D.E., and S.J. Turnovsky. 1980. Optimal exchange market intervention in a simple stochastic macro model. *Canadian Journal of Economics* 13: 269–309.
Tower, E., and Willet, T.D. 1976. *The theory of optimum currency areas and exchange-rate flexibility*. Special studies in international economics No. 11, May. Princeton: Princeton International Finance Section.
Yeager, L. 1976. *International monetary relations: Theory, history, policy*, 2nd ed. New York: Harper & Row.

Optimum Population

J. D. Pitchford

Malthus (1798) had argued that improvements in living standards would almost invariably call forth such an increase in population that wages would eventually be pushed back to subsistence levels. About a hundred years later Cannan (1888), Wicksell (1910) and others were writing of an optimum population, where by implication choice of family size enabled choice of living standards. A variety of measures of birth control had come into use in the 19th century, opening the prospect of permanent escape from the trap of subsistence consumption. The early optimum concept involved a population which, at some specified time, and other things such as the capital stock being held constant, resulted in maximum output per head. Clearly it is associated with the idea of first increasing and later decreasing returns in a given region with given resources and technical knowledge.

These discussions of what population should be gave little consideration to the fact that actual population levels are reached on the basis of private choices of family size, despite the fact that private decisions in this area had been creating a revolution in demographic experience. In the 19th century Europe went through a transition from high to low mortality because of improvements in sanitation and medical science, accompanied and followed by substantial falls in fertility due to a rising age of marriage and the adoption of various methods of birth control.

Later notions of optimum population have produced a variety of specifications of the concept. Meade (1955) argued the merits of the criterion of *total utility*, that is the sum of utilities of all members of the society, rather than the utility of a representative individual implied by the maximization of output or consumption per head. Often discussions of under-or over-population have been based on military, religious or cultural factors, and in the 1970s the quality of the natural environment which the population would enjoy was raised as an important issue. As well as the total utility criterion and the utility of a representative individual, the Rawlsian criterion, requiring maximization of the utility of the worst off members of society has also been used in specifying an optimum. All these criteria involve difficult philosophical issues of the rights of potential future members of the population, which are not pursued here. Production conditions assumed have ranged from constant returns to scale to the two inputs capital and labour, to variable returns to scale depending on the size of the population, capital stock and supply of fixed resources, to the assumption of depletable natural resources.

A wide version of the concept would have it include all the population levels on a path of optimal economic growth where population is chosen at any time subject to demographic constraints, and capital is accumulated according to economic constraints. The path would maximize some social welfare function over the chosen time period which might be infinite. Such problems have been analysed using optimal control techniques such as Pontryagin's Maximum Principle and Dynamic Programming. Solving for the time paths of capital stock and various demographic variables is constrained by the fact that these techniques handle problems involving one of these variables readily, and two or more with great difficulty. Analytical insights from this literature have mainly been gained from examination of numerous versions of one, and occasionally two, dimensional models. For example, Lane

(1975) treats cases in which the total utility is the maximization criterion with constant returns to capital and labour, Pitchford (1974) uses individual utility and examines the consequences of variable returns of scale, and Koopmans (1973) Cigno (1981), and Dasgupta and Mitra (1982) treat the issues raised by exhaustible resources.

The idea of an optimum population has also been associated with the theory of the provision of public goods (see, for instance, Flatters et al. 1974). The possibility of an optimum population in this context arises because an additional worker in a region will reduce the tax burden of the provision of public goods per head, but may lower the marginal product of labour. The theory of local public goods is concerned with the optimal allocation of population amongst the different communities to which these goods are specific. Another closely related issue is the idea of the optimal size of a city. Tradeoffs can occur involving increasing returns to scale in the production of factory goods at the city centre and increasing marginal costs of transport and workers' travel time. Interesting issues are raised by the possibilities of traffic congestion and by workers' preferences about population density in residential districts.

None of these treatments of the subject has resulted in a satisfactory method of empirically computing an optimum population path or level. Few have tried the exercise. Apart from the difficulties of finding data for estimating and solving the underlying relations, there is the problem that the future state of technical knowledge must remain an unknown factor of considerable importance. An alternative approach is to ask why the population levels and growth rates, which are the outcome of individual choices, may be considered nonoptimal. Several classes of reasons can be identified. Firstly, the observer may disagree with the criteria implicitly used in private choices of family size and may wish to see society adopt a population policy based on his own criterion. An obvious example is the desire for a large population for defence reasons, but the espousal of economic criteria such as the various social welfare functions discussed above has aspects of this approach. Secondly, various governmental policies, an example is subsidised education, may be seen to be distorting private choices with respect to family size. Thirdly, individual choices may be thought to involve externalities so not achieving a social optimum. For instance, Pazner and Razin (1980) have shown for a Samuelson consumption–loan model that, if there were perfect capital markets and perfect foresight, private choices of consumption and family size, where parents have preferences for children and have their children's utility as an argument in their own utility functions, will lead to Pareto optimality. A variety of types of externalities may arise from individual choices. Thus individuals may have preferences about the density of population in their region, but this cannot affect economy-wide choices. Again, the output and income levels next period and so the welfare of the next generation will depend on the size of the population, yet parents may be unaware of or unable to calculate the effect of their own and society's current fertility choices on the future size of population. Perhaps the more illuminating way to specify population policy is as a process of recognition and remedy of the possible reasons for divergence between private and social choices regarding family size. Optimal population levels and paths then become a secondary issue, being an outcome of these efforts. Nevertheless, if the concept of an optimum population is chosen in such a way as to represent the underlying population problem, the notion and the related ideas of over- and under-population could be a useful tool for elucidating, diagnosing and treating the problem.

See Also

► Malthus's Theory of Population

Bibliography

Cannan, E. 1888. *Elementary political economy*. London: Oxford University Press.

Cigno, A. 1981. Growth with exhaustible resources and endogenous population. *Review of Economic Studies* 48(2): 281–287.

Dasgupta, P.S., and T. Mitra. 1982. On some problems in the formulation of optimum population when resources are depletable. In *Economic theory of natural resources*, ed. W. Eichhorn, R. Henn, K. Neuman, and R.W. Shephard. Wurzburg/Vienna: Physica-Verlag.
Flatters, F., V. Henderson, and P. Mieszkowski. 1974. Public goods, efficiency, and regional fiscal equalization. *Journal of Public Economics* 3(2): 99–112.
Koopmans, T.C. 1973. Some observations on 'optimal' economic growth. In *Economic structure and development*, ed. H.C. Bos, M. Linnemann, and P. de Wolff. Amsterdam: North-Holland.
Lane, J.S. 1975. A synthesis of the Ramsey–Meade problems when population is endogenous. *Review of Economic Studies* 42(1): 57–66.
Malthus, T.R. 1798. *An essay on the principle of population*, vol. I. London: John Murray.
Meade, J.E. 1955. *The theory of international economic policy*, Trade and welfare, vol. II. London: Oxford University Press.
Pazner, E.A., and A. Razin. 1980. Competitive efficiency in an overlapping-generation model with endogenous population. *Journal of Public Economics* 13(2): 249–258.
Pitchford, J.D. 1974. *Population in economic growth*. Amsterdam: North-Holland.
Wicksell, K. 1910. *Läran om Befolkningen, dess Sammansättning och Förändringar*, The theory of population, its composition and changes. Stockhlom: Albert Bonniers Forlag.

Optimum Quantity of Money

Timothy S. Fuerst

Abstract

The optimum quantity of money is a normative monetary policy conclusion drawn from the long-run properties of a theoretical model. Most famously associated with Milton Friedman, the optimum calls for a zero nominal rate of interest and thus a steady state of price deflation at the long-run real rate of interest. Although this policy prescription has played a minor role in monetary policy implementation, it has had an enormous influence in monetary theory.

Keywords

Bargaining; Deflation; Dynamic new Keynesian models; Fiat money; Friedman rule; Friedman, M.; Hold-up problem; Inflation; Monetary policy; Optimal taxation; Optimum quantity of money; Search-theoretic monetary models; Seigniorage; Transactions role of money

JEL Classifications

E31; E52

The optimum quantity of money is most famously associated with Milton Friedman (1969). The optimum is a normative policy conclusion drawn from the long-run properties of a theoretical model. Friedman posited an environment that abstracts from all exogenous shocks and nominal price and wage sluggishness. The basic logic is then straightforward. One criterion for Pareto efficiency is that the private cost of a good or service should be equated to the social cost of this good or service. The service in question is the transactions role of money. The social cost of producing fiat money is essentially zero. Since fiat money pays no interest, the private cost of using money is the nominal interest rate. Hence, one criterion for Pareto efficiency is that the nominal interest rate should equal zero. Since long-run real rates are positive, this implies that monetary policy should bring about a steady deflation in the general price level. This famous policy prescription is now commonly called the Friedman rule.

Although most closely associated with Friedman's (1969) bold statement of the policy conclusion, the basic idea of the optimum quantity can be found in Tolley (1957), who argues, on similar efficiency grounds, for paying interest on currency. Friedman (1960) credits Tolley with this suggestion, and further notes that an alternative policy would be a steady deflation. It is curious that Friedman (1960) dismisses the 'Friedman rule' deflation as not feasible for practical purposes. Finally, the optimum-quantity result is implicit, but never noted, in Bailey (1956) who

examines the welfare cost of inflation but does not consider the welfare gain of deflations.

In practice, the optimum-quantity result has had remarkably little influence on monetary policy implementation. Although many central banks pursue low inflation rates with an eventual goal of price stability, no central bank has advocated a policy that would bring about a steady price deflation. There are likely several reasons, both judgemental and theoretical, that have led to this lack of influence. I will briefly review both types of objections.

One of the first theoretical objections to the optimum-quantity results was made by Phelps (1973), who argued that Friedman's first-best argument ignored the second-best fact that money growth produces seigniorage revenues for a government, and that all forms of taxation produce distortions of some kind. If 'money' or 'liquidity' is a good like any other, then familiar optimal taxation arguments would suggest that it should be taxed via a steady inflation. This argument seems all the more persuasive given empirical estimates of a fairly low money demand elasticity.

This public finance approach spawned a very large literature. Important contributions include Kimbrough (1986), Guidotti and Vegh (1993), Correia and Teles (1996, 1999), Chari et al. (1996), and Mulligan and Sala-i-Martin (1997). These analyses were much more explicit than Friedman (1969) and considered a fully dynamic theoretical environment with no nominal rigidities. A key relationship in all these models is the transactions or shopping function. The time spent by households shopping (s_t) is a function of the form: $s_t = \phi(c_t, m_t)$, where c_t denotes real consumption and m_t denotes real cash balances. The function ϕ is assumed to be homogenous of degree k, increasing in consumption, and decreasing in real cash balances, the latter effect motivated by the transactions function of money. Money can be thought of as an intermediate good that facilitates consumption purchases. Now suppose a central government needs to finance an exogenous level of spending and can do so only with distortionary taxes on, say, labour income, or the inflation tax on money balances. In this case, is the Friedman rule still optimal?

Most of these papers were supportive of the Friedman rule, concluding that in such a second-best environment the optimal monetary policy is a zero nominal rate. Mulligan and Sala-i-Martin (1997) argued that the result was fragile as it depended on the degree of homogeneity in ϕ and the alternative tax instruments available to the government, for example, income taxes against consumption taxes. These conflicting results have been usefully explained in DeFiore and Teles (2003), who demonstrated that the reason for the divergent conclusions is an inappropriate specification of how consumption taxes are entered in the transactions cost function. They consider a more general environment in which the government has access to both consumption and income taxes. They also consider the case where money is costly to produce at a constant marginal cost of a. Further, they demonstrate that if ϕ is linearly homogenous ($k = 1$) then the optimal interest rate is equal to α. This is a modified Friedman rule in that the private cost and social cost of money are set equal to each other, and is analogous to the Diamond and Mirrlees (1971) optimal taxation result: intermediate goods should not be taxed when consumption taxes are available and the technology is constant returns to scale ($k = 1$). If ϕ is not linearly homogeneous, then the optimal policy involves a tax (or subsidy) on money proportional to α. Since money is essentially costless to produce ($\alpha = 0$) the optimal nominal interest rate is zero. DeFiore and Teles (2003) thus conclude that the Friedman rule is the optimal second-best policy for all homogeneous transactions technologies. Hence, the Phelps (1973) objection appears to be settled in Friedman's favour.

A second theoretical objection to the optimum-quantity result is that, in a world with nominal rigidities, a steady general price deflation would produce unwanted relative price movements since not all nominal prices would be adjusted simultaneously. Strictly speaking this is not a theoretical objection to Friedman (1969), as he assumed a world with perfectly flexible nominal prices and wages. But if one believes that nominal rigidities

are important, and that they matter even in the long run, then this is a relevant objection to the Friedman rule. For example, in the dynamic new Keynesian (DNK) class of models (for example, Woodford 2003) the assumed nominal rigidities have permanent effects so that any departure from price stability causes permanent movements in relative prices. Hence, these models typically suggest that optimal policy is a stable price level, and that a Friedman-rule deflation would be sub-optimal. These DNK models typically abstract from the nominal interest rate distortions that are at the heart of the optimum-quantity result.

A model that combined the DNK nominal rigidities with the nominal rate distortion would presumably result in a long-run optimal nominal interest rate somewhere between zero and the steady-state real rate.

The principle judgemental objection to the Friedman rule is historical. The instances in US history in which deflations occurred are associated with severe recessions, most famously in the 1929–1933 period. A related judgemental concern deals with the zero bound. If the central bank's principal tool to stimulate the economy is a reduction in the nominal rate of interest, then the zero nominal rate prescribed by the Friedman rule apparently leaves no additional ammunition in the monetary policy arsenal (as nominal rates cannot be negative). This nervousness about the Friedman rule was enhanced by the experience of Japan during the 1990s. The Japanese economy performed poorly at a time in which general prices were falling and the short-term nominal rate was zero.

Since central banks have not followed Friedman's (1969) proposal to set the nominal rate to zero, a natural issue is to quantify the welfare costs of being away from Friedman's optimum quantity of money. Following in the footsteps of Bailey (1956) and Lucas (2000) uses a theoretical environment similar to that of Correia and Teles (1996, 1999) to address this question. The welfare cost is approximately the area underneath the money demand curve between the optimal zero nominal rate and the interest rate under question. Lucas reports that the welfare cost of a four per cent nominal rate is between 0.2 per cent and one per cent of annual income, the difference depending upon the assumed behaviour of money demand as the nominal rate approaches zero. Since a zero nominal rate has not been observed in the United States in the post-Second World War period, the data cannot determine which estimate is more accurate. But either estimate suggests a fairly modest welfare cost.

Studies analysing the optimality of the Friedman rule have been reignited by the new class of search-theoretic monetary models. These models are micro-based, replacing the function ϕ in DeFiore and Teles (2003) with a search-based trading environment in which money improves the chances of successfully finding a suitable partner with whom to trade. In an innovative paper, Lagos and Wright (2005) use a search-theoretic environment to address the optimality of the Friedman rule and the welfare consequences of deviating from it. In search models of money the buyer and seller engage in a bargaining game to determine the transactions price at a given meeting. The buyer is carrying money and has thus postponed previous consumption. If sellers have some bargaining power, then there is a hold-up problem because part of the gain associated with the holding of money is received by the seller. This bargaining distortion leads the buyers to economize on money holdings so that they are below the socially efficient level. Lagos and Wright (2005) demonstrate that the optimal policy in this search environment is the Friedman rule (a similar conclusion is reached by Shi 1997). But more interestingly, the welfare cost of being away from the Friedman rule, at say a four per cent nominal rate, is significantly higher than calculated by Lucas (2000). This arises because the positive nominal rate exacerbates an already sub-optimal level of real balances arising from the hold-up problem.

The search models of money have rekindled interest in the optimality of the Friedman rule at just the time when DeFiore and Teles (2003) appear to have settled the issue in the aggregative monetary models. The coming years will probably see further work on the Friedman rule from this search-theoretic perspective. A key issue is

the nature of the bargaining process that arises at trading opportunities. These recent developments testify to the continued prominence of the optimum quantity of money in monetary theory, if not practice. The lasting contribution of the theory is to introduce explicit, utility-based welfare analysis into monetary economics.

See Also

- ▶ Friedman, Milton (1912–2006)
- ▶ Monetary Policy, History of
- ▶ Money and General Equilibrium
- ▶ Real Bills Doctrine

Acknowledgment The author would like to thank Charles Carlstrom and John Hoag for their helpful comments.

Bibliography

Bailey, M.J. 1956. The welfare costs of inflationary finance. *Journal of Political Economy* 64: 93–110.
Chari, V.V., L.J. Christiano, and P. Kehoe. 1996. Optimality of the Friedman rule in economies with distorting taxes. *Journal of Monetary Economics* 37: 202–223.
Correia, I., and P. Teles. 1996. Is the Friedman rule optimal when money is an intermediate good. *Journal of Monetary Economics* 38: 223–244.
Correia, I., and P. Teles. 1999. The optimal inflation tax. *Review of Economic Dynamics* 2: 325–346.
DeFiore, F., and P. Teles. 2003. The optimal mix of taxes on money, consumption and income. *Journal of Monetary Economics* 50: 871–888.
Diamond, P.A., and J.A. Mirrlees. 1971. Optimal taxation and public production. *American Economic Review* 63: 8–27.
Friedman, M. 1960. *A program for monetary stability.* New York: Fordham University Press.
Friedman, M. 1969. *The optimum quantity of money and other essays*. Chicago: Aldine.
Guidotti, P.E., and C.A. Vegh. 1993. The optimal inflation tax when money reduces transactions costs. *Journal of Monetary Economics* 31: 189–205.
Kimbrough, K.P. 1986. The optimum quantity of money rule in the theory of public finance. *Journal of Monetary Economics* 18: 277–284.
Lagos, R., and R. Wright. 2005. A unified framework for monetary theory and policy analysis. *Journal of Political Economy* 46: 463–484.
Lucas, R.E. Jr. 2000. Inflation and welfare. *Econometrica* 68: 247–274.
Mulligan, C.B., and X. Sala-i-Martin. 1997. The optimum quantity of money: Theory and evidence. *Journal of Money, Credit and Banking* 29: 687–715.
Phelps, E.S. 1973. Inflation in the theory of public finance. *Swedish Journal of Economics* 75: 37–54.
Shi, S. 1997. A divisible search model of fiat money. *Econometrica* 65: 75–102.
Tolley, G. 1957. Providing for growth of the money supply. *Journal of Political Economy* 65: 465–485.
Woodford, M. 2003. *Interest and prices*. Princeton: Princeton University Press.

Option Pricing Theory

Jonathan E. Ingersoll, Jr.

Financial contracting is as old as human history. Deeds for the sale of land have been discovered that date to before 2800 BC. The Code of Hammurabi (c1800 BC) regulated, among other things, the terms of credit. Contingent contracting was also common. Under the Code crop failure due to storm or drought served to cancel that year's interest on a land loan. The trading of the first options is probably equally ancient.

Although options have certainly been traded for centuries, it is only in recent years that they have reached any degree of importance. In 1973 the Chicago Board of Trade founded The Chicago Board Options Exchange to create a centralized market for trading call options on listed stock. The American, Pacific, and Philadelphia Stock Exchanges followed suit within a few years. In 1977 the trading of puts on these exchanges began.

By the early 1980s puts and calls could be traded on over 400 listed stocks, and options were available on many other financial instruments such as Treasury bonds and bills, foreign currencies and futures contracts. The volume of trade had grown as well. In terms of the number of shares controlled, option volume often exceeded that on the New York Stock Exchange.

Curiously the recent revolution in option pricing theory also dates to 1973 with the publication by Fischer Black and Myron Scholes of their classic paper on option valuation. In the past decade and a half, the valuation of options or various other contingent contracts has been one of the primary areas of research among financial economists.

Option contracts are examples of derivative securities; that is securities whose values depend on those of other securities or assets. For example, a call option on a share of stock gives the owner the right to purchase a share of that stock at a set price. The value of this right obviously depends on the price per share of the stock on which the option is based.

Terminology

Before discussing the academic study of options, it is useful to consider some terminology. The two most common types of option contracts are puts and calls. A *call* is an option to buy, and a *put* is an option to sell. Puts and calls are contracts between two investors. The purchaser of the option is the party to whom the contract gives certain rights or 'options'. The call's owner is said to have a *long* position. The creator or writer of the call has certain financial obligations if the owner chooses to exercise the option. The writer of an option is said to have a *short* position.

The owner of a call has the right, but not the obligation, to buy a fixed amount (usually 100 shares for exchange listed stock options) of a particular asset on or before a given date, the *maturity* or *expiration date*, upon payment of a stated fee. This fee is called the *exercise price, striking price*, or *contract price*. The owner of the call does not receive any dividends paid by the common stock or have any other rights of ownership until the option is exercised. The owner of the put has the right to sell on similar terms.

When purchasing the option, the amount that the long party pays to the short party is called the *premium*. If the stock price is above the striking price then the difference is the call's *intrinsic value*, i.e., for a call with a striking price of X on a stock with price S, the intrinsic value is Max $(S - X, 0)$. For a put the intrinsic value is the exercise price less the stock price when the former is larger, i.e., Max$(X - S, 0)$. An option's intrinsic value is sometimes called the *whenexercised* value. An option's intrinsic value does not measure its market value. Typically an option sells for more than its intrinsic value.

When options are first written the striking price is usually set near the currently prevailing stock price. The option is then said to be *at-the-money.* As the stock price changes, the option will become *in-the-money* or *out-of-the-money.* A call option is in-the-money when the stock price is above the striking price and out-of-the money when the stock price is below the striking price.

The options just described are *American* options. They can be exercised at any time on or before the expiration date. Options that can only be exercised at maturity are called *European* options. Actually this is a misnomer. While American options are traded on exchanges in the United States and Canada (and Europe), European contracts are not traded on that continent.

A *warrant* is similar to a call option. The primary difference is that a warrant is issued by a corporation against its own stock. When a warrant is exercised, the corporation issues new shares to the owner of the warrant. Warrants typically have maturities of several years or longer. There have even been a few perpetual warrants issued. When they are issued, warrants are usually substantially out-of-the-money.

A *rights issue*, like a warrant, is granted by a corporation against new stock. Usually a rights issue expires in a few weeks to a few months after it is issued. When rights are issued, they are typically substantially in-the-money.

Many other financial contracts contain implicit or explicit options. Convertible bonds, for example, give the owner the right to swap the bonds for shares of stock. This option is like a warrant. Instead of paying a cash exercise price, the bondholder relinquishes the right to the future interest and principal payments. A callable bond includes the company's right to 'repurchase' a bond at a set

call price. Much of the development in option pricing subsequent to the Black–Scholes option pricing model has been in the application of the model to these and other situations.

Preliminary Considerations

Call options are the most common and one of the simplest types of derivative assets so this discussion will be illustrated primarily with calls. Most of the general principles apply with only minor changes to any derivative asset.

A call option with an exercise price of X on a share of stock with a current price per share of S is worth $S - X$ if exercised, for it enables its owner to purchase for X something worth S. To avoid any possibility of arbitrage a call option must sell for at least this difference. In addition because a call has limited liability (that is the owner cannot be forced to exercise when it is not advantageous to do so), it must be worth at least zero. Thus, $C(S, \tau) \geq \text{Max}\,(S - X, 0)$, where $C(S, \tau)$ is the market price of a call with time to maturity of τ.

As a general rule this inequality will be strict, and the call will be worth more 'alive' than when exercised. One exception is at the time a call matures. Then the owner has only two choices–execise the option or let it expire. At this point the preceding relation must hold as an equality, $C(S, 0) = \text{Max}\,(S - X, 0)$. It is this functional relation between the value of the call at maturity and the stock price prevailing at that time that makes the call a derivative asset and allows its price to be determined as a function of the prevailing stock price.

Some general restrictions on option values can be derived with no assumptions beyond the absence of arbitrage opportunities. For example, a call with a low exercise price must be worth at least as much as an otherwise identical call with a high exercise price. The intuition is simple. The owner of the call with the low striking price could exercise whenever the owner of the other call did and would always have a lower cost of doing so. Two important restrictions of this type are Stoll's (1969) put–call parity relation and the proof that a call option on stock which pays no dividend should not be exercised prior to maturity.

The put–call parity relation holds for European puts and calls on stocks not paying dividends. It is

$$P(S, \tau) + S = C(S, \tau) + X/(1 + r)^{\tau}. \quad (1)$$

To prove this relation consider two portfolios. The first holds one share of stock and a put. The second holds a call and a zero coupon bond with a face value of X maturing on the options' expiration date. If the stock price is S_T at the expiration of the option, then the first portfolio is worth $\text{Max}(X - S_T, 0) + S_T = \text{Max}(S_T, X)$. The second is worth $\text{Max}(S_T - X, 0) + X = \text{Max}(S_T, X)$ These values are the same, and neither portfolio makes any interim disbursements. Therefore, absence of arbitrage implies that the current value of the two portfolios must be equal. Equation (1) expresses the equality of these two portfolios' current values. One importance of this relation is that once either the put or call pricing problem has been solved, the answer to the other is also known.

To prove the optimality of holding a call option until maturity consider the following two portfolios. The first holds just one share of stock. The second holds the call and a zero coupon bond with a face value of X. At expiration, the first portfolio is worth S_T. The second is worth $\text{Max}\,(S_T, X)$. As the former value is never larger, the current value of the first portfolio cannot be greater than that of the second, or

$$C(S, \tau) \geq S - X/(1 + r^{\tau}) > S - X. \quad (2)$$

This proves that an option is worth more alive than when exercised. An investor who no longer wishes to hold a call could realize more by selling the option than exercising it.

These two relations do not exhaust the general statements that can be made about option prices. Other propositions, also depending only on the absence of arbitrage, have been proved by Merton (1973) and Cox and Ross (1976b). To go beyond general propositions of this type and derive a precise value for an option, further assumptions must be made.

There were many attempts at a consistent and self-contained model of option valuation. All of these models made assumptions about the distribution of the stock's return (a lognormal distribution was the usual choice) and the absence of market frictions such as taxes, transactions costs, and short sales constraints. Most of the models included unspecified parameters which had to be measured to use the formulae.

This area of research was revolutionized with the 1973 publication of the Black–Scholes option pricing model deriving a formula depending on only five directly observable variables, the stock's price (S), the exercise price (X), the time to maturity (τ), the risk-free rate of interest (r), and the variance of changes in the logarithm of the stock price (σ^2).

Option Models Prior to Black–Scholes

Option pricing theory did not begin with the Black–Scholes model. Many economists had tackled this problem previously. While some of the attempts are flawed by current standards, later developments almost certainly would not have come about without the earlier works. There is room here only to highlight some of the more important steps leading to the Black–Scholes model.

The earliest model of option pricing was probably developed by Louis Bachelier (1900). In examining stock price fluctuations he was led to some aspects of the mathematical theory of Brownian motion five years prior to Einstein's classic paper of 1905. Postulating an absolute Brownian motion without drift and with a variance of σ^2 per unit time for the stock price process, he determined that the expected value of the call option at maturity should be

$$C = S \cdot \Phi\left(\frac{S-X}{\sigma\sqrt{\tau}}\right) - X \cdot \Phi\left(\frac{S-X}{\sigma\sqrt{\tau}}\right) + \sigma\sqrt{\tau} \cdot \phi\left(\frac{S-X}{\sigma\sqrt{\tau}}\right) \quad (3)$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the standard cumulative normal and normal density functions. In keeping with an assumption of a zero expected price change for the stock, he did not discount this expectation to find a present value. This model was rediscovered more than fifty years later by Kruizenga (1956).

By contemporary standards this model must have been very advanced. The model is only lacking in two primary areas. The use of absolute Brownian motion allows the stock price to become negative – a condition at odds with the assumption of limited liability. The assumption of a mean expected price change of zero ignores a positive time value for money, the different risk characteristics of options and the underlying stock, and risk aversion. Despite these shortcomings, the formula is actually quite good at predicting the prices of short-term calls. It fails at long maturities, however, by requiring the option price to grow proportionally to the square root of maturity.

Most of the developments in option pricing for the next half century or more were *ad hoc* econometric models. Typical of this type is the model of Kassouf (1969) who estimated call prices with the formula

$$C = X\left[\left[(S/X)^{\gamma} + 1\right]^{1/\gamma} - 1\right], 1 \le \gamma < \infty. \quad (4)$$

This formula does bound the call price above by the stock price and below by its intrinsic value, Max(-$S - X$, 0). It also gives correct maturity values for calls when the parameter γ is set to ∞. Kassouf fit his model by estimating the parameter γ using time to maturity, dividend yield, and other variables.

Major new developments in option pricing began in the 1960s. Sprenkle (1961) assumed a lognormal distribution for the stock price with a constant mean and variance (although not specifically a diffusion) and allowed for a positive drift in the stock's price. His equation for a call value can be written as

$$C = e^{\alpha\tau} S \cdot \Phi\left[\frac{\ln(S/X) + (\alpha + \frac{1}{2}\sigma^2)\tau}{\sigma\sqrt{\tau}}\right] - (1-\pi) \times X \cdot \Phi\left[\frac{\ln(S/X) + (\alpha - \frac{1}{2}\sigma^2)\tau}{\sigma\sqrt{\tau}}\right]. \quad (5)$$

The parameter π was an adjustment for the market 'price for leverage'. Sprenkle did not discount this expectation to determine the option value. (Note that if π is set to zero, (5) gives the expected terminal value for the option.)

Boness's (1964) model was very similar. He also assumed a stationary lognormal distribution for stock returns, and recognized the importance of risk premiums. For tractability he assumed that '[i]nvestors are indifferent to risk'. He used this last assumption to justify discounting the expected final option value by α, the expected rate of return on the stock. His final model was

$$C = S \cdot \Phi\left[\frac{\ln(S/X) + (\alpha + \frac{1}{2}\sigma^2)\tau}{\sigma\sqrt{\tau}}\right] - e^{-\alpha\tau} X \cdot \Phi\left[\frac{\ln(S/X) + (\alpha - \frac{1}{2}\sigma^2)\tau}{\sigma\sqrt{\tau}}\right]. \quad (6)$$

This equation is identical in form to the Black–Scholes formula described below. Its only difference is its use of α, the expected rate of return on the stock, rather than the risk-free rate of interest. If Boness had carried his assumption that investors are indifferent to risk to its logical conclusion that $\alpha = r$, he would have derived the Black–Scholes equation. Of course, his derivation would still have been based on the *assumption* of risk neutrality.

Samuelson (1965) recognized that the expected rates of return on the option and stock would generally be different due to their different risk characteristics. He posited a higher (constant) expected rate of return for the option, β, although recognizing that a 'deeper theory would deduce the value of [the expected rate of return]'. He also realized that this assumption would mean that it might be optimal to exercise a call option prior to its maturity but was unable to solve for the optimal exercise policy except in the case of perpetual calls. His model for a European call was

$$C = e^{(\alpha-\beta)\tau} S \cdot \Phi\left[\frac{\ln(S/X) + (\alpha + \frac{1}{2}\sigma^2)\tau}{\sigma\sqrt{\tau}}\right] - e^{-\beta\tau} X \cdot \Phi\left[\frac{\ln(S/X) + (\alpha - \frac{1}{2}\sigma^2)\tau}{\sigma\sqrt{\tau}}\right]. \quad (7)$$

Boness's equation above is a special case of this model for $\alpha = \beta$.

Samuelson and Merton (1969) examined option pricing in a simple equilibrium model of portfolio choice that allowed them to determine the stock's and option's expected rates of return endogenously. They verified that the option problem could be stated in 'utili-probability' terms in a function form identical to the problem statement in terms of the true probabilities. When stated in this fashion, the adjusted expected rates of return on the stock and option were the same. This approach anticipated the development of the risk-neutral or preference-free method of valuing options that is now accepted as a matter of course.

The Black–Scholes Option Pricing Model

The Black–Scholes option pricing model is based on the principle that there should be no arbitrage opportunities available in the market. The following simple model, due to Cox et al. (1979), can be used to illustrate the principle behind the Black–Scholes model.

Assume that over a single period the stock price can change in only one of two ways. From its current level S, the stock price can increase to hS or fall to kS. Let $C(S,n)$ denote the value of a call option on the stock when the stock price is S and there are n of these 'steps' remaining before the option matures.

Consider a portfolio that is short one call option and long N shares of stock. This portfolio is currently worth $NS - C(S,n)$. After one period this portfolio will be worth either $NhS - C(hS, n-1)$ or $NkS - C(kS,n-1)$. Suppose N is chosen so that these last two quantities are equal; i.e.,

$$N = \frac{C(hS, n-1) - C(kS, n-1)}{(h-k)S} \quad (8)$$

then after one period the portfolio will be worth

$$\frac{kC(hS, n-1) - hC(kS, n-1)}{(h-k)} \quad (9)$$

with certainty. To avoid an arbitrage opportunity the current value of the portfolio must be equal to this value discounted at $(1 + R)$ where R is the risk-free rate of interest (not annualized) over the time of a single step in the stock price. That is,

$$C(S,n) = \frac{1}{1+R}\left[\frac{1+R-k}{h-k}C(hS,n-1)+\frac{h-1-R}{h-k}C(kS,n-1)\right]. \tag{10}$$

This equation relates the value of a n step call option to the value of a $n - 1$ step call. At the time it matures, the value of a call with an exercise price of X is $C(S, 0) = \text{Max}\ (S - X, 0)$. As this functional form is known, (10) can be used to derive the value of a one-period call for different stock prices. Given these values, (10) can be used again to derive the value of a two-period call. The value of any call can be computed by using (10) recursively.

The resulting formula for a n step call is

$$C(S,n) = (1+R)^{-n} \times \sum_{i=1}^{n} \frac{n!}{i!(n-i)!} q^i (1-q)^{n-i} \left(Sh^i k^{n-1} - X\right) \tag{11}$$

where $q \equiv (1 + R - k)\ (h - k)$ and I is the smallest integer for which $Sh^i k_{n-i} \geq X$.

The fraction and the next two terms involving q in the summation can be recognized as the probability of i successes in n trials with a success probability of q from a binomial distribution. Thus the formula in (11) can be rewritten as

$$C(S,n) = (1+R)^{-n} E^*[\text{Max}(S_n - X, 0)] \tag{12}$$

where S_n is the random stock price after n steps and $E^*[\cdot]$ denotes the expectation using the artificial probabilities q and $1 - q$ for the up and down steps. Similarly equation (10) can be expressed as

$$C(S,n) = \frac{1}{1+R}[qC(hS,n-1)+(1-q)C(kS,n-1)] = \frac{1}{1+R}E^*[C(S,n-1)]. \tag{13}$$

Again an 'artificial' expectation has been taken. It should be noted that q is not the actual probability that the stock price will change from S to hS – in fact this true probability has not be used here at all.

In deriving their model Black and Scholes did not assume that the stock price followed this binomial step process. They used instead a geometric or lognormal Brownian motion process. Geometric Brownian motion can be constructed as the limit of this type of binomial process as the step sizes $h - 1$ and $k - 1$ shrink to zero while the number of steps per unit time goes to infinity.

Taking these limits in (10) gives the Black–Scholes partial differential equation

$$\tfrac{1}{2}\sigma^2 S^2 C_{SS} + rSC_s - rC + C_t = 0 \tag{14}$$

where r is the continuously-compounded (annualized) rate of interest on a risk-free asset, σ^2 is the variance of changes in the logarithm of the stock price per unit time and subscripts on C denote partial differentiation. Applying the limits to (11) yields the Black–Scholes call option pricing formula

$$C(S,\tau) = S \cdot \Phi\left[\frac{\ln(S/X) + (r + \frac{1}{2}\sigma^2)\tau}{\sigma\sqrt{\tau}}\right] - e^{-r\tau} X \cdot \Phi\left[\frac{\ln(S/X) - (r + \frac{1}{2}\sigma^2)\tau}{\sigma\sqrt{\tau}}\right]. \tag{15}$$

where $\Phi\ (\cdot)$ is the standard cumulative normal distribution function and $\tau \equiv T - t$ is the time until maturity. (Black and Scholes derived this differential equation and its solution working directly with the continuous time diffusion and not by taking limits.)

The Black–Scholes formula is identical to Samuelson's with $\alpha = \beta = r$ and to Boness's with $\alpha = r$. In fact the most remarkable feature about the model is that the resulting formula does not depend on the stock's or the option's expected rates of return or any measure of the market's risk aversion. Only five variables determine the option's price: S, τ, r, X and. σ^2 Except for the variance, each of these variables is known, and

the variance can be measured with a high degree of certainty.

The absence of the expected rates of return or any measure of risk aversion from the Black–Scholes model was at first troubling. This puzzle was explained by Cox and Ross (1976a) and Merton (1976) who introduced the risk neutral or martingale representation. This idea was later developed more formally by Harrison and Kreps (1979) and others.

The fact that a hedging argument can be used to derive (10), which does not include explicitly expected rates of return, investor preferences, or probabilities means that given the stock price and the interest rate, the value of the option cannot depend *directly* on these either. To solve for the option price, then, we need only find the equilibrium solution in some world where returns, preferences, and probabilities are consistent with the actual stock price process and interest rate. The solution obtained will then be generally applicable.

The most convenient choice of equilibrium is often an economy with risk neutral investors. In such an economy all expected rates of return must be equal to the risk-free rate. If the stock price has a lognormal distribution, then Boness's model applies with $\alpha = r$.

In the risk neutral economy the Black–Scholes formula has an interpretation identical to that in (12). The cumulative normal in the second term in (14) is the risk neutral 'probability' that the option will mature in-the-money. Thus, the second term is the discount factor multiplied by the 'expected' exercise payment. The first term is the discounted value of the expectation of the stock's price at expiration conditional on $S_T > X$.

Extensions of the Black–Scholes Model

The derivation of the Black–Scholes model rests on six assumptions: (i) There are no transactions costs, taxes or restrictions on short sales. (ii) The risk-free rate of interest is constant. (iii) The stock pays no dividends. (iv) The stock price evolution is geometric Brownian motion. (v) The market is open continuously for trading. (vi) The option is European.

Subsequent modifications of the basic model have shown that it is quite robust with respect to relaxations of these assumptions. Thorpe (1973) examined the short sale constraint. Leland (1985) allowed for transactions costs. Ingersoll (1976) and Scholes (1976) considered the effects of differing tax rates on capital gains and dividends. Merton (1973) generalized the model to allow for dividends and a stochastic interest rate. He also proved that assumption (vi) was not necessary if the stock did not pay dividends. Cox and Ross (1976a) and Merton (1973) utilized alternative stochastic processes. Cox and Ross (1976a) and Merton (1976) considered the option problem when the stock's price evolution did not have a continuous sample path. Rubinstein (1976) and Brennan (1979) obtained the Black–Scholes solution with discrete-time trading by imposing conditions on the utility function of the representative investor.

Other types of options have also been valued using the same methods or extensions of them. Some examples are European puts by Black and Scholes (1973), 'down-and-out' options by Merton (1973), commodity options by Black (1976) and interest rate options by Cox et al. (1985b). To solve these or similar problems, the Black–Scholes partial differential equation (14) is used.

While (10) and, therefore (14), were developed to price call options, the characteristics of the call are captured entirely by the condition at maturity $C(S, 0) = \text{Max}(S - X, 0)$. Thus, this equation is a general one that can be used to price calls, puts, or any other derivative asset whose value depends on just the price of the primitive asset.

To solve this equation for other problems the appropriate boundary condition is required

$$C(S, T) = H(S). \tag{16}$$

$H(\cdot)$ specifies a contractual or otherwise known payment at the derivative asset's maturity. If the derivative asset's value arises solely from this payment at maturity, then the formal solution to (14) with boundary condition (16) is

$$C(S, t) = \mathrm{e}^{-r(T-t)} E^*[H(S)]. \tag{17}$$

For some contracts a portion or all of the value may be due to payments that are received at random times prior to maturity. In this case (17) does not measure the full value. For example, a down-and-out option is a call contract that is cancelled if and when the stock price falls below the 'knock-out' price. At this point a partial rebate is usually given. Let K and R denote the knock-out price and rebate. Then the conditions imposed to value this option are

$$\begin{aligned} C(K,u) &= R \qquad \forall u < T \\ C(S,T) &= \mathrm{Max}(S-X,0) \\ &\quad \text{if } S(u) > K \quad \text{for } t < u < T. \end{aligned} \tag{18}$$

The value of the down-and-out option is

$$\begin{aligned} C(S,t) &= RE^*\left[\mathrm{e}^{-r(U-t)} I(U \le T)\right] \\ &\quad + \mathrm{e}^{-r(T-t)} E^*[\mathrm{Max}(S_T - X, 0) I(U > T)]. \end{aligned} \tag{19}$$

Here U is a random variable that takes on the value u if the first time that the stock price drops to K is u. $I(\cdot)$ is an indicator function with the value one if its argument is true and zero otherwise. The first expectation is taken over the random variable U. This term measures the value contributed by the receipt of the rebate. The second expectation is taken over both random variables U and S_T. This term measures the value contributed by the right to exercise if it was not cancelled.

The pricing of the American put has a similar feature. The payment received upon exercise, $X - S$, is known (conditional on the stock price at that time) but its timing is not. In addition, unlike the timing of the rebate in the previous problem, the timing of the exercise is not contractually stated. It is chosen by the put's owner.

Suppose that the put owner chooses a rule for exercising. This rule will generate a random time U at which the option is exercised. The random variable U must be a Markov time; that is, whether or not exercise occurs at a particular time can depend on information known at that time but cannot in any way anticipate the future. For a given rule U, the put's value is

$$E^*\left[\mathrm{e}^{-r(U-t)}(X - S_U)\right]. \tag{20}$$

As the owner of the put has the choice, the rule chosen will be that which maximizes the value of the option

$$P(S,t) = \sup_U E^*\left[\mathrm{e}^{-r(U-t)}(X - S_U)\right]. \tag{21}$$

In principle the American put could be valued by solving (20) for all exercise rules and choosing that one which maximized the value. Samuelson (1965) conjectured and Merton (1973) proved that in such problems the value and the optimal exercise rule could be determined simultaneously by imposing the 'high contact' condition.

The partial differential equation (14) is solved subject to the maturity condition $P(S, T) = \mathrm{Max}(X - S, 0)$ and

$$P[K(t), t] = X - K(t) \tag{22a}$$

$$\left.\frac{\partial P(S,t)}{\partial S}\right|_{S=K(t)} = -1. \tag{22b}$$

$K(t)$ denotes the optimal exercise policy; that is if the stock price falls to $K(t)$ at time t, then the put is exercised. Equation (22a) is the standard condition at exercise. Equation (22b) is the high contact condition.

The high contact requirement assures that for the optimal policy the slope of the pricing function, $P(\cdot)$ is equal to the slope of the payoff function (-1 in the relevant region of exercise). This is just the usual tangency condition at an optimum.

No analytical solution to the American put problem has yet been derived. Brennan and Schwartz (1977), Parkinson (1977) and others have described numerical techniques for these problems and other contracts for which there are no analytical solutions.

Applications of Option Pricing to Valuing Corporate Securities

After deriving their call option formula Black and Scholes make an observation that may be one of

the most important in the field of finance. They argue that the same methods can be used to value other contingent claims, in particular the components of a firm's capital structure. This observation has led to an enormous amount of research. Option pricing techniques have been applied to a wide variety of financial instruments and contracts including corporate bonds, futures, variable rate mortgages, insurance, investment timing advice, and the tax code.

For the simplest problems the call formula can be applied directly. Consider a firm with assets whose value, V, evolves according to a geometric Brownian process. The firm's capital structure consists of common stock and single issue of zero coupon bonds with an aggregate face value of B which mature at time T. At that time the firm will be liquidated.

If $V_T \geq B$, then the bondholders can be paid and the equity will be worth $V_T - B$. If $V_T < B$, the assets will be insufficient to pay the bondholders, and there will be nothing left for the shareholders. Thus, the payoff to the common shares is $\text{Max}(V_T - B, 0)$. This is just like a call option so currently the equity must be worth $C(V, T - t; B)$. By the Modigliani–Miller irrelevancy theorem, the value of the debt and equity must sum to V so the debt is worth

$$D(V, T-t; B) = V - C(V, T-t; B). \tag{23}$$

This same valuation applies even if the firm is not to be liquidated. To repay the bondholders, the firm must raise B dollars. Selling assets to do this is the same as a liquidation. The only other way to raise this money is by a new offering of securities. To raise B dollars the firm will have to offer a security that is worth B. If the firm's assets are not worth at least B, this cannot be done. If they are, then again by the Modigliani–Miller theorem the original equity will be worth $V_T - B$.

A zero coupon convertible bond can be priced similarly. Suppose there are N shares of common outstanding and the convertibles can be exchanged for n shares in aggregate. If all the bondholders convert, then they will own the fraction $\gamma \equiv n/(N+n)$ of the equity. Clearly the bondholders will convert if $\gamma V_T > B$. Otherwise they will receive B, unless the firm is insolvent, in which case they will get just V_T. Thus, the bondholders will receive

$$\begin{aligned}\text{Max}[\gamma V, \text{Min}(V, B)] = {} & \text{Max}(\gamma V - B, 0) \\ & + [V - \text{Max}(V - B, 0)].\end{aligned} \tag{24}$$

This is the payoff to an option plus an ordinary zero coupon bond so the convertible's value must be $C(\gamma V, T - t; B) + D(V, T - t; B)$. If the convertible is also callable, as most are, then methods used to determine the optimal exercise policy for and the value of an American put must be used. This problem has been solved by Ingersoll (1977).

Most corporate securities receive periodic coupons or dividends. While a default-free coupon bond can be valued as a portfolio of zero coupon bonds, this method will not work when there is default risk because the omission of one coupon puts the whole bond in default. These securities can be priced as a series of options, however.

Consider a company with common stock on which it is not paying dividends and a single issue of coupon bonds with aggregate periodic coupons of c, at times $T_1, \ldots, T_n$, and an aggregate par value of B, repaid at T_n. Once the next to last coupon is paid only a single payment remains $B + c$. Therefore, just after the next to last payment the bond can be treated like a zero coupon bond. Its value at that time is

$$D_{n-1}V, T_{n-1} = D(V, T_n - T_{n-1}; B + c).$$

Between times T_{n-2} and T_{n-1} the company makes no payments to the holders of its securities so the standard Black–Scholes equation (14) applies. The solution for the bond's value at time T_{n-2} is

$$\begin{aligned}& D_{n-2}(V, T_{n-2}) \\ & = \mathrm{e}^{-r(T_{n-1} - T_{n-2})} E^*[D(V_{T_{n-1}}, T_{n-1})]\end{aligned} \tag{25}$$

as given in (17). The price at earlier times can be determined by a recursive application of (25). Geske (1977) addresses this compound option problem.

Another way to price claims with coupons or dividends is to approximate the sequence of

payments as continuous flows. The general problem is to value a particular claim, $F(S, t)$, when the price evolution of the firm's value is

$$dV = [\alpha V - \Delta(V,t)]dt + \sigma V d\omega. \quad (26)$$

$\Delta(V,t)$ is the total flow of all disbursements (dividends, coupons, etc.) paid by the firm and $d\omega$ is the increment to a Wiener process.

The equilibrium price process for the claim is

$$dF(V,t) = [\beta(V,t)F - \delta(V,t)]dt + (F_v/F)\sigma V d\omega \quad (27)$$

where $\beta(\cdot)$ is the (endogenous) expected rate of return on the derivative asset and $\delta(\cdot)$ is the portion of the total disbursement received by the owners of the derivative asset.

Ito's Lemma is used to determine the expected rate of price appreciation which is equated to the rate of capital gains required in equilibrium to earn β.

$$\tfrac{1}{2}\sigma^2 V^2 F_{vv} + [\alpha V - \Delta(V,t)]F_V + F_t = \beta(V,t)F(V,t) - \delta(V,t). \quad (28)$$

The equivalent risk neutral processes replace α and β by r, the risk-free rate. Thus, the general valuation equation is

$$\tfrac{1}{2}\sigma^2 V^2 F_{VV} + [rV - \Delta(V,t)]F_V - rF + F_t - \delta(V,t) = 0. \quad (29)$$

Equation (29) is the fundamental valuation equation for the financial claims against a firm. It can be used for any situation when the standard Black–Scholes conditions hold and the value of the claim to be priced depends solely on time and the value of the assets of the firm. The basic requirement for this second condition is that there be no other sources of uncertainty beyond that affecting the value of the assets. Thus, the interest rate cannot be stochastic, the dividend policy must be a known function of the firm value and time, the firm cannot alter its investment or financing policies in unanticipated ways.

If this second requirement is not met, then the value of the claim being priced will depend on other variables as well – variables that measure the overall state of the economy. Cox et al. (1985a) have developed a theoretical context in which all these pricing problems can be handled. The basic Black–Scholes method is still valid, but the pricing equation will include these additional state variables.

Other Applications of Option Pricing

In recent years option pricing techniques have been used in a great variety of situations. PBGC insurance and the effects of ERISA on corporate pension plans have been considered as have FDIC insurance and the implicit insurance in government loan guarantees. The asymmetries of the tax code and their effects on corporations and investors have been analysed. Option pricing methods have been used to value market timing advice and to examine the efficiency of dynamic portfolio strategies such as contingent immunization. More on the applications of option pricing and extensive bibliographies can be found in the survey articles by Mason and Merton (1985) and Smith (1976) and in the texts by Cox and Rubinstein (1985) and Ingersoll (1987).

It should be clear that the realm of applications goes far beyond the more obvious corporate securities. A bibliography of the published papers alone would be extensive, and working papers are continually added. Option pricing theory has become an important element in our understanding of financial contracting and a practical tool in widespread applications.

See Also

▶ Finance

Bibliography

Bachelier, L. 1900. Théorie de la speculation. *Annales de l'Ecole Normale Superieure*. Trans. A.J. Boness in *The random character of stock market prices*, ed. P.-H. Cootner. Cambridge, MA: MIT Press, 1967.

Black, F. 1976. The pricing of commodity contracts. *Journal of Financial Economics* 3(1–2): 167–179.
Black, F., and M.J. Scholes. 1973. The pricing of options and corporate liabilities. *Journal of Political Economy* 81(3): 637–654.
Boness, A.J. 1964. Elements of a theory of stock option value. *Journal of Political Economy* 72(2): 163–175.
Brennan, M.J. 1979. The pricing of contingent claims in discrete time models. *Journal of Finance* 34(1): 53–68.
Brennan, M.J., and E.S. Schwartz. 1977. The valuation of American put options. *Journal of Finance* 32(2): 449–462.
Cox, J.C., and S.A. Ross. 1976a. The valuation of options for alternative stochastic processes. *Journal of Financial Economics* 3(1–2): 145–166.
Cox, J.C., and S.A. Ross. 1976b. A survey of some new results in financial option pricing policy. *Journal of Finance* 31(2): 383–402.
Cox, J.C., and M. Rubinstein. 1985. *Options markets*. Englewood Cliffs, NJ: Prentice-Hall.
Cox, J.C., S.A. Ross, and M. Rubinstein. 1979. Option pricing: A simplified approach. *Journal of Financial Economics* 7(3): 229–263.
Cox, J.C., J.E. Ingersoll, and S.A. Ross. 1985a. An intertemporal general equilibrium model of asset prices. *Econometrica* 53(2): 363–384.
Cox, J.C., J.E. Ingersoll, and S.A. Ross. 1985b. A theory of the term structure of interest rates. *Econometrica* 53(2): 385–407.
Geske, R. 1977. The valuation of corporate liabilities as compound options. *Journal of Financial and Quantitative Analysis* 12(4): 541–552.
Harrison, J.M., and D. Kreps. 1979. Martingales and arbitrage in multiperiod securities markets. *Journal of Economic Theory* 20(3): 381–408.
Ingersoll, J.E. 1976. A theoretical and empirical investigation of the dual purpose funds: an application of contingent-claims analysis. *Journal of Financial Economics* 3(1–2): 83–123.
Ingersoll, J.E. 1977. A contingent-claims valuation of convertible securities. *Journal of Financial Economics* 4(3): 289–322.
Ingersoll, J.E. 1987. *Theory of financial decision making*. Totowa, NJ: Rowman and Littlefield.
Kassouf, S.T. 1969. An econometric model for option price with implications for investors' expectations and audacity. *Econometrica* 37(4): 685–694.
Leland, H.E. 1985. Option pricing and replication with transactions costs. *Journal of Finance* 40(5): 1283–1301.
Mason, S.P., and R.C. Merton. 1985. The role of contingent claims analysis in corporate finance. In *Recent advances in corporate finance*, ed. E.I. Altman and M.G. Subramanyam. Homewood: Richard D. Irwin.
Merton, R.C. 1973. The theory of rational option pricing. *Bell Journal of Economics* 4(Spring): 141–183.
Merton, R.C. 1976. Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics* 3(1–2): 125–144.
Parkinson, M. 1977. Option pricing: The American put. *Journal of Business* 50(1): 21–36.
Rubinstein, M. 1976. The valuation of uncertain income streams and the pricing of options. *Bell Journal of Economics* 7(2): 407–425.
Samuelson, P.A. 1965. Rational theory of warrant pricing. *Industrial Management Review* 6(2): 13–32.
Samuelson, P.A., and R.C. Merton. 1969. A complete model of warrant pricing that maximizes utility. *Industrial Management Review* 10(Winter): 17–46.
Scholes, M.J. 1976. Taxes and the pricing of options. *Journal of Finance* 31(2): 319–332.
Smith, C.W. 1976. Option pricing: A review. *Journal of Financial Economics* 3(1–2): 3–51.
Sprenkle, C.M. 1961. Warrant prices as indicators of expectations and preferences. *Yale Economic Essays* 1(2): 178–231. Reprinted in *The random character of stock market prices*, ed. P.H. Cootner. Cambridge, MA: MIT Press, 1967.
Stoll, H.R. 1969. The relationship between put and call option prices. *Journal of Finance* 24(5): 801–824.
Thorpe, E.O. 1973. Extensions of the Black–Scholes option model. *Bulletin of the International Statistical Institute, Proceedings of the 39th Session*, 522–529.

Options

Robert C. Merton

Abstract

An option is a security whose owner has a right to buy (sell) it at a specified price on a specified date (or, with an American-type option, on or before the specified date). Trading of options on common stock began in 1973 and has since spread to other commodities. Option pricing theory provides a unified theory for the pricing of corporate liabilities. Of its more recent extensions, perhaps the most significant is its application in the evaluation of operating or 'real' options in the capital budgeting decision problem.

Keywords

Arbitrage; Bachelier, L.; Black, F.; Capital budgeting; Deposit insurance; Forward contracts; Futures contracts; Government loan guarantees; Ito's lemma; Merton, R. C.; Option

pricing theory; Options; Pension fund insurance; Probability; Scholes, M.

JEL Classifications
G1

A 'European-type call (put) option' is a security that gives its owner the right to buy (sell) a specified quantity of a financial or real asset at a specified price, the 'exercise price', on a specified date, the 'expiration date'. An American-type option provides that its owner can exercise the option on or before the expiration date. If an option is not exercised on or before the expiration date, it expires and becomes worthless.

Options and forward or futures contracts are fundamentally different securities. Both provide for the purchase (or sale) of the underlying asset at a future date. A long position in a forward contract obliges its holder to make an unconditional purchase of the asset at the forward price. In contrast, the holder of a call option can choose whether or not to purchase the asset at the exercise price. Thus, a forward contract can have a negative value whereas an option contract never can.

The first organized market for trading options was the Chicago Board Options Exchange (CBOE) which began trading options on common stocks in 1973. The initial success of the CBOE was followed by an expansion in markets to include options on fixed-income securities, currencies, stock and bond indices, and a variety of commodities. Although these markets represent an increasingly larger component of total financial market trading, options are still relatively specialized financial securities. Option pricing theory has, nevertheless, become one of the cornerstones of financial economic theory.

This central role for options analysis derives from the fact that option-like structures pervade virtually every part of the field. Black and Scholes (1973) provide an early example: shares of stock in a firm financed in part by debt have a payoff structure which is equivalent to a call option on the firm's assets where the exercise price is the face value of the debt and the expiration date is the maturity date of the debt. Option pricing theory can thus be used to price levered equity and, therefore, corporate debt with default risk.

Identification of similar isomorphic relations between options and other financial instruments has led to pricing models for seniority, call provisions and sinking fund arrangements on debt; bonds convertible into stock, commodities, or different currencies; floor and ceiling arrangements on interest rates; stock and debt warrants; rights and stand-by agreements. In short, option pricing theory provides a unified theory for the pricing of corporate liabilities.

The option-pricing methodology has been applied to the evaluation of noncorporate financial arrangements including government loan guarantees, pension fund insurance and deposit insurance. It has also been used to evaluate a variety of employee compensation packages including stock options, guaranteed wage floors, and even tenure for university faculty.

Perhaps the most significant among the more recent extensions of option analysis is its application in the evaluation of operating or 'real' options in the capital budgeting decision problem. For example, a production facility which can use various inputs and produce various outputs provides the firm with operating options that it would not have with a specialized facility which uses a fixed set of inputs and produces a single type of output. Option-pricing theory provides the means of valuing these production options for comparison with the larger initial cost or lower operating efficiency of the more flexible facility. Similarly, the choice among technologies with various mixes of fixed and variable costs can be treated as evaluating the various options to change production levels, including abandonment of the project. Research and development projects can be evaluated by viewing them as options to enter new markets, expand market share or reduce production costs.

As these examples suggest, option analysis is especially well suited to the task of evaluating the 'flexibility' components of projects. These, corporate strategists often claim, are precisely the components whose values are not properly measured by traditional capital-budgeting techniques. Hence, option-pricing theory holds for

the promise of providing quantitative assessments for capital budgeting projects that heretofore were largely evaluated qualitatively. Survey articles by Smith (1976) and Mason and Merton (1985) provide detailed discussion of these developments in option analysis along the extensive bibliographies.

The lineage of modern option pricing theory began in 1900 with the Sorbonne thesis, 'Theory of Speculation', by the French mathematician Louis Bachelier. The work is rather remarkable because, in analysis the problem of option pricing, Bachelier derives much of the mathematics of probability diffusions; this, five years before Einstein's famous discovery of the theory of Brownian motion. Although, from today's perspective, the economics and mathematics of Bachelier's work are flawed, the connection of his research with the subsequent path of attempts to describe an equilibrium theory of option pricing is unmistakable. It was not, however, until nearly 75 years later with the publication of the seminal Black and Scholes article (1973), that the field reached a sense of closure on the subject and the explosion in research on option pricing applications began.

As with Bachelier and later researchers, Black and Scholes assume that the dynamics for the price of the asset underlying the option can be described by a diffusion process with a continuous sample path. The breakthrough nature of the Black–Scholes analysis derives from their fundamental insight that the dynamics trading strategy in the underlying asset and a default-free bond can be used to hedge against the risk of either a long or short position in the option. Having derived such a strategy, Black and Scholes determine the equilibrium option price from the equilibrium condition that portfolios with no risk must have the same returns as a default-free bond. Using the mathematics of Ito stochastic integrals, Merton (1973, 1977) formally proves that with continuous trading, the Black–Scholes dynamic portfolio will hedge all the risk of an option position held until price exercise or expiration, and therefore, that the Black–Scholes option price is necessary to rule out arbitrage.

Along the lines of the derivation for general contingent claims pricing in Merton (1977), a sketch of the arbitrage proof for the Black–Scholes price of a European call option on a nondividend-paying stock in a constant interest rate environment is as follows.

Assume that the dynamics of the stock price, $V(t)$, can be described by a diffusion process with a stochastic differential equation representation given by:

$$dV = \alpha V dt + \alpha V dz \tag{1}$$

where α is the instantaneous expected return on the stock; σ^2 is the instantaneous variance per unit time of the return, which is a function of V and t; dz is a standard Wiener process. Let $F[V, t]$ satisfy the linear partial

$$0 = \frac{1}{2}\sigma^2 V^2 F_{11} + rVF_1 - rF + F_2 \tag{2}$$

where subscripts denote the partial derivatives and r is the interest rate. Let F be such that it satisfies the boundary conditions:

$$\begin{aligned} F/V \leq 1; \quad F(0,t) = 0; \quad F[V,T] \\ = \max[0, V - E]. \end{aligned} \tag{3}$$

Note from (3) that the value of F on these boundaries are identical to the payoff structure on a European call option with exercise price E and expiration date T. From standard mathematics, the solution to (2) and (3) exists and is unique.

Consider the continuous-time portfolio strategy which allocates the fraction $w(t) \equiv F_1[V, t]\, V(t)/P(t)$ to the stock and $1-w(t)$ to the bond, where $P(t)$ is the value of the portfolio at time t. Other than the initial investment in the portfolio at there are no contributions or from the portfolio until it is liquidated at $t = T$.

The prescription for the portfolio strategy for each time t depends only on the first derivative of the solution to (2)–(3) and the current values of the stock and the portfolio. It follows from the prescribed allocation $w(t)$ that the dynamics for the value of the portfolio can be written as:

$$dP = w(t)P\ dV/V + [1 - w(t)]rP\ dt$$
$$= F_1 dV + r[P - F_1 V]dt. \quad (4)$$

As a solution to (2), F is twice-continuously differentiable. Hence, we can use Ito's Lemma to express the stochastic process for F as:

$$dF = \left[\frac{1}{2}\sigma^2 V^2 F_{11} + \alpha V F_1 + F_2\right]dt + F_1 \sigma V dz \quad (5)$$

where F is evaluated at $V = V(t)$ at each point in time t. But, F satisfies (2). Hence, we can rewrite (5) as:

$$dF = F_1 dV + r[F - F_1 V]dt. \quad (6)$$

Define $Q(t)$ to be the difference between the value of the portfolio and the value of the function $F[V, t]$ evaluated at $V = V(t)$. From (4) and (6), we have that $dQ = rQ\ dt$ which is a nonstochastic differential equation with solution $Q(t) = Q(0)\exp[rt]$ and $Q(0) = P(0) - F[V(0), 0]$. Hence, if the initial investment in the portfolio is chosen so that $P(0) = F[V(0), 0]$ then $Q(t) = 0$ and $P(t) = F[V(t), t]$ for all t.

Thus, we have constructed a dynamic portfolio strategy in the stock and a default-free bond that exactly replicates the payoff structure of a call option on the stock. The solution of (2) and (3) for F and its first derivative F_1 provides the 'blueprint' for that construction. The standard no-arbitrage condition for equilibrium prices holds that two securities with identical payoff structures must have the same price. It follows, therefore, that the equilibrium price of the call option at time t must equal the Black–Scholes price, $F[V(t), t]$.

The extraordinary impact of the Black–Scholes analysis on financial economic research and practice can in large part be explained by three critical elements: (1) the relatively weak assumptions for its valid application; (2) the variables and parameters required as inputs are either directly observable or relatively easy to estimate, and there is computational ease in solving for the price; (3) the generality of the methodology in adapting it to the pricing of other options and option-like securities.

Although framed in an arbitrage type of analysis, the derivation does not depend on the existence of an option on the stock. Hence, the Black–Scholes trading strategy and price function provide the means and the cost for an investor to create synthetically an option when such an option is not available as a traded security. The findings that the equilibrium option price is a twice continuously differentiable function of the stock price and that its dynamics follow an Ito process are derived results, not assumptions.

The striking feature of (2) and (3) is not the variables and parameters that are needed for determining the option price but rather, those not needed. Specifically, determination of the option price and the replicating portfolio strategy does not require estimates of either the expected return on the stock, α or investor risk preferences and endowments. In contrast to most equilibrium models, the pricing of the option does not depend on price and joint distributional information for all available securities. The only such information required is about the underlying stock and default-free bond. Indeed, the only variable or parameter required in the Black–Scholes pricing function that is not directly observable is the variance rate function, σ^2. This observation has stimulated a considerable research effort on variance-rate estimation in both the academic and practising financial communities.

With some notable exceptions, equations (2) and (3) cannot be solved analytically for a closed-form solution. However, powerful computational methods have been developed to provide high-speed numerical solutions of these equations for both the option price and its first derivative.

As in the original Black and Scholes article, the derivation here focuses on the pricing of a European call option. Their methodology is, however, easily applied to the pricing of other securities with payoff structures contingent on the price of the underlying stock. Consider, for example, the determination of the equilibrium price for a European put option with exercise price E and expiration date T. Suppose that in the original derivation we change the boundary conditions specified for F in (3) so as to match the payoff structure of the put option on

these boundaries. That is, we now require that F satisfy $F \leq E$; $F[0, t] = E \exp[-r(T-t)]$; $F[V, T] = \max[0, E-V]$. Once F and its derivative are specified, the development of the replicating portfolio proceeds in identical fashion to show that $P(t) = F[V(t), t]$. With the revised boundary conditions, the portfolio payoff structure will match that of the put option at exercise or expiration. Thus, $F[V(t), t]$ is the equilibrium put option price.

As shown in Merton (1977), the same procedure can be used to determine the equilibrium price for a security with a general contingent payoff structure, $G[V(T)]$, by changing the boundary conditions in (3) so that $F[V, T] = G[V]$. A particularly important application of this procedure is in the determination of pure state-contingent prices.

Let $\pi[V, t; E, T]$ denote the solution of (2) subject to the boundary conditions:

$$\pi/V \leq 1; \quad \pi[0, t; E, T] = 0; \quad \pi[V, T; E, T] = \delta(E-V)$$

where $\delta(x)$ is the Dirac delta function with the properties that

$$\delta(x) = 0 \text{ for } x \neq 0$$

and $\delta(0)$ is infinite in such a way that

$$\int_a^b \delta(x)\, \mathrm{d}x = 1 \text{ for } a < 0 < b.$$

By inspection of this payoff structure, it is evident that this security is the natural generalization of Arrow–Debreu pure state securities to an environment where there is a continuum of states defined by the price of the stock and time. That is loosely, $\pi[V, t; E, T]\mathrm{d}E$ is the price of a security which pays \$1 if $V(T) = E$ at time T and \$0, otherwise.

As is well known from the Green's functions method of solving differential equations, the solution to equation (2) subject to the boundary condition $F[V, T] = G[V]$ can be written as:

$$F[V, t] = \int_0^\infty G[E]\pi[V, t; E, T]\mathrm{d}E. \tag{7}$$

Thus, just as with the standard Arrow–Debreu model, once the set of all pure state-contingent prices, $\{\pi\}$ are derived, the equilibrium price of any contingent payoff structure can be determined by mere summation or quadrature.

To underscore the central importance of call option pricing in the general theory of contingent claims pricing, consider a portfolio containing long and short positions in call options with the same expiration date T where each 'unit' contains a long position in an option with exercise price $E-\varepsilon$; a long position in an option with exercise price $E + \varepsilon$; and a short position in two options with exercise price E. If one takes a position in $1/\varepsilon^2$ units of this portfolio, the payoff structure at time T with $V(T) = V$ is given by:

$$\{\max[0, V + \varepsilon - E] - 2\max[0, V - E] + \max[0, V - \varepsilon - E]\} \times /\varepsilon^2. \tag{8}$$

The limit of (8) as $\varepsilon \to 0$ is $\delta(E-V)$ which is the payoff structure to a pure contingent-state security. If $F[V, t; E, T]$ is the solution to (2) and (3), then it follows from (8) that:

$$\begin{aligned}\pi[V, t; E, T] &= \lim_{\varepsilon \to 0} \{F[V, t; E-\varepsilon, T] \\ &\quad -2F[V, t; E, T] + F[V, t; E+\varepsilon, T]\}/\varepsilon^2 \\ &= \frac{\partial^2 F[V, t; E, T]}{\partial E^2}.\end{aligned} \tag{9}$$

Hence, once the call-option pricing function has been determined, the pure state- contingent prices can be derived from (9).

For further discussion of options, see especially the January/March 1976 issue of the *Journal of Financial Economics;* the October 1978 issue of the *Journal of Business;* and the excellent book by Cox and Rubinstein (1985).

See Also

- ▶ Bachelier, Louis (1870–1946)
- ▶ Finance
- ▶ Options (New Perspectives)

Bibliography

Bachelier, L. 1900. *Théorie de la speculation*. Paris: Gauthier-Villars. English translation in *The random character of stock market prices,* ed. P. Cootner, revised ed, 17–78. Cambridge, MA: MIT Press, 1967.

Black, F., and M. Scholes. 1973. The pricing of options and corporate liabilities. *Journal of Political Economy* 81: 637–659.

Cox, J., and M. Rubinstein. 1985. *Options markets*. Englewood Cliffs: Prentice-Hall.

Mason, S., and R.C. Merton. 1985. The role of contingent claims analysis in corporate finance. In *Recent advances in corporate finance*, ed. E.I. Altman and M.G. Subrahmanyan, 7–54. Homewood: Richard D. Irwin.

Merton, R.C. 1973. Theory of rational option pricing. *Bell Journal of Economics and Management Science* 4 (Spring): 141–183.

Merton, R.C. 1977. On the pricing of contingent claims and the Modigliani-Miller theorem. *Journal of Financial Economics* 5: 241–250.

Smith, C.W. 1976. Option pricing: A review. *Journal of Financial Economics* 3 (1/2): 3–51.

Options (New Perspectives)

Thaleia Zariphopoulou

Abstract

This article provides an overview of risk-neutral valuation methodology and presents historical milestones in the development of quantitative finance. It also discusses current challenges and new perspectives in model choice, pricing and hedging.

Keywords

Arbitrage; Bachelier, L.; Brace-Gatarek-Musiela model; Barrier options; Call options; Continuous-time models; Copulas; Credit default obligations; Credit default swaps; Credit risk; Derivatives; Exotics; Heath–Jarrow–Morton model; Hedging; Incomplete markets; Libor; Martingales; Model calibration; Model specification; Option valuation; Options; Real options; Reduced-form models of default; Risk measures; Risk-neutral pricing; Risk-neutral valuation; Stochastic integration theory; Structural models of default; Swap market models; Term structure models; Value at risk; Vanilla options; Volatility smile; Yield curve

JEL Classifications

G1

In 1973, Black, Scholes and Merton developed a method for the valuation of a European option based on the idea of perfect replication of its payoff. Their approach demonstrates how to act in an uncertain environment so that relevant risks are controlled. Around the same time, trading of options on common stocks started in the Chicago Board Options Exchange. Theory met practice and an exciting and fruitful journey started on the crossroads of economics, finance and mathematics. Its impact was phenomenal in both academia and industry. New areas of research were created, and numerous educational and training activities were established. The derivatives market grew at an unprecedented rate and influenced the development of other markets. Complex mathematical modelling and technical sophistication, predominant elements in theory and applications in engineering and natural sciences, now entered the theory and practice of finance. This was not the first time that stochastic modelling touched finance. At the beginning of the twentieth century, in his pioneering doctoral work, Bachelier (1900) proposed a stochastic model, based on normality assumptions on their returns, for stock prices. In many aspects, however, his work was ahead of its time and had no impact for years to come.

What was the Black, Scholes and Merton option valuation approach? A European call option is a contract that gives its owner the right to buy the underlying stock at a given price, K and a given maturity, T. Their model, powerful and simple, assumed a liquid market environment consisting of a non-defaultable bond and a stock. The bond yields constant interest rate r, while the stock price, S_t, is modelled as a log-normal

diffusion process having constant mean rate of return, μ, and volatility parameter, σ. Applying Ito's formula – a fundamental result of modern stochastic calculus – they were able to build a dynamic self-financing portfolio, (α_t,β_t), $0 \le t \le T$, that replicates the option payoff, that is, for which $\alpha_T + \beta_T = (S_T - K)^+$. For all t, the option price, v_t, is, then, given by the current portfolio value, $v_t = \alpha_t + \beta_t$. Stochastic and differential arguments yield the price process representation $v_t = C(S_t, t)$, with the function C satisfying the partial differential equation

$$C_t + \frac{1}{2}\sigma^2 S^2 C_{SS} + rSC_S = rC \qquad (1)$$

and the terminal condition $C(S, T) = (S - K)^+$. The components of the replicating portfolio turn out to be $\alpha_t = S_t\, C_S\, (S_t, t)$ and $\beta_t = C(S_t, t) - \alpha_t$, representing the amounts invested, respectively, in the stock and bond.

The construction of the price and hedging policies, as well as the specification of various sensitivity indices (greeks), thus amount to solving linear partial differential equations, a relatively easy task given the existing technical body in mathematical analysis.

The industry rapidly adopted the Black and Scholes model as a standard for the valuation of simple (vanilla) options. Soon after, more complex products were created and traded, like options on fixed-income securities, currencies, indices and commodities. Gradually, the options market experienced great growth and its liquidity reached very high levels (for a concise exposition see, for example, Musiela and Rutkowski 2005).

In parallel, substantial advances in research took place. In 1979, Harrison and Kreps laid the foundations for the development of the risk-neutral pricing theory. They created a direct link between derivative valuation and martingale theory. For a finite number of traded securities and under general assumptions on their price processes and related payoffs, they established that the price of a replicable contingent claim corresponds to the expected value, calculated under the risk-neutral probability of the (discounted) claim's payoff. These results were further developed and presented by Harrison and Pliska (1981). In the years that followed, the theory was extended and a model-independent approach for pricing and risk management emerged. In a generic derivatives model, the (discounted) prices of primary assets are represented by a vector-valued semi-martingale $\boldsymbol{S}_s = \left(S_s^1,\ldots,S_s^m\right)$, defined in a probability space $(\Omega, \mathscr{F}, (\mathscr{F}_t), \mathbb{P})$ where $\mathbb{P}$ is the historical measure. The (discounted) payoff, C_T, is taken to be an $\mathscr{F}_T$-measurable random variable.

The derivative price, discounted under the same numeraire as $\boldsymbol{S}$ and C_T, is given by the conditional expectation

$$v_t\,(C_T\,) = E_{\mathbb{Q}}(C_T/\mathscr{F}_t). \qquad (2)$$

The pricing measure $\mathbb{Q}$ is equivalent to $\mathbb{P}$ and, under it, the (discounted) price processes become martingales, that is, $E_{\mathbb{Q}}\,(S_s|\mathscr{F}_t) = S_t$, $t \le s \le T$. The derivative prices, themselves martingales under $\mathbb{Q}$, are linear with respect to their payoffs, time and numeraire consistent and independent of their holder's risk preferences.

Fundamental questions in risk-neutral valuation are related to existence and uniqueness of the derivative price. Uniqueness turns out to be equivalent to the replicability of all claims in the market. Such a market is classified as complete. Stochastic integration theory was used to establish that market completeness is equivalent to uniqueness of the risk-neutral martingale measure $\mathbb{Q}$. In this case, the price is given by (2) and, thus, exists and is unique. If, however, the market is not complete there is multiplicity of equivalent martingale measures. In this case, perfect replication is abandoned and absence of arbitrage becomes the key requirement for price specification and model choice. In an arbitrage-free model, a judicious choice of the pricing measure is made and the price is still represented as in (2). In many aspects, market completeness and absence of arbitrage are complementary concepts. Their relationship has been extensively studied with the use of martingale theory and functional analysis. Important results in this direction are formulated in the First and Second Fundamental Theorems of Asset Pricing (see, among others, Bjork 2004; Delbaen and Schachermayer 2006).

The risk-neutral valuation theory, built on a surprising fit between stochastic calculus and quantitative needs, revolutionized the derivatives industry. But its impact did not stop there. Because the theory provides a universal approach to price and manage risks, the option pricing methodology has been applied in an array of applications. Indeed, corporate and non-corporate agreements have been analysed from an options perspective. Option techniques have also been applied to the valuation of pension funds, government loan guarantees and insurance plans. In a different direction, applications of the theory resulted in a substantial growth of the fields of real options and decision analysis. Complex issues related, for example, to operational efficiency, financial flexibility, contracting, and initiation and execution of research and development projects were revisited and analysed using derivative valuation arguments (see the review article of Merton 1998).

Since the 1970s, theoretical developments, technological advances, modelling innovations and creation of new derivatives products have been proceeding at a remarkable rate. During this period, theory and practice have been shaping each other in a unique challenging and intense interaction. The rest of the article is, mainly, dedicated to this dimension.

Theory and Practice in Derivatives Markets

The Black and Scholes model included various assumptions that are not valid in practice. Interest rates and volatilities are not constant, trading is not continuous, defaults occur and information is not complete. How did academic research and industry reality react to and handle these issues? Albeit there are very distinct priorities, needs and goals, shortcomings of the theory not only did not limit its applicability but prompted a remarkable progress between the theoretical and the applied worlds. Models were developed and innovative computational techniques were invented, and used in practice, for new complex products (exotics). Progress did not occur simultaneously. While theory developed mostly in bursts, practice continued the use of basic models which often involved self-contradictory assumptions. However, despite internal modelling inconsistencies, industry applications offered valuable intuition and feedback to the abstract theoretical developments.

The first revisited assumption was that the (short) interest rate is constant. Models of stochastic interest rates started appearing, and a major breakthrough occurred in 1992 with the work of Heath, Jarrow and Morton. Moving away from modelling directly the short rate, their novel approach was focused on the dynamics of the entire (instantaneous and continuously compounded) forward curve $f(t, T)$, defined by

$$f(t,T) = -\frac{\partial}{\partial T} \ln B(t,T)$$

where $B(t, T)$ represents the price, at time t, of a zero-coupon discount bond with maturity T. To facilitate the analysis of the forward curve, Musiela (1993) introduced an alternative parametrization, namely, $r(t, x) = f(t, t + x)$, which exhibited the importance of infinite dimensional diffusions and stochastic partial differential equations in finance. This helped to find answers to a number of practical questions related to the yield curve dynamics. Indeed, the issue of consistency between the yield curve construction and its evolution was resolved. Additionally, the support of the yield curve distribution has been studied and the mean reversion, or, more mathematically, stationarity of the entire yield curve dynamics has been addressed.

Clearly, the infinite dimensional analysis was useful in a study of the dynamics of the forward rates for all maturities. There was, however, still a problem that needed to be looked at, namely, that the forward rates $f(t, T)$ are not traded in the market, and the Libor and swap rates are together with options on them. Moreover, information contained in these option prices should be taken into account in the specification of the yield curve dynamics. Because the market trades caps and swaptions in terms of their Black and Scholes volatilities, it would be advantageous to develop

a term structure model that is consistent with such practice, a task seen by many academics at that time, as impossible for its apparent internal inconsistency.

In a series of papers by Miltersen, Sandmann, Sondermann, Brace, Gatarek, Musiela, Rutkowski and Jamshidian (see Part II of Musiela and Rutkowski 2005, for a detailed exposition of these works), a new modelling framework for term structure dynamics was put in place. The so-called Libor, also known as BGM (Brace–Gatarek–Musiela), and swap market models resolved the outstanding issue of the link between the traded instruments and the mathematical description of their dynamics. In essence, they provided a model-independent framework for the analysis of the interest rates dynamics when coupled with the advances – taking place in parallel – in the modelling of volatility smile dynamics. The latter issue is discussed next.

The Black and Scholes model assumes constant volatility and hence, within this model, a call option with arbitrary strike is priced with the same volatility. However, call options of different strikes are priced differently by the market which 'allocates' into the Black and Scholes formula a strike-dependent volatility generating the so-called volatility smile. This is clearly inconsistent with the assumption of the model. It turns out, however, that a complete collection of call prices, for all strikes and maturities, uniquely determines the one-dimensional distributions of the underlying forward price process, under a probability measure which should be interpreted as a forward measure to the option maturity. In a series of papers, Dupire (1993) shows how to construct martingale diffusions with a given set of one-dimensional distributions, demonstrating, once more, that the market practice is theoretically sound and internally consistent when analysed from the perspective of the appropriate model. The Black and Scholes model is used only to convert the quoted volatility into a price and it is no longer used for the pricing of vanilla options. Moreover, there are many ways of constructing martingales with a given set of one-dimensional marginals, and the question is not so much how to construct one but, rather, which one to choose and under which criteria. The important message here is that, again, one can now look at the problem in a completely model-independent way, provided all objects – namely, the underlying assets, the associated probability measures and the relevant market information – are correctly interpreted.

Obviously, the theory and practice, at least in the equity, foreign exchange and interest rates derivatives markets, have moved to a different level and reached a certain degree of maturity. Of course, important challenges remain but experience since the 1970s defines clearly a path to follow.

Current Challenges and Perspectives

Credit Risk

A fundamental assumption of the Black and Scholes model is that the underlying securities do not default. However, default is a realistic element of financial contracts and very relevant to any firm's performance. Credit-linked instruments have, by now, become a central feature in derivatives markets. These are financial products that pay their holders amounts contingent on the occurrence of a default event ranging from bankruptcy of a firm to failure to honour a financial agreement. Examples include, among others, credit default swaps (CDS), credit default obligations (CDO) and tranches of indices. Their market has grown more than eightfold in recent years and, undoubtedly, credit risk is, today, one of the most active and challenging areas in quantitative finance.

There are various issues that make the problems in credit risk difficult, from both the modelling and the implementation point of view. The first challenge is how to model the time of default. In academic research, there are two well-established approaches, the structural and the reduced. In the structural models, it is postulated that uncertainty related to default is exclusively generated by the firm's value. Modelling default, then, amounts to building a good model for the company's assets and determining when the latter will fall below existing liabilities. However, such default times are, typically, predictable which is

not only unrealistic, but, also, difficult to implement due to limited public information about the firm's prospects. In the other extreme, the reduced-form models, the default time is associated with a point process with an exogenously given stochastic jump intensity. The intensity essentially measures the instantaneous likelihood of default. Reduced models are more tractable for pricing and calibration but the default times are completely sudden (totally inaccessible), a non-realistic feature. Recently, efforts have been made to bridge the two approaches by incorporating the limited information the investors might have about the firm's value. This information-based approach is gradually emerging but a number of modelling and technical serious issues remain to be tackled. See, among others, Bielecki and Rutkowski (2002) and Schönbucher (2003).

Even though the above models are theoretically sound, their practical implementation is so difficult that it makes them, effectively, inapplicable. The main problem stems from the high dimensionality and inability to develop computational methods that track 'name by name' the valuation outputs. For this reason, the focus in the industry has shifted to an alternative direction centred on modelling the joint distribution of default times. An important development in this direction is the use of a copula function, a concept introduced in statistics by Sklar (1959). The aim is to define the joint distribution of a family of random variables when their individual marginal distributions are known. Such marginal distributions may be, frequently, recovered from the market, as is the case with CDS that yield implicit information on the underlying name's default time. Today, the most widely used copula is the one-factor Gaussian one, proposed by Li (2000). Its popularity lies in the ability to obtain the sensitivity, and thus information on hedging, of the derivative price in a name by name correspondence.

Model Specification

As has been mentioned earlier, the theory has long departed from perfect replication, and practice never relied on it. Absence of arbitrage is the underlying pricing criterion in the derivatives market. However, a plethora of pricing issues and model specifications arise every day. Derivatives markets have been growing very rapidly, and high liquidity in vanilla options on a large number of underlyings including, among others, single stocks and equity indices, interest rates, foreign exchange and commodities, has been achieved. The users benefit from competitive prices, quoted at very tight spreads, for the protection they need. This, in itself, brings another challenge to the providers of such services and products, namely, the models that are currently under development need to reflect this liquidity before they can be used for the pricing of less liquid products. This process is known in the industry as model calibration. To a large extent, one can assume that the market gives the prices for simple derivatives like calls and puts and, hence, pricing considerations dissolve. However, more exotic options need to be priced and this must be done in a way consistent with the basic products (vanilla).

To provide some intuition, consider the case of the so-called first generation exotic, namely, a down and out call option. This is a barrier option that reduces to a simple call option when the likelihood of crossing the barrier is very small. Consequently, a model to price such an option must return the market price of a call in such a scenario. Call prices will be liquid for all strikes up until a certain maturity, say, 18 months or two years for currency options. However, there may be a need to price products with embedded currency options of very long maturity, like up to 50 years in dollar–yen exchange rate. In this case, a suitable model needs to be developed that accommodates short- and long-term issues. On one hand, the model must fit the short-dated foreign exchange (FX) calls and puts. On the other, it has to be consistent with the interest rates volatilities and must capture correctly the dependence structure between the dollar and yen interest rates curves, their volatilities and the spot FX.

A standard approach for solving such problems consists of writing a continuous-time model and trying to fit it to the liquid prices. This task is often very difficult to complete. Indeed, as more market information must be put into a model, the more complicated the model gets, the more difficult and time consuming the calibration procedure

becomes, and the more time it takes to produce accurate prices and stable sensitivity reports. To a large extent, model calibration is identical to the specification of one-dimensional distributions of the underlying process. Model specification, on the other hand, can be identified with the specification of an infinite dimensional copula function defining the joint distribution of the entire path, given the marginal distributions that can be deduced from the call prices. At this point, it is important to recall that, often, option payoffs depend solely on a finite dimensional distribution of the underlying process. Consequently, the need to specify the continuous-time dynamics remains valid only if one wants to link the concept of price with perfect replication of the payoff, a requirement that is, in any case, not met in practice.

Seen from this perspective, a new modelling path emerges, namely, one can take the marginals as given by the call prices and choose a copula function in such a way that the joint distribution is consistent with an arbitrage-free model. For example, if one wants to price a forward start option, the distributions of the underlying asset at two different dates are given. Then, only the joint distribution needs to be specified but in such a way that the martingale property is preserved. Clearly, there is an infinite number of ways to build such a martingale, and the choice should be based on additional information – for example, not on the smile as seen today but on the assumptions one might want to make about the smile dynamics.

Risk Measures

As was previously discussed, absence of arbitrage is the fundamental ingredient in derivative pricing. Absence of perfect replication remains, however, a major issue and dictates the creation of financial reserves. To this effect, regulatory policies have been in place for few years now.

These requirements prompted the axiomatic analysis of the so-called risk measures, which are nonlinear indices yielding the capital requirement of financial positions. The theory of coherent risk measures was proposed by Artzner et al. (1999). A popular risk measure is the 'value at risk', which, despite its widespread use, neither promotes diversification nor measures large losses accurately. Since the mid-1990s a substantial research effort has been invested in further developing the theory. Relaxing a scaling assumption in the coherent case has led to the development of convex risk measures. The next step has been the axiomatic construction of dynamic risk measures that are time consistent, an indispensable property of any pricing system.

See Also

- ▶ Hedging
- ▶ Options

Bibliography

Artzner, P., F. Delbaen, J.M. Eber, and D. Heath. 1999. Coherent measures of risk. *Mathematical Finance* 9: 203–228.

Bachelier, L. 1900. Théorie de la spéculation. Ph.-D. dissertation L' Ecole Normale Supérieure. English translation in *The Random Character of Stock Market Prices*, ed. P.H. Cootner. Cambridge, MA: MIT Press.

Bielecki, T.R., and M. Rutkowski. 2002. *Credit risk: Modeling, valuation and hedging*. Berlin: Springer.

Bjork, T. 2004. *Arbitrage theory in continuous time*. 2nd ed. Oxford: Oxford University Press.

Black, F., and M. Scholes. 1973. The pricing of options and corporate liabilities. *Journal of Political Economy* 81: 637–654.

Delbaen, F., and W. Schachermayer. 2006. *The mathematics of arbitrage*. Berlin: Springer.

Dupire, B. 1993. Pricing and hedging with a smile. *Journées Internationales de Finance*. La Baule: IGR–AFFI.

Harrison, M., and D.M. Kreps. 1979. Martingales and arbitrage in multi-period security markets. *Journal of Economic Theory* 20: 381–408.

Harrison, M., and S. Pliska. 1981. Martingales and stochastic integrals in the theory of continuous trading. *Stochastic Processes and Applications* 11: 215–260.

Heath, D., R.A. Jarrow, and A. Morton. 1992. Bond pricing and the term structure of interest rates: A new methodology for contingent claim valuation. *Econometrica* 60: 77–105.

Li, D.X. 2000. On default correlation: A copula function approach. *Journal of Fixed Income* 9: 43–54.

Merton, R. 1973. Theory of rational option pricing. *Bell Journal of Economics and Management Science* 4: 141–183.

Merton, R. 1998. Applications of option-pricing theory: Twenty-five years later. *American Economic Review* 88: 323–349.

Musiela, M. 1993. Stochastic PDEs and term structure models. *Journées Internationales de Finance*. La Baule: IGR-AFFI.
Musiela, M., and M. Rutkowski. 2005. *Martingale methods in financial modelling*. 2nd ed. Berlin: Springer.
Schönbucher, P.J. 2003. *Credit derivatives pricing models. Model, pricing and implementation*. Chichester: Wiley.
Sklar, A. 1959. Fonctions de répartition à n dimensions et leures marges. *Publications de l'Institut de Statistique de l' Université de Paris* 8: 229–231.

Orderings

Charles Blackorby

Keywords

Choice; Compensation criteria; Happiness; Interpersonal utility comparisons; Ordering; Preference orderings; Preferences; Transitivity; Utility measurement; Welfare economics; Well-being

JEL Classifications

C0

An ordering (also called a complete preordering or a weak ordering) is a binary relation which is reflexive, transitive and complete, that is, it is a preordering that is complete.

A binary relation R defined on a set S is a set of ordered pairs of elements of S, that is, a subset of the Cartesian product of S with itself, $S \times S$. One writes xRy (or $(x, y) \in R$) to mean that $x \in S$ stands in relation R to $y \in S$. An ordering is a binary relation, R, which satisfies three properties: (i) reflexivity: for all $x \in S$, xRx; (ii) transitivity: for $x, y, z \in S$, if xRy and yRz, then xRz; and (iii) completeness for all x, $y \in S$, xRy or yRx, where 'or' is used in its non-exclusive sense.

A simple example results from letting S be the real line and R the greater than or equal to relation so that xRy if and only if $x \geq y$. The most common use of orderings in economics is in preference theory where S is a commodity space and R stands for 'at least as desirable as'. Every ordering can be separated into its symmetric and asymmetric factors, respectively, as follows:

xIy if and only if xRy and yRx
and
xPy if and only if xRy and not yRx.

In the case of preference theory, these correspond to indifference and strict preference relations.

In consumer theory orderings first appeared in the work of Wold (1943–4). In an attempt to put utility theory on a more solid foundation, Wold posited the existence of an ordering with certain properties and demonstrated that this could be represented by a continuous real-valued function, thus making absolutely clear that this was an ordinal concept. Perhaps the most innovative and useful aspect of Wold's argument was an insightful definition of a continuous ordering. (An ordering is continuous if the sets $x|xRy, y \in S$ and $x|yRy, y \in S$ are closed.)

The first modern treatment of the subject appears in Arrow (1951). Agents as well as society as a whole are characterized by their orderings over spaces of alternative. That the choices of society be consistent with an ordering, and understanding the implications of that requirement, has been particularly important in welfare economics. For example, various compensation criteria have been shown to fail transitivity (see Gorman 1955) and hence be unsuitable for public decision-making. In addition, by representing agents and society by their orderings, Arrow made the first step toward unravelling a long-standing confusion between the measurability of utility on the one hand and interpersonal comparability on the other. This step was critical if social decision-making was to rest on solid ground; for an accessible discussion of these issues see Blackorby et al. (1984).

It is common in economics to represent agents by their preference orderings. This leads to a set of complicated and somewhat unresolved issues: what are the relationships among the notions of preference, choice and happiness or well-being. Either a preference ordering or the choices of an individual may be viewed as a primitive and they may or may not be mutually consistent; the issues

at stake can, however, be characterized quite precisely. The relationship between either of these and some notion of happiness or well-being is much less clear; for a good introduction to these problems see Sen and Williams (1982).

Bibliography

Arrow, K.J. 1951. *Social choice and individual values*. New York: Wiley. 2nd edn, 1963.

Blackorby, C., D. Donaldson, and J. Weymark. 1984. Social choice with interpersonal utility comparisons: A diagrammatic introduction. *International Economic Review* 25: 327–356.

Gorman, W. 1955. The intransitivity of certain criteria used in welfare economics. *Oxford Economic Papers* 7: 25–35.

Sen, A., and B. Williams, eds. 1982. *Utilitarianism and beyond*. Cambridge: Cambridge University Press.

Wold, H. 1943–4. A synthesis of pure demand analysis, I–III. *Skandinavisk Aktuarietidskrift* 26: 85–118; 220–63; 27: 69–120.

Oresme, Nicholas (1325–1382)

Barry Gordon

Keywords

Aquinas, St. Thomas; Biel, G.; Buridan de Bethune, J.; Debasement; Langenstein, H. von; Money as standard of value; Oresme, N.; William of Ockham

JEL Classifications

B31

A noted mathematician and physicist and Bishop of Lisieux (1377–82), Oresme was a close friend and adviser of Charles V of France. Because of the frequent currency debasements of his era, he contributed a most influential treatise on money, *Tractatus de origine natura jure, et mutationibus monetarum* (*c*1360). Other works by Oresme of interest to economists are his commentaries on Aristotle's politics, economics, and ethics.

Oresme's tract on money owes much to the ideas of Jean Buridan de Bethune, who was Rector of the University of Paris (*c*1327) and may have taught Oresme. Another to take up Buridan's line of monetary thought was Heinrich von Langenstein (1325–97), a German theologian who taught at both Paris and Vienna. In the next century, the doctrines of Buridan and Oresme were developed by Gabriel Biel (1430–95), a founder of the University of Tübingen and its first professor of theology (from 1484). Biel wrote the outstanding *Tractatus de Potentate et Utilate Monetarum.*

Each of the foregoing Schoolmen wrote in the nominalist tradition deriving from the Oxford philosopher William of Ockham (1285–1347), a tradition that stands in opposition to St Thomas Aquinas on money as on much else. The nominalists question the Thomistic understanding of money as a standard of value established by the Prince (that is, by the ruler of the state). In their view, the Prince's right with respect to setting the standard is a limited right.

According to Oresme and the other nominalists, who are reacting against the princely practice of debasement, a particular currency is likely to be an effective medium of exchange only if the nominated values of the units of that currency are acceptable to the citizens who are the users of the medium. They add that the users of the money are the real owners of it, and so have the right to be consulted by the Prince concerning appropriate arrangements.

Such ideas were revolutionary in terms of much earlier Western thought. One notable aspect of the revolution is the shifting of the grounds for thinking about money. Earlier scholastic discussion had concentrated on the morality of individual transactions but here the operation of a monetary system as a whole begins to come into view.

Selected Works

1956. *The De Moneta of Nicholas Oresme and English Mint documents.* Introduction and translation by C. Johnson. London/New York: Nelson.

Bibliography

Biel, G. 1930. In *Treatise on the power and utility of moneys*, ed. R.B. Burke. Philadelphia: University of Pennsylvania Press.
Estrup, H. 1966. Oresme and monetary theory. *Scandinavian Economic History Review* 14(2): 97–116.
Gordon, B. 1975. *Economic analysis before Adam Smith*. London: Macmillan.
Langholm, O. 1983. *Wealth and money in the Aristotelian tradition*. Bergen: Universitetsforlaget.
Menut, A.D. 1957. Introduction and English translation to 'Maistre Nicole Oresme: Le Livre de Yconomique d'Aristotle'. *Transactions of the American Philosophical Society*, N.S. 47: 783–853.
Monroe, A.E. 1923. *Monetary theory before Adam Smith*. Cambridge, MA: Harvard University Press.
Noonan, J.T. 1957. *The scholastic analysis of Usury*. Cambridge, MA: Harvard University Press.

Organic Composition of Capital

Anwar Shaikh

The distinction between labour value transferred and labour value added is crucial to Marx's theory of value. For the capitalist system as a whole, the abstract labour-time previously materialized in machinery and materials (c) merely reappears in the total product. The capital expended for the purchase of c is therefore constant-in-value. On the other hand, whereas the capital expended for the engagement of workers is determined by the labour value of their means of consumption (v), their actual employment results in a quantity of abstract labour-time (l) which is generally different from v. Thus capital expended for the purchase of labour-power is intrinsically variable-in-value. Indeed, the secret of capitalist production is contained precisely in this variability, since surplus value ($s = 1 - v$) only exists to the extent that l is greater than v. It follows from this that for any given total capital expended ($c + v$), its *composition* between c and v is of the utmost importance, because only v expands total capital value from $c + v$ to $c + l = v + s$ (Marx 1867, pp. 421, 571).

The ratio c/v, the *value composition*, is the immediate measure of the composition of capital. But since c represents the value of machines and materials and v the value of labourpower, the (vectors of) technical proportions in which various machines and materials combine with labour (the *technical composition* of capital) clearly stand behind the value composition c/v (Marx 1863, ch. 33; and Marx 1894, ch. 45). That is to say, the technical composition is the inner measure of the composition of capital. Similarly, since $c + v$ materializes itself as $c + l$, we can view the ratio c/l as the outer measure of the composition of capital – the *materialized composition* of capital (Marx 1894, ch. 8). At a more concrete level each of the above value measures acquires a corresponding price counterpart, and each element of any price/value pair is in turn differentiated into stock/flow measures. We shall see that these distinctions can play an important role at times. None the less, because the value relations are so fundamental to the basic argument, we will concentrate our attention on this level.

It is evident that the technical, value, and materialized compositions of capital are intrinsically related. Indeed, it was one of Marx's central claims that the *movements* of all three are dominated by one overriding force: the mechanization of labour process, which is 'the distinguishing historic feature' of the capitalist mode of production.

To see how this works, we begin by reducing the technical composition vector to a scalar measure TC by valuing the current vector elements at time t in terms of the unit values of means of production in some base year t_0. Suppressing the current time subscript t, let $k_j =$ the j th means of production per worker, λ_1, λ_2 indexes of the unit values of means of production and wage goods respectively, $w =$ an index of the real wage per worker, $h =$ the number of hours worked by each worker, all at time t; while λ_{j0}, $\lambda_{i0} =$ the unit values of means of production and wage goods, respectively, and $v_0 =$ a constant representing the labour value of a unit of labourpower, all in the base year t_0. Then

$$K = [K_j] = \text{the technical composition} = \text{a vector of means of production per worker} \quad (1)$$

O

TC = a scalar measure of the technical composition of capital

$$= \sum_j \lambda_{j0} k_j. \tag{2}$$

Next, note that $c/v = c'/v'$ and $c/l = c'/h$ where c' and v' are per worker, and h is the length of the working day. Then

$$c' \equiv \sum_j \lambda_j k_j = \left[\frac{\sum_j \lambda_j k_j}{\sum_j \lambda_{j0} k_j} \right] \sum_j \lambda_{j0} k_j = \lambda_1 TC$$

where λ_1 = the term in brackets = an index of the current unit value of means of production. Similarly,

$$v' \equiv \sum_i \lambda_i w_i = \left[\frac{\sum_i \lambda_i w_i}{\sum_i \lambda_{i0} w_i} \right] \left[\frac{\sum_i \lambda_{i0} w_i}{\sum_i \lambda_{i0} w_{i0}} \right] \left[\sum_i \lambda_{i0} w_{i0} \right] = \lambda_2 w v_0$$

where the terms in brackets are respectively:

λ_2 = an index of the current unit value of means of production

w = an index of the real wage

v_0 = the base year value of labour-power

$$c/v = (TC/v_0)(\lambda_1/\lambda_2)(1/w) \tag{3}$$

$$c/l = (TC/v_0)(\lambda_1)(v_0/h). \tag{4}$$

Now, according to Marx's argument, mechanization is a continual process of increasing the productivity of labour through the use of ever greater quantities of machines and materials per worker. In a mathematical sense, this means a secular rise in most but not necessarily all of the elements of the technical composition vector (which will itself grow in dimension). It is therefore easy to see why the technical composition measure TC will tend to rise secularly, and why, other things being equal, this in turn will transmit an upward tendency to both c/v and c/l through their common term TC/v_0 (equations (3)–(4)). Because this latter term is both the direct gauge of the effect of a rising technical composition on c/v and c/l and also itself a constant-value measure of the current year's value composition, Marx calls it the *organic composition* of capital (Fine and Harris 1976; Shaikh 1978; Weeks 1981).

Accordingly, we write

$$OC = TC/v_0 = \text{the organic composition of capital.} \tag{5}$$

The organic composition OC is evidently the critical link between the technical composition and the value and materialized compositions. But since the latter two have other determinants as well, we need to consider the specific influence of these other factors. In this regard, Marx argues that these other factors act as counter-tendencies which may slow down, but do not negate, the basic upward trend produced by the tendency toward a rising technical composition of capital (Rosdolsky 1977, part V, appendix).

Consider the above expression for the value composition c/v (equation (3)). Here, we see that in addition to the organic composition OC, it depends also on the ratio λ_1/λ_2, and on the real wage w. But the former factor will serve primarily to create fluctuations around the basic trend produced by the rising organic composition, because the diffusion of technical change will tend to confine the variations in λ_1/λ_2 within a fairly narrow range. Therefore, it is only a secularly rising real wage which can cause the trend of the value composition to lag systematically behind that of the organic composition (though at the same time it accelerates the growth of organic composition by enhancing the scope of mechanization) (Marx 1867, ch. 15). The trend of the organic composition is thus an upper bound to that of the value composition. A corresponding lower bound can then be found by noting that the value composition is related to the materialized composition through the rate of surplus value:

$$c/v = (c/l)(l/v) = (c/l)[(v+s)/v] = (c/s)(1 + s/v) \tag{6}$$

On the question of the rate of surplus value, Marx argued that workers could not generally capture all of the gains in productivity achieved through mechanization, so that over time real wages would normally rise more slowly than productivity and the rate of surplus value would tend to rise (Rosdolsky 1977). In the equation (6) above, this in turn immediately implies that the trend of *c/l* will be the lower bound to that of *c/v*.

This brings us to the trend of *c/l* itself. Here, the central theme of Marx's argument is that for individual capitalists the principal purpose of mechanization is to lower their unit production costs and thereby raise their profitability. But the gain of reduced units (flow) costs generally carries with it a corresponding requirement of the increased *capitalization* of production, i.e., a corresponding increase in the scale of investment required per unit output (and hence in unit fixed costs). This familiar tradeoff between unit variable and unit fixed costs (Pratten 1971, pp. 306–7; Weston and Brigham 1982, pp. 145–7) turns out to be a sufficient condition for the rise in the organic composition *OC* to dominate the falling unit value of means of production λ_1, so that the net result is a secularly rising *c/l* (Shaikh 1978, pp. 239–40). And once it has been established that *c/l* rises over time, it follows from our earlier discussion concerning equation (6) that *c/v* also rise secularly. We can therefore say that under the conditions Marx sees as characteristic of capitalist industrialization, the resulting mechanization and capitalization of production expresses itself in a rising technical and hence organic composition *OC*, a less rapidly rising materialized composition *c/l*, and a value composition *c/v* which rises more slowly than the organic composition but more rapidly than the materialized composition.

All of this brings us to the implications of levels and movements of the various measures of the composition of capital. Marx distinguishes three major domains in which these factors are of critical importance. First, there is the domain of price/value relations, in which he uses the inter-industrial dispersion of organic compositions in any given period to derive the principal difference between prices of production and prices proportional to labour values. Here, the crosssectional dispersion in organic compositions is initially taken to reflect the underlying variations in (the vectors of) technical compositions. Marx notes (but does not pursue) the fact that his results would undoubtedly be somewhat modified by the additional complications which arise when one distinguishes the dispersion of value compositions from that of the technical compositions, and the further dispersion of the price (transformed) compositions from that of the value (untransformed) compositions (Marx 1894, chs 9, 45). Much of the subsequent debate surrounding the relation between values and prices of production (the Transformation Problem) has in fact centred around the complexity of the latter set of differences, with the dominant position being that such considerations effectively negate Marx's original formulations (Steedman 1977, chs 1–2). Yet recent work shows that the empirical differences between Marx's prices of production and the conventional (Bortkiewicz–Sraffa) 'correct' ones are generally very small, that both are good predictors of actual market prices (as are labour values also, all with R^2's between 93 and 96 per cent), and that there are sound mathematical reasons why the basic value categories dominate the overall results – as Marx quite correctly perceived from the start (Shaikh 1984; Ochoa 1984).

The second domain in which the composition of capital plays a central role is in the maintenance of a reserve army of labour. Marx points out that while the accumulation of total capital $c + v$ increases the demand for labour, the attendant growth in the value composition of capital *c/v* in turn decreases the demand for labour. Where the net effect is negative, the reserve army grows. And where it is positive, the resulting shrinkage in the reserve army eventually puts pressure on the labour market and accelerates the growth in real wages. This rise in real wages then slows down accumulation on one hand, while on the other it accelerates the pace of mechanization and hence the growth of *c/v*. In this way, the growth of the value composition automatically adjusts so as to maintain a reserve army of labour. When capitalism is viewed on the world scale, this phenomenon assumes great significance.

The third, and perhaps most important application of the concept of the composition of capital arises in connection with what Marx calls the 'one of the most striking phenomena of modern production', which is the tendency of the rate of profit to fall. The central variable in this case is the stock/flow materialized composition of capital *C/l*, because any sustained rise in *C/l* can be shown to give rise to an actual falling rate of profit, *no matter how fast the rate of surplus value is rising*. Writing the rate of profit r in terms of s, v, $l = v + s$, and $C =$ total (constant and circulating) capital advanced, we get

$$r = \frac{s}{C} = \frac{s/v}{C/v} = \frac{s/v}{(C/l)(l/v)} = \frac{s/v}{1 + (s/v)} \frac{1}{(C/l)}. \quad (7)$$

It is evident from equation (7) that as the rate of surplus value rises, the term $s/l = (s/v)/(1 + s/v)$ rises at an ever decreasing rate, since in the limit it approaches 1. Thus, no matter how fast the rate of surplus value rises, the rate of profit eventually falls at a rate asymptotic to the rate of fall of *l/C* (Rosdolsky 1977, chs. 16, 17, 26 and part V, appendix).

But the matter does not end there, because this issue recently sparked a fresh round of debates. On one side was an argument based on the (essentially neoclassical) theory of perfect competition, in which capitalists are assumed to invest in new methods only if these raise their own rate of profit, on the grounds that they would otherwise prefer to continue using their existing plant and equipment; and on the opposite side, an argument based on Marx's notion of competition-as-war, in which capitalists are driven to invest in those methods which lower their unit production costs, because the first ones to do so can cut prices and thereby expand their total profits through larger market shares. In the former case, the result is that the general rate of profit will necessarily rise, other things being equal; in the latter, the general rate of profit will tend to fall (as outlined above), provided that the new methods generally embody higher unit fixed costs.

In the original debates, the focus was on the differing implications of two apparently contradictory investment criteria; profit rate maximizing versus unit cost minimizing (profit margin maximizing). However, a subsequent contribution by Nakatani effectively dissolved this apparent opposition by showing that *both criteria are equivalent to selecting the highest projected rate of profit*. The principal difference then arises from the fact that in the case of perfect competition it is assumed that firms neither anticipate nor engage in price-cutting behaviour, while in the case of competition-as-war, firms are assumed to necessarily do both (Nakatani 1979). With this step, the issue reverts back to the two opposing conceptions of capitalism which lie behind these different notions of competition.

See Also

► Value and Price

Bibliography

Fine, B., and L. Harris. 1976. Controversial issues in Marxist economic theory. In *Socialist register*, ed. R. Miliband and J. Saville. London: Merlin.

Marx, K. 1858. *Grundrisse*. London: Penguin.

Marx, K. 1863. *Theories of surplus value, Part I*. Moscow: Progress Publishers.

Marx, K. 1867. *Capital*, vol. I. London: Penguin, 1976.

Marx, K. 1894. *Capital*, vol. III. New York: Vintage, 1981.

Nakatani, T. 1979. Price competition and technical choice. *Kobe University Economic Review* 25: 67–77.

Ochoa, E. 1984. Labor values and prices of production: An inter-industry study of the US economy, 1947–1972. PhD dissertation. Graduate Faculty, New School for Social Research.

Pratten, C.F. 1971. *Economies of scale in manufacturing industry*. Cambridge: Cambridge University Press.

Rosdolsky, R. 1977. *The making of Marx's capital*. London: Pluto Press.

Shaikh, A. 1978. Political economy and capitalism: Notes on Dobb's theory of crisis. *Cambridge Journal of Economics* 2(2): 233–251.

Shaikh, A. 1984. The transformation from Marx to Sraffa. In *Ricardo, Marx, Sraffa*, ed. E. Mandel. London: Verso.

Steedman, I. 1977. *Marx after Sraffa*. London: New Left Books.

Weeks, J. 1981. *Capital and exploitation*. Princeton: Princeton University Press.

Weston, J.F., and E.F. Brigham. 1982. *Essentials of managerial finance*, 6th ed. Chicago: Dryden Press.

Organization of the Petroleum Exporting Countries (OPEC)

James L. Smith

Abstract

Since the 1960s, the Organization of the Petroleum Exporting Countries (OPEC) has dominated the world oil market by exercising physical control over a large portion of the world's oil reserves. Coordinated production restraint among OPEC members has artificially limited the supply of oil and succeeded in pushing oil prices far above the competitive level. Despite its past success, OPEC faces three basic problems that, in the long run, tend to undermine all cartels: coordination failures, opportunistic cheating, and the entry of competing producers who manage to find and bring alternative supplies to the market.

Keywords

Barriers to entry; Cartels; Cheating; Coordination; Cournot oligopoly; Entry; Free-rider problem; Organization of Petroleum Exporting Countries (OPEC); Prisoner's Dilemma; Stackelberg dominant-firm models

JEL Classification

F33

The Organization of the Petroleum Exporting Countries (OPEC), an international cartel of oil-producing states, affects the price of nearly all crude oil traded in the world economy and has done so since the early 1970s.

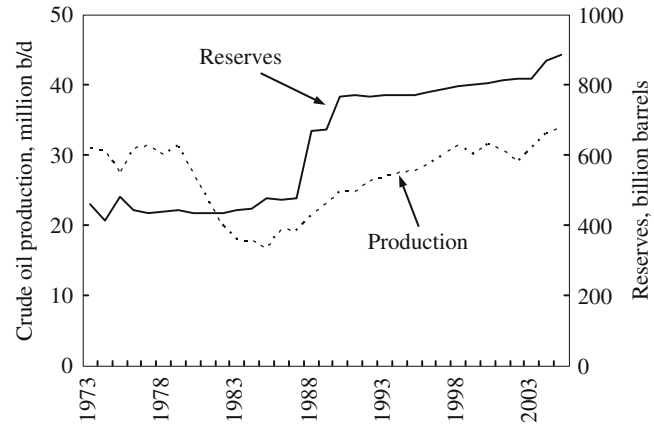
Founded in 1960, OPEC initially consisted of five member states (Iran, Iraq, Kuwait, Saudi Arabia and Venezuela) which together accounted for 38% of total world production of crude oil. The founders sought to coordinate national petroleum policies and forge a more united front in dealings with the multinational oil companies that operated within their borders. Although membership has grown to 12, OPEC's share of global crude oil production still amounts to only about 44%. Coordinated restraints on output (especially since 1973) have deliberately held OPEC's market share in check.

During its first decade (1960–1970), OPEC's principal objective was to secure for its members a larger share of the profits derived from the production and sale of their oil – the stated goal being to raise government take from 50% to 80% of total profit. Beginning with the so-called Teheran–Tripoli Agreements of 1970–1971, OPEC turned to what has become its main purpose: manipulating the level of world oil prices by restricting productive capacity and output. Initially, this was attempted without assigning individual production quotas to the respective members. Only after the downturn in world oil prices that began in 1982 did OPEC introduce a formal system of production allocations – which remained in force as of 2007. The members meet at regular intervals (and sometimes on an emergency basis) to review market conditions and adjust individual production ceilings as needed to maintain a target price. Adelman (1995) and Parra (2004) describe the intriguing economic and political challenges faced by the members of OPEC in dealing with the market and with each other.

There is no question that OPEC members have restricted production in ways that are unrelated to the physical scarcity of oil. Even though OPEC's proved oil reserves in 2007 were double those of 1973, the cartel initiated sharp output cuts that by 1985 had removed nearly half of their previous production from the market, as shown in Fig. 1. Not until 2005 did OPEC production regain (barely) the level of 1973. Over that same period, worldwide consumption of crude oil grew by 50% and production from non-OPEC producers (who faced much higher marginal costs) managed to increase by 70%.

Economic Models of OPEC Behaviour

Early economic analyses of OPEC behaviour questioned whether the output reductions might reflect competitive or other forms of



Organization of the Petroleum Exporting Countries (OPEC), Fig. 1 OPEC Production and reserves, 1973–2005 (Sources: Production, U.S. Energy Information Administration. Reserves, *Oil & Gas Journal*)

non-cooperative conduct (for example, oligopoly), as opposed to outright collusion. Mead (1979) and Johany (1980) proposed a 'property rights' explanation that linked the production cuts to the wave of nationalizations that swept through the global oil industry in the early 1970s. Property rights in oil reserves were transferred, via nationalizations, from the multinational corporations (with higher presumed discount rates) to OPEC states (with lower presumed discount rates and therefore greater patience in extracting the oil). However, this explanation is belied by the fact that, throughout the 1960s, these same host governments had repeatedly exhorted the multinational companies to increase, not decrease, their rates of production (Adelman 1982).

Teece (1982) and Crémer and Salehi-Isfahani (1980) advanced the idea that the limited domestic revenue needs ('absorptive capacity') of some OPEC members imposed an indirect restriction on production. The higher the price, the lower the volume of oil exports required to achieve a requisite amount of revenue. The result would be a backward-bending supply curve that links lower oil output to higher prices in a manner that implies no coordination among OPEC members. One problem with this argument, as Adelman (1982) pointed out, is that the absorptive capacities of OPEC members seemed to increase faster than export revenues. Griffin's (1985) subsequent empirical tests found little statistical support for the target revenue hypothesis.

Distinguishing between the various models of OPEC behaviour has been complicated by the fact that cooperative and non-cooperative models share many similar predictions. Thus, the same body of evidence has been interpreted in ways that are consistent with a variety of competing models. By focusing on one aspect of producer behaviour (short-run reactions to cost shocks) that more clearly distinguishes between models, Smith (2005) found a degree of parallelism among OPEC producers that can be accounted for only as the result of cooperative behaviour, not competition or mere interdependence among producers, as in the Cournot oligopoly or Stackelberg dominant-firm models.

Future Challenges Facing OPEC

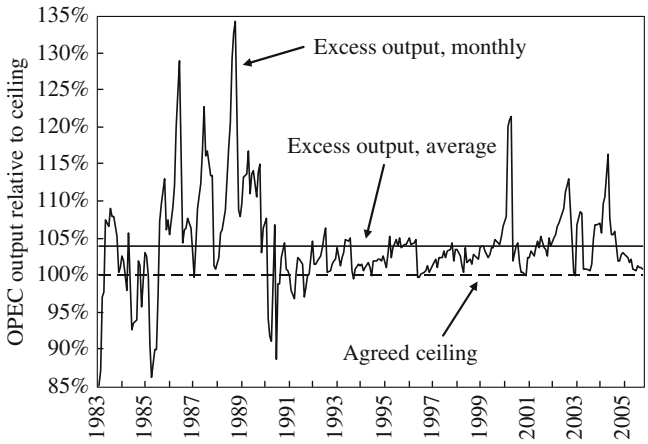
Levenstein and Suslow (2006) identify three critical problems that any cartel must solve if it is to endure: coordination, cheating and entry. In the case of OPEC, the last of these has been the easiest. OPEC is protected by barriers to entry that stem from ownership and control of low-cost oil reserves. Roughly 75% of the world's proved reserves of crude oil are located in OPEC nations. Additional reserves are discovered and developed each year, but this process has become increasingly difficult and expensive – even more so outside OPEC than within. Thus, production of crude oil from non-OPEC sources does expand when the cartel cuts production and pushes prices up, but the scope for this is limited and will remain so.

The problem of cheating has been more difficult for OPEC. Any system of output restraints is vulnerable to the free-rider problem. Although OPEC as a whole may benefit by restricting total output, individual members are tempted to produce beyond their assigned quotas. Cartel membership is most beneficial to those members who do *not* cut production. Without a system to detect and punish cheating, the cartel is hampered by a Prisoner's Dilemma in which the dominant strategy for most, if not all, members is to ignore their assigned quotas.

It is common, as in Gately (2004), to distinguish between 'core' (low cost, high compliance) and 'non-core' (high cost, low compliance) members of OPEC. In fact, compliance with the quota by members of both groups has been sporadic, as shown in Fig. 2. Since the inception of the formal quota system in 1983, total OPEC production of crude oil through 2005 has exceeded the ceiling by 4% on average, but on numerous occasions the excess has run to 15% or more. In general, full compliance has been achieved only during episodes (like 2005–2006) when the production ceiling itself tested the limits of each member's available production capacity, such that cheating was not feasible.

The third problem – coordination among members – presents further difficulties. Due to economic and demographic heterogeneity, the interests of individual OPEC members do not naturally align behind a single 'correct' price or production target. In part this is due to the fact that OPEC has limited means by which to redistribute earnings among members. Therefore, any given set of quotas determines not only the overall profit of OPEC but also the individual revenues that accrue to each member. Moreover, coordination requires agreement not only about how aggregate output is parcelled out to individual members, but also about the amount of oil to be produced by OPEC in total. Members with low-cost, long-lived reserves may be more reluctant to have OPEC pursue severe output cuts since too-high prices would induce technological development and new forms of energy (or energy conservation) that will eventually compete with OPEC. Members that possess smaller reserves and shorter horizons are less affected by this and may prefer deeper production cuts. Internal divisions between 'price hawks' and 'price doves' have been observed previously and will likely surface within OPEC again.

In terms of longevity, OPEC is already far beyond the mean lifetime (5 years) of contemporary international cartels (Levenstein and Suslow 2006). In terms of economic impact, it is sufficient to note that crude oil is among the most valuable commodities exchanged in international trade, with total daily receipts in 2007 in excess of $1 billion. Thus, by exerting even a small impact on the market price, the cartel effects an enormous transfer of wealth between consumers and producers of crude oil, and creates a substantial allocative inefficiency of the type that arises whenever the price of a product deviates from its marginal cost. As of 2007, no one has attempted to reckon the full



Organization of the Petroleum Exporting Countries (OPEC), Fig. 2 OPEC compliance with the production ceiling, 1983–2005 (Sources: Ceilings, *OPEC Annual Statistical Bulletin*. Actual production, U.S. Energy Information Administration)

magnitude of welfare losses that may be associated with OPEC's manipulation of the world oil market.

See Also

- ▶ Cartels
- ▶ Concentration Measures

Bibliography

Adelman, M.A. 1982. OPEC as a cartel. In *OPEC behavior and world Oil prices*, ed. J.M. Griffin and D.J. Teece. London: George Allen and Unwin.

Adelman, M.A. 1995. *The genie out of the bottle: World oil since 1970*. Cambridge, MA: MIT Press.

Crémer, J., and D. Salehi-Isfahani. 1980. *A theory of competitive pricing in the oil market: What does OPEC really do?* Working Paper No. 80–4, CARESS, University of Pennsylvania.

Gately, D. 2004. OPEC's incentives for faster output growth. *The Energy Journal* 25(2): 75–96.

Griffin, J.M. 1985. OPEC behavior: A test of alternative hypotheses. *American Economic Review* 75: 954–963.

Johany, A.D. 1980. *The myth of the OPEC cartel*. New York: Wiley.

Levenstein, M.C., and V.Y. Suslow. 2006. What determines cartel success? *Journal of Economic Literature* 44: 43–95.

Mead, W.J. 1979. The performance of government in energy regulation. *American Economic Review* 69: 352–356.

Parra, F. 2004. *Oil politics: A history of modern petroleum*. London: I.B. Taurus.

Smith, J.L. 2005. Inscrutable OPEC? Behavioral tests of the cartel hypothesis. *The Energy Journal* 26(1): 51–82.

Teece, D.J. 1982. OPEC behavior: An alternative view. In *OPEC behavior and world oil prices*, ed. J.M. Griffen and D.J. Teece. London: George Allen & Unwin.

Organization Theory

Thomas Marschak

Since all the social sciences deal with human organizations (families, bureaucracies, tribes, corporations, armies), the term 'organization theory' appears in all of them. What has distinguished the economists' pursuit of organization theory from that of sociologists, of political scientists and of psychologists (say those psychologists working in the field called 'organizational behaviour')? First, the real organizations that have inspired the theorizing of economists are the economy, the market and the firm. Second, economists, with their customary taste for rigour, have sought to define formally and precisely the vague terms used in informal discourse about organizations, in such a way as to capture the users' intent. They have sought to test plausible propositions about organizations – either by proving that they follow from simple, reasonable and precisely stated assumptions, or (rarely) by formulating the propositions as statements about observable variables on which systematic rather than anecdotal data can be collected, and then applying the normal statistical procedures of empirical economics. (Here we shall only consider testing of the first type). Third, much of the economists' organization theory is not descriptive but normative; it concerns not what is, but what could be. It takes the viewpoint of an organization *designer*. The organization is to respond to a changing and uncertain environment. The designer has to balance the 'benefits' of these responses against the organization's *informational costs*; good responses may be costly to obtain. In addition, the designer may require the responses to be *incentive-compatible*: each member of the organization must *want* to carry out his/her part of the total organizational response in just the way the designer intends.

The design point of view has old and deep roots in economics. Adam Smith's 'invisible hand' proposition is a statement about the achievements of markets as resource-allocating devices. If one reinterprets it as a comparative conjecture about alternative designs for a resource-allocating organization – namely, that a design using prices is superior to other possible designs – then it becomes an ancestor of the organization-design point of view. In any case, that point of view appears very clearly in Barone's 'The Ministry of Production in the Collective State' (1908), and in the debates about 'the possibility of socialism' (i.e., of a centrally directed

economy) in the 1930s and 1940s (Hayek 1935; Lange 1938; Dobb 1940; Lerner 1944).

Nearly all the debaters agreed that if the designer of resource-allocating schemes for an economy has a clean slate and can construct any scheme at all, then he must end up choosing some form of the price mechanism; for example, a scheme of the Lange–Lerner sort. Here a Centre announces successive trial prices; in response to each announcement, profit-maximizing demands are anonymously sent to the Centre by managers, and utility-maximizing demands are sent by consumers; in response to the totals of intended demands, the Centre announces new prices; the final announced prices are those which evoke zero excess demands, and the corresponding intended productions and consumptions are then carried out. The debate dealt largely with the informational virtues of such a price scheme as compared to an extreme centralized alternative scheme. The alternative scheme (never made very explicit) appears to be one wherein managers and consumers report technologies, tastes and endowments to the Centre, which thereupon computes the economy's consumptions and productions; those become commands to be followed.

In retrospect, the extreme centralized alternative seems an unimaginative straw man, since one can imagine a whole spectrum of designs lying between extreme centralization, on the one hand, and the price scheme, on the other; namely, designs in which some of the agents' private information is centrally collected (or pooled), but not all of it. In any case, the debaters agreed that the price scheme is informationally superior to the centralized alternative because (1) in the former, small computations are performed simultaneously by very many agents (though possibly many times), whereas in the latter an immense central computation is required (though required only once), and (2) the messages required in the former (prices and excess demands) are small (though sent many times) while in the latter a monstrously large information transmission is required (though only once).

Persuasive as this claim may appear, a moment's thought reveals how very many gaps need to be filled before the claim becomes provable or disprovable. If a proposed scheme is to be operated afresh at regular intervals (in response, say, to new and randomly changing tastes, technologies and endowments), then what is the designer's measure of a proposed allocation scheme's gross performance (against which a scheme's cost must be balanced)? Is it, for example, the expected value of the gross national product in the period which follows each operation of the scheme? Or is it perhaps a two-valued measure which takes the value one when the scheme's final allocation is Pareto-optimal and individually rational (i.e., every consumer ends up with a bundle at least as good as his/her endowment) and takes the value zero otherwise? When is the scheme to be terminated if it comprises a sequence (possibly infinite) of steps? What interim action (resource allocation) is in force while the proposed scheme is in operation and before it yields a final action? For alternative investments in information-processing facilities, how long does the sequence's typical step take? (The longer a step takes, the longer one waits until a given terminal step is reached and the longer an unsatisfactory interim action is in force.)

Once such gaps are filled in, the claim becomes, in principle, a verifiable conjecture. Without venturing to fill them in, economists were nevertheless sufficiently intrigued by the intuitive (but quite unverified) informational appeal of the Lange–Lerner scheme so that they proceeded to construct many more schemes of a similar kind in a variety of settings, including multidivisional firms, for example, as well as planned economies with technologies less well behaved than the classic (convex) ones (see Heal 1986). These efforts were partly stimulated by (and, in turn, stimulated) the development of algorithms for general constrained optimization, which often had a natural interpretation as schemes wherein a 'Centre' makes announcements and other 'persons' respond without directly revealing their private information. (One can so interpret, for example, certain gradient methods for constrained optimization, as well as the 'decomposed' version of the simplex algorithm for linear programming.)

If the informational appeal of schemes of the Lange–Lerner type was powerful but unverified, what of the incentive side? Here the 'possibility-of-socialism' writers were divided. A sceptic like Hayek (1935, pp. 219–20) asked why a manager would want to follow the Lange–Lerner rules. One (unsupported) reply – hinted at in various places in the debate – is that to induce a manager to follow the rules we need only pay him a reward which is some nondecreasing function of his enterprise's profit. The incentive question becomes acute when one turns to the scheme that is the analogue of the Lange–Lerner scheme if there are public goods; namely, the Lindahl scheme (Lindahl 1919), when that is given a central-price-announcer interpretation. (The scheme was developed before the possibility-of-socialism debates but appears to have been unknown to the debaters). For here, as Samuelson (1954) was the first to note, the prospective consumer of a public good may perceive an advantage in falsifying his demand for it; that is, in disobeying the designer's rules. (In fact, it turned out later (Hurwicz 1972) that the same difficulty can arise without public goods; that is, in the original Lange–Lerner scheme itself). It took about three decades after the possibility-of-socialism debate until one had the framework to study with precision the question of when incentive-compatible schemes of the price-announcer type – or indeed of any type – can be constructed for economies or for organizations in general.

On the informational side of the design question, a 1959 paper by Hurwicz (Hurwicz 1960) proved to be a major step towards precise conjectures (as opposed to broadly appealing but unverifiable claims) about the informational merits of alternative resource-allocating schemes for economies, or indeed alternative designs for organizations in general. The key notion is that of an *adjustment process*, to be used by an n-person organization confronting a changing environment $e = (e_1, \ldots, e_n)$, lying always in some set E of possible environments. Here e_i is that aspect of the environment e observed by person i. Assume that the possible values of e_i comprise a set E_i and that $E = E_1 \times \ldots \times E_n$. If, for example, the organization is an exchange economy, then e_i is composed of i's endowment and i's preference ordering on alternative resource allocations; if $n = 2$, then E might be the set of classic Edgeworth-box economies. An adjustment process is a quadruple, $\pi = (M, m_0, f, h)$,where M is a set called a *language* and is the cartesian product of n *individual language* M_i ; f is an n-tuple $(f_1, \ldots, f_n)$; f_i is a function from $M \times E_i$ to M_i; $m_0 = (m_{01},\ldots,m_{0n})$ is an *initial* message n-tuple in M; h is a function, called the *outcome function*, from $M \times E$ to A; and A is a set or organization *actions* or *outcomes* (e.g., resource allocations). Imagine the environment to change at regular intervals. Following each new environment, person i emits the initial message $m_{1i} = f_i(m_0,\varepsilon_i)$ in $\boldsymbol{M}_i$. At step 1, person i emits the message $m_{1i} = f_i(m_0,\varepsilon_i)$ in $\boldsymbol{M}_i$ and at the typical subsequent step t, person i emits $m_{ti} = f_i(m_{t-1}, \varepsilon_i)$, where $m_{ti} \in M_i$ and m_{t-1} denotes an element of M; namely,$(m_{t-1},1,\ldots,m_{t-1,n})$. At a terminal step T, the organization takes the action (or puts into effect the outcome) $h(m_T, e)$ in A which is its final response to the environment e. The process is *privacy-preserving* in the sense that e enters i's function f_i only through e_i which is i's private knowledge. One might require a similar property for h, that is, that h be an n-tuple $(h_1,\ldots,h_n)$ where h_i is a function from $M \times E_i$ to a set A_i of possible values of i's *individual action* (thus A is the cartesian product $A_1 \times \ldots \times A_n$). In the useful special case of a 'non-parametric' outcome function, where h does not depend on e at all, such privacy-preservation for action selection holds trivially.

Note that we can endow person i with a memory. To do so, let every element m_i of the set M_i be a *pair* (m_i^*, m_i^{**}), where m_i^* denotes memory and m_i^{**} denotes a message sent to (noticed by) others; specify that for $k \neq i$, f_k is insensitive to (its value does not depend on) m_i^*. By making the set in which m_i^* lies sufficiently large, we can let i remember, at every step, all that he has observed of the organization's messages thus far. We can, moreover, let i send messages always to j and to no one else by specifying that for $k \neq i$, $k \neq j$, f_k is insensitive to the ith component of m. We can let i send a message to j and to no one else *at some specific step* t^* by specifying that when all persons'

memories tell them that t^* has been reached, then for $k \neq i, k \neq j, f_k$ is insensitive to the ith component of m.

The adjustment process, as the object to be chosen by the designer, is a concept sufficiently broad and flexible to accommodate all the economists' iterative resource allocation schemes for economies as well as a rich variety of designs for other organizations. All organizations, after all, respond to a changing environment of which each member observes only some aspect in which he/she is the specialist, and the environment's successive values are unknown to the designer when a design is to be chosen. If those values *were* known (e.g., if the environment were constant), then there would be no need for message exchanges at all: each member could simply be programmed once and for all to take a correct (a best) action or sequence of actions. In all organizations, moreover, members engage in dialogue that evèntually yields an organizational response to the current environment (an action).

With regard to the classic claim that price schemes are informationally superior designs when the organization is an economy, the adjustment-process concept has permitted a first rigorous test. The test takes the view that we can (as a reasonable starting place) ignore the pre-equilibrium performance of a price scheme (formulated as an adjustment process), and can focus entirely on its *equilibrium* achievements. For any e in E let M^{e} denote the set of *equilibrium messages*; that is, every $m^{\mathrm{e}} = \left(m_i^e, \ldots, m_n^e\right)$ in M^{e} satisfies $f_{\mathrm{i}}(m^e, e_{\mathrm{i}}) = m_1^e$ for all i. Confine attention to processes with non-parametric outcome functions h (i.e., h depends only on m, not on e) and, for the case where E is a set of exchange economies, formulate the competitive (the Walrasian) mechanism as a non-parametric process, say $\pi^* = \left(M^*, m_0^*, f^*, h^*\right)$. The typical element m of M^* comprises a vector of proposed prices and an $(n-1)$-tuple of proposed trade vectors; f_i yields i's intended trade vector – or, in an alternative version, a *set* of acceptable trade vectors – at the just-announced prices; and h is a projection function yielding the 'trade' portion of m. For the process π^* and for every e in a classical set E, all the *equilibrium outcomes for e* – that is, all those allocations (trade $(n-1)$-tuples) a satisfying $a = h^*(m)$ for all m in M^{*e} – are Pareto-optimal and individually rational. One now asks the following question: does there exist any other process $\pi = (M, m_0, f, h)$ such that (i) for all e in the same set E every equilibrium outcome is again Pareto-optimal and individually rational, and (ii) the process π is informationally 'cheaper' than π^*? A natural starting place for the assessment of informational cost is size of the language. If one confines oneself to processes π in which M is in a finite Euclidean space, then a natural measure of language size is dimension. But then the question just posed has a trivial Yes as its answer, since one can always code a message of arbitrary dimension as a one-dimensional message. To rule out such coding, one imposes 'smoothness' on the process π. For example, one considers the mapping t from A (the set of outcomes), to the subsets of E, such that for every e in $t(a)$, a is an equilibrium outcome for e, and one requires that t contain a Lipschitzian selection. It turns out that for classic sets E and for language dimension as the cost measure, no smooth process satisfying (i) and (ii) exists (Hurwicz 1972). The result extends (for more general sorts of smoothness requirements) to processes with non-Euclidean languages and language-size measures more general than dimension (Mount and Reiter 1974; Walker 1977; Jordan 1982).

These results are clearly a first step towards vindicating the classic claim that the price process is informationally superior. To go further, one would like to consider pre-equilibrium outcomes – so that the final allocation is the one attained at a fixed, but well-chosen, terminal step – and to take account of the change in the time required to reach that terminal step as one varies the investment in the information-processing facilities available for carrying out the typical step. It seems plausible that a version of the competitive process that converges rapidly to its equilibrium messages will rank high relative to other processes once this complication is added. One would like the 'smoothness' requirement to arise naturally from a model of a wellbehaved information technology rather than being introduced (as at present) in an ad-hoc manner. One would like to leave the

O

setting just sketched, wherein messages and outcomes are points of a continuum, to see whether analogous results hold when both messages and outcomes (allocations) have to be rounded off to a chosen precision. (A limited analogue of the dimensional-minimality result just sketched has in fact been obtained in such a discrete setting (Hurwicz and Marschak 1985).

For organizations in general, the requirements of Pareto-optimality and individual rationality are replaced by some given set of desired (and equally acceptable) responses to every possible given environment. The problem facing a designer who is unconcerned about incentive aspects can then be put as follows. Given a set E and a *desired-performance correspondence* ϕ from E to the subsets of an outcome (action) set A, find an adjustment process $\pi = (M, m_0, f, h)$ which *realizes* ϕ that is, which satisfies $a \in \phi(e)$ if $a = h(m, e)$ and $m \in M^e$ – and *whose* informational costs (suitably measured) are no less than those of any other process which realizes ϕ.

Note that a far more ambitious task could be given the designer instead. Let the designer have preferences over alternative environment/outcome/cost triples and let the preferences be represented by a utility function. The ambitious task is then to find a process π, and an accompanying selection function, which chooses a unique equilibrium outcome in the set M^e for every e, so as to maximize the designer's expected utility (expectation being taken with respect to the random variable e). It seems clear that such unbounded designer's rationality is too ambitious a standard; organization theory would freeze in its tracks if it adopted such a standard. The realization of a given performance correspondence at minimum informational cost is a reasonable step towards bounded rationality, especially if the performance correspondence is not stringent. (Thus ϕ might assign to e all outcomes which are within a certain specified distance of an outcome that is 'ideal' for e – say an outcome that maximizes some pay-off function).

The preceding bounded-rationality version of the designer's task can again be modified by allowing some 'dynamics'; that is, permitting choice of terminal step rather than focusing on equilibrium outcomes. Whether we do so or not, we now have a precise version of the general performance-versus-cost problem which we claimed at the start to be a distinctively 'economists', contribution to organization theory. (The problem is surveyed in more detail in Marschak 1986).

When one turns to incentive issues, a certain 'contraction' of the adjustment-process concept has proven useful. The object chosen by the designer now becomes a *game form* (S, g), where $S = S_1 \times \ldots \times S_n$; S_i is the set of person i's possible *strategies* s_i; and g is an outcome function from S to A (the set of organizational actions or outcomes). Person i's local environment e_i specifies (among other things) i's preferences over the alternative organizational outcomes. The set of Nash-equilibrium strategy n-tuples $s = (s_1,\ldots,s_n)$ such that given $e = (e_1,\ldots,e_n)$ each person i regards the outcome $g(s)$ to be at least as good as the outcome, $g(s_1, \ldots, s_{i-1}, \bar{s}_i, s_{i+1}, \ldots, s_n)$ for all $\bar{s}_i$ in S_i. Suppose the designer is again given a desired-performance correspondence ϕ from E to the subsets of A. Then the incentive problem may be put this way: find a game form (S, g) such that for every e in E and every s in $N_{sg}(e)$, the outcome $g(s)$ is contained in the set $\phi(e)$. Such a game form *Nash-implements* ϕ. We can trivially find an adjustment process(M,m_0,f,h) whose equilibrium outcomes for every e comprise exactly the set $\{a{:}a = g(s); s \in N_{sg}(e)\}$ (To do so, let $M = M_1 \times \ldots \times M_n = S_1 \times \ldots \times S_n$; let f_i satisfy $f_i()(s_1,\ldots,s_n),e_i) = s_i$ if and only if, given e_i, i regards the outcome $g(s)$ to be at least as good as the outcome $g(s_1, \ldots, s_{i-1}, \bar{s}_i, s_{i+1}, \ldots, s_n)$ for all $\bar{s}_i$ in S_i; and let $h(s) = g(s)$.) Much has now been learned about what sorts of performance functions ϕ (including economically interesting ones) can be implemented and what sorts cannot (for a survey, see Hurwicz 1986). We again have the 'dynamic' shortcoming noted before: if, for every e, an outcome in the set $\{a \in N_{gs}(e){:}s \in S\}$ is indeed to be reached by operating an adjustment process (as in the economists' allocation mechanisms), then the behaviour of the process prior to equilibrium must be studied. Doing so may, moreover, introduce quite new strategic considerations, since a fresh incentive problem may arise at each step of the process: at

each step a member may ask whether carrying out the designer's instructions (applying f_i) is what he/she really wants to do.

Thus both on the informational and the incentive sides, a very large research agenda stretches before the economic organization theorist. Moreover, the abstract theorizing we have sketched is very far indeed from making good contact with the institutional facts about real organizations. One may take the design point of view, but even a designer is constrained by those facts.

In particular, the notion of *hierarchy* (the 'organization chart'), which appears so often in popular discourse, is very hard indeed to pin down in the adjustment-process framework. To define 'hierarchy', we first have to define 'authority'. When does an adjustment process have the property that person 1 is in authority over person 2? Probably the best one can hope for (Hurwicz 1971) is this: person 1 is in authority over person 2 if (1) at the terminal step T,m_{T2} depends only on $m_{T-1,1}$, and (2) $m_{T-1,1}$ is sensitive to $e1$. If we did not add requirement (2), then person 1's apparent terminal instruction to person 2 (embodied in the pre-terminal message $m_{T-1,1}$) might in fact be a robot-like repetition (perhaps in recoded form) of a 'command' that 2 gave to 1 at step $T-2$. On the other hand, we might satisfy the sensitivity required by (2) in such a trival way that we have not really succeeded in ruling out person 2 as the 'true' (though somewhat disguised) commander. Authority is, in short, a very fragile concept from a formal point of view.

Yet it is a central concept in influential writings like those of Williamson (1975). His book is a rich source of institutionally motivated conjectures about how organizations work, but it teems with terms, concepts and conjectures that the formal theorist must struggle mightily to make precise. The task of precise pinning down is so daunting that the stage of testing the conjectures (trying to prove them) seems unlikely to be reached. The book argues for these conjectures nevertheless, and many of them appear, at some level, to be plausible. Here is one example: 'it is elementary that the advantages of centralization vary with the degree of independence among the members, being ... almost certainly great in an integrated task group' (p. 51). To the formal theorist, that is not 'elementary' at all. One requires five or six definitions before one even knows what is being claimed.

Nevertheless, such informal but insightful institution-based essays are an essential challenge to formal theory. The economists' organization theory of the future will grow out of the tension between highly imprecise but widely believed and institutionally grounded claims and the harsh demands of formal argument.

See Also

- Decision Theory
- Efficient Allocation
- Exchange
- Game Theory
- Rank

Bibliography

Barone, E. 1908. The ministry of production in the collectivist state. In *Collectivist economic planning*, ed. F.-A. von Hayek. London: Routledge, 1935, 245–290.

Dobb, M.H. 1940. *Political economy and capitalism*. New York: Macmillan.

von Hayek, F. (ed.). 1935. *Collectivist economic planning*. London: Routledge.

Heal, G. 1986. Planning. In *Handbook of mathematical economics*, vol. III, ed. K.J. Arrow and M.-D. Intriligator. Amsterdam: North-Holland.

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

Hurwicz, L. 1971. Centralization and decentralization in economic processes. In *Comparison of economic systems*, ed. A. Eckstein. Berkeley: University of California Press.

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

Hurwicz, L. 1972b. On the dimensional requirements of informationally decentralized Pareto-satisfactory processes. In *Studies in resource allocation processes*, ed. K.J. Arrow and L. Hurwicz. Cambridge: Cambridge University Press, 1977.

Hurwicz, L. 1986. Incentive aspects of decentralization. In *Handbook of mathematical economics*, vol. III, ed. K.-J. Arrow and M.D. Intriligator. Amsterdam: North-Holland.

Hurwicz, L. and Marschak, T. 1985. Discrete allocation mechanisms: Dimensional requirements for resource-allocation mechanisms when desired outcomes are unbounded. *Journal of Complexity*.
Jordan, S.J. 1982. The competitive allocation process is informationally efficient uniquely. *Journal of Economic Theory* 28: 1–18.
Lange, O. 1936–7. On the economic theory of socialism. In *On the economic theory of socialism*, ed. B. Lipincott. Minneapolis: University of Minnesota Press, 1938.
Lerner, A.P. 1944. *The economics of control*. New York: Macmillan.
Lindahl, E. 1919. Just taxation: A positive solution. In *Classics in the theory of public finance*, ed. R. Musgrave and A. Peacock. London: Macmillan, 1958.
Marschak, T. 1986. Organization design. In *Handbook of mathematical economics*, vol. III, ed. K.J. Arrow and M.D. Intriligator. Amsterdam: North-Holland.
Mount, K., and S. Reiter. 1974. The informational size of message spaces. *Journal of Economic Theory* 8(2): 161–192.
Samuelson, P.A. 1954. The pure theory of public expenditure. *Review of Economics and Statistics* 36: 387–389.
Walker, M. 1977. On the informational size of message spaces. *Journal of Economic Theory* 15(2): 366–375.
Williamson, O.E. 1975. *Markets and hierarchies, analysis and antitrust implications: A study in the economics of internal organizations*. New York: Free Press.

Ortes, Giammaria (1713–1790)

Ugo Rabbeno

Keywords

Division of labour; Free exchange; Malthus, T. R.; Mathematical method; Money; Ortes, G.; Population growth; Wealth

JEL Classifications

B31

A Venetian monk, Ortes left his cloister on the entreaties of his mother after his father's death, but remained in holy orders and was ever a strenuous defender of the clergy. It is with this purpose that he wrote his *Errori popolari intorno al- l'Economia nazionale,* his *Lettere sulla religione* and his treatise *Dei Fide-commessi a famiglie e a chiese*, with the scope of upholding the existence of clerical property in Mortmain.

In his *Economia nazionale* (vols xxi, xxii, and xxiii, of Custodi's *Scrittori classici italiani di economia politica*, Milan, 1802–1816) Ortes endeavours to demonstrate that as

> the wealth of a nation is determined by the (previous) wants of its members, the riches of one of them cannot increase unless at the expense of another one; the bulk of existing riches is in each nation measured by its wants, and cannot by any means whatever exceed this measure. (*Discorso preliminare*)

From this rather startling proposition, Ortes, who certainly was an original thinker, deduces the condemnation of the principles on which mercantilism was based.

> Money is only a sign of wealth, and must never be considered as being wealth itself. The error of those who mistake money for wealth, proceeds from a confusion between the equivalent of a thing and the thing itself, or between two equivalents which they consider as identical things, although they are not. (ch. ix)

In his *Riflessioni sulla popolazione* (Venice, 1790, and vol. xxiv of Custodi) Ortes controverts the prevailing opinion that an increase of population must necessarily increase the wealth of a nation, and maintains that 'in any nation whatever the population is compelled to keep within fixed limits, which are invariably determined by the necessity of providing for its subsistence' (*Prefazione*). In his very first chapter he asserts that, if natural instincts were allowed full play, population would increase in a *geometrical* progression (doubling every 30 years), and calculates that a group of 7 persons composed of three old people, two young men and two young women of 20, would be the ancestors at the end of 150 years of 224 living persons.

150 years of	224 living persons
300 years of	7, 1688 living persons
450 years of	229, 376 living persons
900 years of	7, 516, 192, 768 living persons

Sheer violence keeps down the numbers of animals within the necessary limits, but among men, 'generation is limited by reason' (ch. iii), especially by voluntary celibacy, which affords Ortes an occasion of extolling the provident discipline of the Roman Catholic Church. Ortes is a harbinger of Malthus; first by his law of the geometrical increase of population, and secondly by the influence which he ascribes to human reason as a prudential check against overpopulation.

Ortes was a fervent mathematical student, and expresses himself in algebraical formulae in his *Calcolo sopra il Valore delle Opinioni umane* (vol. xxiv, Custodi). In the same work he illustrates his meaning by curves, which, if not actually traced, are at least minutely described.

Edward Cannan

Ortes is undoubtedly the most eminent of the Venetian economists of the 18th century; his genius, original and sometimes paradoxical, is often opposed to the general tendency of the ideas of his time, and though his researches are occasionally faulty in their method, he has left a deep impress on the history of economic theory. He regards economic laws as immutable, like those of nature; he maintains this in opposition to the opinion usually accepted in his time, which regarded economics only in relation to special interests. Perhaps it is this idea which leads him to distrust the action of the state, considering it is not adapted to promote the wealth of a country.

While Ortes applied a mathematical method to economics, his arguments are based throughout on abstract theory, disregarding the study both of facts and of history as not appertaining to economic science. This detracts from the value of his labours. Still his works are of weight in the history of economic theory. He did not adopt the doctrines of the Physiocrats, and he also recognizes the importance of division of labour, and the important place taken by production in economic theory. Contrary to the prevailing ideas of his day, Ortes upholds universal free exchange.

Selected Works

n.d. *Calcolo sopra il valore delle opinioni umane.*
1771. *Errori popolari intorno all'economia nazionale.*
1790. *Reflessioni sulla popolazione delle nazioni per rapporto all'economia nazionale.* Venice.
1802–16. Many works in *Scrittori classici italiani di economia politica*, ed. P. Custodi. Milan.

Ostrom, Elinor (1933–2012)

Paul Dragos Aligica and Peter Boettke

Abstract

Elinor Ostrom, a recipient of the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2009, had a foundational contribution to the Public Choice movement and to the rise of the new institutionalism and has been a key figure in the resurgence of political economy. Her studies of common pool resources, economic governance and institutional diversity are an attempt to transcend the 'markets vs. states' dichotomy and are marked by a distinctive approach, relying on multiple methods, interdisciplinary collaborative teamwork and the primacy of empirical observations in field and laboratory settings.

Keywords

Common pool resources; Economic governance; Institutionalism; Public choice; Public economies; Political economy

JEL Classifications

B31

Elinor Ostrom is Distinguished Professor and Arthur F. Bentley Professor of Political Science, Indiana University, Bloomington; Senior Research Director, Workshop in Political Theory and Policy Analysis, Indiana University, Bloomington; and Founding Director, Center for the Study of Institutional Diversity, Arizona State University, Tempe. She was born in Los Angeles in 1933 and received her Ph.D. in Political Science from UCLA in 1965. Ostrom is a recipient of the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2009, made a foundational contribution (together with her husband, and co-founder of the Bloomington Workshop in Political Theory and Policy Analysis, Vincent Ostrom) to the Public Choice movement (President of the Public Choice Society, 1982–1984), and has been associated with efforts leading to the resurgence of political economy in economics and political sciences (President of the American Political Science Association, 1996–1997). Also, she has been recognised as a key figure in the rise of the new institutionalism and as an influential advocate of a specific form of methodological pluralism that emphasises intensive empirical work and interdisciplinarity.

Ostrom's contribution has been complex, prolific and multifaceted, but several themes have gained widespread recognition for her work: her contribution to a better understanding of the nature of economic governance; the development of the notion of 'public economy' involving a challenge of the 'markets vs. states' dichotomy; her role in the metropolitan governance reform debate; the efforts to develop analytical frameworks for the study of action situations and institutional arrangements; and indeed, her contribution to the study of the 'commons' and self-governance. In addition, her distinctive approach to social research, relying substantially on the primacy of empirical observations in field and laboratory settings as well as collaborative teamwork, deserves a special note. This article will briefly outline some elements of these themes.

Beyond Hobbes and Smith

As Elinor Ostrom described it herself, her work is a systematic attempt to transcend the basic dichotomy of modern political economy. On the one hand, there is the tradition defined by Adam Smith's theory, focused on the pattern of order and the positive consequences emerging out of the independent actions of individuals pursuing their own interests within a given system of rules. On the other hand, there is the tradition rooted in Thomas Hobbes' theory, in which individual actors, pursuing their own interests and trying to maximise their welfare, behave in ways that lead inevitably to chaos and conflict. From that is derived the necessity of a single centre of power imposing order. These two theories were assumed to be able to answer all important questions. In this context, when confronted with a question such as 'how far the logic of market organisation can be applied to the organisation of productive activities beyond strictly private goods' the answer was given by introducing concepts such as market failure and by prescribing a centralised authority to provide for collective goods. In other words, Smith's concept of market order was considered applicable for all private goods and Hobbes's conception of the single centre of power and decision for all collective goods (Ostrom 1998b).

But what if the domains of modern political–economic life could not be understood or organised by relying only on the concepts of markets or states? Answering that challenge is probably one of the best ways to see Ostrom's work: a theoretically informed, empirically based contribution to a larger and bolder attempt to build an alternative to the basic dichotomy of modern

political economy. 'The presence of order in the world', Elinor Ostrom (1998a) writes, 'is largely dependent upon the theories used to understand the world. We should not be limited, however, to only the conceptions of order derived from the work of Smith and Hobbes'. We need a theory that 'offers an alternative that can be used to analyze and prescribe a variety of institutional arrangements to match the extensive variety of collective goods in the world'.

In response to that need, Ostrom has explored a new domain of the complex institutional reality of social life – the rich institutional arrangements that are 'neither states nor markets'. They are small and large, multi-purpose or just focused on one good or service: suburban municipalities, neighbourhood organisations, condominiums, churches, voluntary associations, or informal entities like those solving the common-pool resources dilemmas. As such they could be seen as a 'third sector' ('public economy' was one of the suggested names for it), related to, but different from, both 'the state' and 'the market'. Irrespective of how this domain is named, the fact is that a theoretical perspective that takes it into account is substantially different from one based on the classical dichotomy.

If that important aspect is considered, one could get a more nuanced view of Ostrom's place in Public Choice economics – an intellectual movement with which her work was associated from the very beginning Ostrom 1968, 1986; Ostrom and Ostrom 1971. Buchanan (1977) and Tullock (1970), argue convincingly that state failure is even more systematic and perverse than market failure. The Public Choice theory of Buchanan and Tullock is a theory of state failure. The state's efficiency must be proved, not postulated. Ostrom – while initially also contributing to the typical arguments regarding 'state failure' – went beyond the Buchanan and Tullock demonstration of the fact that in numerous cases the state is far from being 'the solution', for her emphasis was not on the 'bad news' but on the 'good news': a demonstration that, even in the case of public goods and services that the market and the state cannot supply efficiently, people can solve complex cooperation and coordination problems of governance and can develop complex institutional arrangements in order to produce and distribute precisely those goods and services. Self-governance is possible.

Governance and Public Choice

From the very beginning, Ostrom's work was grounded in the incipient Public Choice revolution. Her doctoral dissertation was an empirical extension of the pathbreaking article by Ostrom et al. (1961) 'The organization of government in metropolitan areas: a theoretical inquiry', an article that used modern economic theory to challenge the mainstream views regarding centralised administration and governance. She also drew on The Calculus of Consent by Buchanan and Tullock (1962) and Stigler's (1962) work on the functions of local government, as well as on Schumpeter's (1942) discussion of entrepreneurship. Ostrom's dissertation focused on the collective management of groundwater basins in Southern California and examined the processes used by those individuals that formed water associations to cope with the problem of water availability and quality when no political jurisdiction had the same boundaries as the groundwater basins.

This dissertation experience drew her attention to how disparate individuals could collectively band together to protect a common resource – what would become a defining theme for Ostrom's work. The experience was also instrumental in shaping her attitude towards 'heavy duty' empirical research through case studies and field work, another defining feature of her approach for the rest of her career: 'Undertaking this study, she wrote, gave me a deep respect for individual case studies based on intensive fieldwork. (. . .) Individual case studies are a very important method to include along with larger- n field studies, meta-analysis, formal models, and experimental research. None of these should be viewed as the only way or best way to do research' (Ostrom 2010).

One of the best illustrations of Ostrom's work is the metropolitan governance reform debate, a

debate that she engaged in enthusiastically soon after being offered an Assistant Professor position at Indiana University Bloomington in 1965. Conventional wisdom had been that a metropolitan region should be organised as one large administrative unit functionally integrated by bureaucratic hierarchies. Advocates of metropolitan reform argued in favour of centralisation and against what they called 'fragmentation' of urban services (Zimmerman 1970; McGinnis 1999).

Ostrom's early work (developed together with Vincent Ostrom) challenged the basic tenets of the 'reformers'. She argued that the optimum scale of production is not the same for all urban public goods and services. Some services may be produced 'more efficiently on a large scale while other services may be produced more efficiently on a small scale'. Therefore, the existence of multiple agencies interacting and overlapping, far from being a pathological situation, 'may be in fact a natural and healthy one', the result of the fact that scale efficiencies and the principles of division of labour, cooperation and exchange function in the public sector too. 'One need not assume a priori that competition among public agencies is necessarily inefficient', she wrote. 'Duplication of functions is assumed to be wasteful and inefficient. Yet we know that efficiency can be realized in a market economy only if multiple firms serve the same market. Overlapping service areas and duplicate facilities are necessary conditions for the maintenance of competition in a market economy. Can we expect similar forces to operate in a public economy?' (Ostrom and Ostrom 1965).

In addition, Ostrom demonstrated that the variety of relationships between governmental units, public agencies and private businesses coexisting and functioning in metropolitan areas can be coordinated through patterns of inter-organisational arrangements that 'would manifest market-like characteristics and display both efficiency-inducing and error-correcting behavior'. Coordination in the public sector, she argued, 'need not rely exclusively upon bureaucratic command structures controlled by chief executives. Instead, the structure of inter-organizational arrangements may create important economic opportunities and evoke self-regulating tendencies' (Ostrom 1983; Ostrom and Ostrom 1965).

Empirical Research on 'Public Economies'

Among the most distinctive aspects of Ostrom's work has been her approach to empirical research. The series of studies produced as part of the metropolitan governance debate are exemplary in this respect. Among the key issues in the metropolitan debate was the impact of the size of a government unit producing a service: was the size affecting positively or negatively the output and efficiency of service provision? To test the competing hypotheses, Ostrom (1972) and her team built an entire research programme. She selected one governmental function (the police) and started to gather the data needed to measure the relationship between department size and the efficiency of policing. At the most basic level, the complex research design concentrated on large centralised police departments versus smaller departments serving similar neighbourhoods observed across multiple indicators. The investigation (based on field teams and participant observation) started in Indianapolis, continued with the Chicago Police Department, and was followed by a massive survey and field study in the St Louis metropolitan area, while replications were undertaken in Grand Rapids, Michigan and in the Nashville–Davidson County area of Tennessee. To test for external validity, the team drew on a large survey of citizens living in 109 cities with populations of more than 10,000, conducted by the National Opinion Research Center and on data from the Municipal Year Book.

In this full set of investigations, writes Ostrom, 'no one found a single case where a large centralized police department was consistently able to outperform smaller departments serving similar neighborhoods across multiple indicators'. The study challenged on empirical grounds the notion that larger urban governments would always produce superior public services. The presumption that economies of scale were prevalent was wrong; the presumption that you needed a single

police department was wrong; and the presumption that individual departments wouldn't be smart enough to work out ways of coordinating was wrong (Ostrom 2010). Most aspects of police work in fact carried diseconomies of scale.

Ostrom (1976a, b) approached in a similar way the related question of the impact of the number of governments providing a service in a metropolitan area. With the support of the National Science Foundation, she developed a large study of the organisation of service delivery in metropolitan areas. Conventional wisdom and most prior studies had stressed the 'chaos' resulting from multiple units of government producing urban services in the same region. In what was to become one of her trademarks, Ostrom's empirical results were again challenging the 'self-evident truths' (Ostrom 2000; Ostrom et al. 2007). But, even more important, they led to a series of empirical analysis-based insights regarding the nature of institutional arrangements that individuals and communities use in order to produce, deliver and consume goods and services. Ostrom et al. 1973, 1978a, b; Ostrom and Parks 1973, 1999.

Her investigations demonstrated that, even in the case of public goods and services that the market and the state cannot supply efficiently, people can develop institutional arrangements in order to produce and distribute. That was not a theoretical possibility but an empirical reality that challenged the theory-based conventional wisdom. For instance, using a conceptual framework based on two variables (the feasibility of exclusion and jointness of use) as well as the ensuing typology of goods (public, private, toll and common pool) she discovered situations when the units of government were 'collective consumption units' whose first order of business was to articulate and aggregate demands for those goods that are subject to joint consumption where exclusion is difficult to attain. In such situations, relationships are coordinated among collective consumption and production units by contractual agreements, cooperative arrangements, competitive rivalry and mechanisms of conflict resolution. In a similar way, there are situations in which larger governmental entities are optimal. No single centre of authority is responsible for coordinating all relationships in such a 'public economy'. Market-like mechanisms can develop competitive pressures that tend to generate higher efficiency than can be gained by enterprises organised as exclusive monopolies and managed by elaborate hierarchies of officials.

To sum up, through an intensive empirical investigation combining multiple methods and teamwork, a complex system was revealed in which not only markets and hierarchies but also more hybrid and peculiar arrangements, including social networks and informal relations, were combined to generate a special institutional architecture. To name it, the notion of 'public economy' was introduced. The notion was purposefully chosen 'to save the concept of "public" from the false notion that "public" meant "the State" (or "centralized systems of governance") and to make clear the difference from the market economy'. In other words, 'to show that it is possible to have systems that are neither markets nor states, and which preserve the autonomy and the freedom of choice of the individual' (Ostrom and Ostrom 1977; Ostrom et al. 1992).

Social Dilemmas and the Commons

This is the context in which one could also read Ostrom's celebrated studies of the 'commons'. As the Nobel Prize 2009 press release put it, 'Ostrom has challenged the conventional wisdom that common property is poorly managed' by showing that 'resource users frequently develop sophisticated mechanisms for decision-making and rule enforcement to handle conflicts of interest, and she characterizes the rules that promote successful outcomes'.

Her work in this respect is better understood if we note that one of the most interesting and enduring features of her scholarship was a fascination with the dilemmas and paradoxes of social cooperation – 'action situations' that imply theoretical and empirical puzzles. A good introduction to this theme is the so-called 'service paradox': the conjecture that the increasing professionalisation of public services is accompanied by serious erosion in the quality of those services. To deal with

this puzzle, Ostrom mobilised the usual combination of Public Choice conceptual instruments (in this case, the theory of goods) and hard-nosed empiricism. Her investigations revealed a whole series of cases wherein the collaboration between those who supplied a service and those who used it was the factor determining the effective delivery of the service. In other words, in many instances the users of services also function as co-producers, for production was not separated from consumption. And thus one gets to the solution of the 'service paradox': The standard assumption of the separation of production from consumption blinded everybody from identifying its source. When professional personnel, writes Ostrom, 'presume to know what is good for people rather than providing people with opportunities to express their own preferences, we should not be surprised to find that increasing professionalization of public services is accompanied by a serious erosion in the quality of those services'. Hence a policy implication: the organisation of a public economy that 'gives consideration to economies of consumption as well as of production and provides for the co-ordination of the two is most likely to attain the best results' (Ostrom and Ostrom 1977).

In a similar way, Ostrom was fascinated by the collective action dilemmas identified by Mancur Olson (1965) and Garrett Hardin (1968). These dilemmas recognised that those harvesting from a common-pool resource (pasture, fisheries or groundwater basins) have incentives to harvest for individual gain as much as (and as fast as) they can, leading to depletion and long-term losses for all. Various empirical studies were indicating the capacity of local users to solve problems of the commons, contrary to the standard rational choice theory predictions. However, it was generally believed to be impossible for individuals involved in such situations to overcome the problems.

Challenged by Reinhard Selten and D. C. North, Ostrom decided to try to understand why some users overcame the tragedy of the commons, while others were unable to do it. In the typical manner, she started with large-scale empirical research. Together with her team from the Workshop in Political Theory and Policy Analysis, she identified over 1000 documented cases related to diverse resources (fisheries, forests, irrigation systems etc.) in many regions of the world. She developed with them the Common-Pool Resource (CPR) database, at that point the largest in the world. That in itself was a remarkable task: 'Several years were devoted to screening cases to assess the quality and extensiveness of data collected, to record those cases with substantial information, to check with case authors when feasible to improve data quality, and to undertake careful analysis' (Ostrom 1990, 1999, 2010; Ostrom et al. 1994). Field work in, among other countries, Nepal, Nigeria and Kenya, was combined with in-depth comparative case studies, formal modelling, statistical analysis and experimental studies to test specific hypotheses. In the end, this line of research not only generated a robust body of knowledge regarding CPR governance but also a series of insights about the potential and limits of institutional design principles.

In addition, the CPR research also led to contributions to the theory of property rights, broadly defined. For instance, contrary to the theory stressing the centrality of alienation rights (Demsetz 1967), Schlager and Ostrom (1992) found that user rights of access, withdrawal, management, exclusion and alienation were all important rights and were cumulative. 'This led to a new conceptual terminology for analyzing bundles of rights within a hierarchy of possible rights (. . .) and demonstrated, among others, that users did not need alienation rights in order to manage a resource sustainably' (Ostrom 2010). This conception of property rights is now generally accepted as the main framework for the analysis of property rights systems around the world.

Institutional Analysis: Frameworks and Methods

Throughout her career, an important part of Ostrom's theoretical effort was dedicated to the development of operational analytical frameworks for the study of institutions (Aligica and Boettke 2009; Ostrom and Walker 2003). The

most important in this respect is the Institutional Analysis and Development (IAD) framework (Ostrom 2005), a multi-tier conceptual map based on two elements: first, the distinction between three tiers of decision making (constitutional, collective choice, and operational) and the relations among them; and second, the operationalisation of a conceptual unit – called an action situation – meant to analyse behaviour within institutional arrangements at any of the three tiers of decision making. Ostrom's endeavour to develop 'a cumulative syntax that would enable future work on institutional analysis to have a common foundation' has also lead to a typology of rules that potentially affect action situations. To illustrate the usefulness of analysing the rules that underlie action situations, she examined, among others, familiar models of the bargaining between elected and public bureaucratic officials over the output-cost combination to serve their citizens (Downs 1957; Niskanen 1971; Romer and Rosenthal 1978; McGuire et al. 1979). All these authors derived different predictions about the equilibrium outcome in a bargaining game. 'Controversy existed', writes Ostrom (2010), 'as to whose model was correct. My analysis showed that they were all correct, given the differences in the underlying rules of each model. I was thus able to demonstrate that digging under competing models to examine the specific rules assumed by scholars, explained why the predictions made for the "same" bargaining situation differed so widely'.

Ostrom's recent efforts have been dedicated to how individual case studies, meta-analyses of multiple cases, large-scale comparative field-based analysis, formal theory, experimental research and agent-based models could be combined in collaborative research practice (Poteete et al. 2010). By highlighting the multiple methods approach, she restates a basic but subtle tenet underlying her philosophy of social research as well as the success of her career: a deep conviction that it is both desirable and possible to build an alternative that goes beyond mechanical applications and formal interpretation and thus to reclaim the spirit of genuine empirical research based on data collection, observation and in-depth analysis.

See Also

- ▶ Buchanan, James M. (Born 1919)
- ▶ Common Property Resources
- ▶ Economic Governance
- ▶ Market Institutions
- ▶ New Institutional Economics
- ▶ 'Political Economy'
- ▶ Public Choice
- ▶ Public Goods
- ▶ Property Rights
- ▶ Rational Choice and Political Science
- ▶ Tragedy of the Commons

Bibliography

Aligica, P.D., and P. Boettke. 2009. *Challenging institutional analysis and development: The Bloomington school*. New York: Routledge.

Buchanan, J. 1977. *Democracy in deficit (with R. Wagner)*. New York: Academic.

Buchanan, J.M., and G. Tullock. 1962. *The calculus of consent: Logical foundations of constitutional democracy*. Ann Arbor: University of Michigan Press.

Demsetz, H. 1967. Toward a theory of property rights. *American Economic Review* 57 (2): 347–359.

Downs, A. 1957. *An economic theory of democracy*. New York: Harper & Row.

Elinor Ostrom Nobel Prize web page. 2009. http://nobelprize.org/nobel_prizes/economics/laureates/2009/index.html

Hardin, G. 1968. The tragedy of the commons. *Science* 162: 1243–1248.

McGinnis, M.D., ed. 1999. *Polycentricity and local public economies: Readings from the workshop in political theory and policy analysis*. Ann Arbor: University of Michigan Press.

McGuire, T., M. Coiner, and L. Spancake. 1979. Budget maximizing agencies and efficiency in government. *Public Choice* 34 (3/4): 333–359.

Niskanen, W.A. 1971. *Bureaucracy and representative government*. Chicago: Aldine-Atherton.

Olson, M. 1965. *The logic of collective action*. Cambridge, MA: Harvard University Press.

Ostrom, E. 1968. Some postulated effects of learning on constitutional behavior. *Public Choice* 5: 87–104.

Ostrom, E. 1972. Metropolitan reform: Propositions derived from two traditions. *Social Science Quarterly* 53: 474–493.

Ostrom, E., ed. 1976a. *The delivery of urban services: Outcomes of change*. Beverly Hills: Sage.

Ostrom, E. 1976b. Size and performance in a federal system. *Publius: The Journal of Federalism* 6 (2): 33–73.

Ostrom, E. 1983. A public choice approach to metropolitan institutions: Structure, incentives, and performance. *The Social Science Journal* 20 (3): 79–96.

Ostrom, E. 1986. An agenda for the study of institutions. *Public Choice* 48 (1): 3–25.
Ostrom, E. 1990. *Governing the commons: The evolution of institutions for collective action*. New York: Cambridge University Press.
Ostrom, E. 1998a. *The comparative study of public economies. Presented upon acceptance of the Frank E. Seidman distinguished award in political economy.* Memphis: P.K. Seidman Foundation.
Ostrom, E. 1998b. A behavioral approach to the rational choice theory of collective action. *American Political Science Review* 92 (1): 1–22.
Ostrom, E. 1999. Coping with tragedies of the commons. *Annual Review of Political Science* 2: 493–535.
Ostrom, E. 2000. The danger of self-evident truth. *PS: Political Science & Politics* 31 (1): 33–44.
Ostrom, E. 2005. *Understanding institutional diversity.* Princeton: Princeton University Press.
Ostrom, E. 2010. A long polycentric journey. *Annual Review of Political Science* 13: 1–23.
Ostrom, V., and E. Ostrom. 1965. A behavioral approach to the study of intergovernmental relations. *The Annals of the American Academy of Political and Social Science* 359: 137–146.
Ostrom, V., and E. Ostrom. 1971. Public choice: A different approach to the study of public administration. *Public Administration Review* 31 (2): 203–216.
Ostrom, V., and E. Ostrom. 1977. Public goods and public choices. In *Alternatives for delivering public services. Toward improved performance*, ed. E.S. Savas, 7–49. Boulder: Westview Press.
Ostrom, E., and R.B. Parks. 1973. Suburban police departments: Too many and too small? In *The urbanization of the suburbs*, ed. L.H. Masotti and J.K. Hadden, 367–402. Beverly Hills: Sage.
Ostrom, E., and R.B. Parks. 1999. Neither gargantua nor the land of lilliputs: Conjectures on mixed systems of metropolitan organization. In *Polycentricity and local public economies: Readings from the workshop in political theory and policy analysis*, ed. M.D. McGinnis, 284–305. Ann Arbor: University of Michigan Press.
Ostrom, E., and J. Walker, eds. 2003. *Trust and reciprocity: Interdisciplinary lessons from experimental research*. New York: Russell Sage Found.
Ostrom, V., C.M. Tiebout, and R.L. Warren. 1961. The organization of government in metropolitan areas: A theoretical inquiry. *American Political Science Review* 55: 831–842.
Ostrom, E., W. Baugh, R. Guarasci, R.B. Parks, and G.P. Whitaker. 1973. *Community organization and the provision of police services*. Beverly Hills: Sage.
Ostrom, E., R.B. Parks, and G.P. Whitaker. 1978a. *Patterns of metropolitan policing*. Cambridge, MA: Ballinger.
Ostrom, E., R.B. Parks, G.P. Whitaker, and S.L. Percy. 1978b. The public service production process: A framework for analyzing police services. *Policy Studies Journal* 7: 381–389.
Ostrom, E., J. Walker, and R. Gardner. 1992. Covenants with and without a sword: Self-governance is possible. *American Political Science Review* 86 (2): 404–417.
Ostrom, E., R. Gardner, and J. Walker. 1994. *Rules, games, and common-pool resources*. Ann Arbor: University of Michigan Press.
Ostrom, E., M. Janssen, and J. Anderies. 2007. Going beyond panaceas. *PNAS* 104 (39): 15176–15178.
Poteete, A., M. Janssen, and E. Ostrom. 2010. *Working together: Collective action, the commons, and multiple methods in practice*. Princeton: Princeton University Press.
Romer, T., and H. Rosenthal. 1978. Political resource allocation, controlled agendas, and the status quo. *Public Choice* 33 (4): 27–43.
Schlager, E., and E. Ostrom. 1992. Property-rights regimes and natural resources: A conceptual analysis. *Land Economics* 68 (3): 249–262.
Schumpeter, J.A. 1942. *Capitalism, socialism and democracy.* New York: Harper Torchbooks.
Stigler, G.J. 1962. The tenable range of functions of local government. In *Private wants and public needs*, ed. E.S. Phelps, 167–176. New York: Norton.
Tullock, G. 1970. *Private wants, public means: An economic analysis of the desirable scope of government*. New York: Basic Books.
Zimmerman, J.F. 1970. Metropolitan reform in the U.S.: An overview. *Public Administration Review* 30: 531–543.

Outliers

William S. Krasker

Nearly all empirical investigations in economics, particularly those involving linear structural models or regressions, are subject to the problem of anomalous data, commonly called outliers. Roughly speaking, there are three sources of outliers. First, the distribution of the model's random disturbances often has longer tails than the normal distribution, resulting in a greatly increased chance of larger disturbances. Second, the data set may contain erroneous numbers, or 'gross errors'. The data bases most prone to gross errors are large cross sections, particularly those compiled from surveys; gross errors can result from misinterpreted questions, incorrectly recorded answers, keypunch errors, etc. Third, the model itself, typically linear in (transformations of) the variables, is only an approximation to reality. It is apt to be a poor representation of the process

generating the data for extreme values of the explanatory variables. This source of outliers applies even to, say, macroeconomic time series, where the likelihood of gross errors is minimal.

Outliers resulting from heavy-tailed but still symmetric disturbance distributions can greatly decrease the efficiency of least squares, while gross errors can in addition cause substantial biases. These potentially damaging effects of anomalous data have been recognized for many years; indeed, the first published work on least squares (Legendre 1805) recommended that outliers be removed from the sample before estimation. The wisdom of this and other approaches that give the observations unequal weights was debated throughout the 19th century.

Despite considerable evidence that error distributions tend to be heavy tailed, many statisticians were reluctant to modify least squares, which was known to be optimal when the disturbances are normally distributed. There were notable exceptions, however, such as Simon Newcomb, an astronomer and mathematician as well as an economist. Newcomb (1886) introduced the idea of modelling the disturbance distribution as a mixture of normal distributions with differing variances; the implied marginal distribution then has heavier tails than the normal. Newcomb also proposed a 'weighted least squares' alternative that, it turns out, is similar to a 1964 proposal of Peter Huber, discussed below, which has numerous desirable robustness properties. (The contributions of Newcomb and other late-19th and early 20th-century statisticians are discussed in more detail by Stigler 1973.)

There was a rapid increase in interest in robustness in the mid 1900s, in part due to the work of John Tukey (see, e.g., Tukey 1960). Robustness research benefited greatly in the 1960s from the formalization of certain desirable robustness properties of estimators. The first is 'efficiency robustness': one would like an estimator to maintain a high efficiency for all symmetric disturbance distributions that are 'close to' the normal distribution. Peter Huber (1964) found a one-parameter family of estimators, indexed by $c > 0$, that have a certain optimal minimax efficiency-robustness property. Suppose the regression model is $y_i = x_i\beta + u_i$ $(i = 1,\ldots, n)$, where y_i is the ith observation on the dependent variable, x_i is the k-dimensional row vector containing the ith observation on the explanatory variables, u_i is the ith disturbance, and β is the k-vector parameters to be estimated. Then the Huber estimate b is the vector that solves the equations

$$0 = \psi_c(y_i - x_i b)\, x_{ij} \ (j = 1, \ldots, k),$$

where $\psi_c(t) \equiv \max[-c, \min(t, c)]$ and where the choice of the parameter c depends on the scale of the disturbance distribution and the desired tradeoff between robustness and efficiency. As $c \to \infty$, the Huber estimator reduces to ordinary least squares, whereas if c is never zero, the estimator is similar to the method of least absolute residuals, which had been studied as early as Laplace (1818) and which gained some popularity in the 1950s (see Taylor 1974). The Huber estimator and over sixty others were compared for small samples in the 'location' problem (regression on just a constant term) in an extensive 1970–71 Monte Carlo study (Andrews et al. 1972). The results suggested that the asymptotic properties hold quite well in samples as small as twenty.

Though the Huber estimators, and others designed for efficiency robustness, maintain a high efficiency even for heavy-tailed disturbance distributions, they are not resistant to other sources of outliers, such as low-probability gross errors. A second desirable robustness property, introduced by Hampel (1968, 1971) and corresponding to the mathematical concept of uniform continuity, is that if gross errors are generated with small probability, then, irrespective of the distribution of those gross errors, the estimator's bias should be small. Estimators having this property are called 'qualitatively robust'. Hampel quantified this relationship by means of an estimator's 'sensitivity', which he defined as the right-hand derivative of the maximum possible bias, with respect to the probability of gross errors, evaluated at probability zero.

Modifications of the Huber estimator designed to make it qualitatively robust were proposed by

several researchers in the 1970s (see Krasker and Welsch (1982) for further discussion). They have the general form

$$0 = v(x_i)\psi_c((y_i - x_i b)/(w(x_i)) \times x_{ij} \quad (j = 1, \ldots, k)$$

where w and v are non-negative weight functions that allows for the downweighting of observations with outlying values for the explanatory variables, called 'leverage points'. The proposals of Krasker (1980) and Krasker and Welsch (1982) also have a certain efficiency property among estimators with the same sensitivity to gross errors. The idea of finding an estimator that has maximum efficiency subject to a bound on the sensitivity was developed by Hampel (1968).

If an estimator is qualitatively robust, its asymptotic bias will be small provided the probability of gross errors is sufficiently small. However, this property does not tell us how the estimator will behave if the gross errors are, say, ten per cent of the data. One crude measure of this behaviour, introduced by Hampel (1968, 1971) and called the 'breakdown point', is the smallest probability of gross errors that can cause the asymptotic bias to be arbitrarily large. Equivalently, it is the largest fraction of gross errors in the data that the estimator can handle before it becomes totally unreliable. By the 1980s it was clear that the most common qualitatively robust regression estimators, such as those listed earlier, have low breakdown points when k, the number of parameters, is large. Several alternative estimators $\frac{1}{2}$, the largest possible value. Examples are the 'repeated medians' estimator of Siegel (1982), the projection-pursuit approach of Donoho and Huber (1983), and the estimator proposed by Rousseeuw (1984), which minimizes the median of the squared residuals (rather than their sum). However, all of these estimators are computationally burdensome unless k is small, and in fact, it appears that the computational difficulties are an inherent feature of high-breakdown multivariate procedures that transform naturally under linear changes in the coordinate system.

One of the most important uses for high-breakdown procedures is simply to facilitate the identification of outliers, which are often masked by non-robust estimators. For example, in a simple regression, a single outlier associated with an extreme value of the explanatory variable can have so much influence on the least-squares estimate that its own residual is very small. Thus, mere examination of the residuals from a non-robust fit can fail to reveal the anomalous observations. This problem becomes much more severe in higher dimensions, where even many qualitatively robust estimators can break down due to a small cluster of outlying observations. Belsley et al. (1980) have proposed a variety of methods for identifying outliers in regression.

For statistical inference, as opposed to data analysis, identification of the outliers is only a small part of the problem. An important difficulty is that it is often impossible to determine solely from the data whether an outlying observation results from aberrant data, or whether the true regression function is slightly non-linear. Typically either of these possibilities will 'explain' the outlier, but for inference their implications may be very different. In these circumstances it seems essential to place a prior on the amount of curvature in the regression function, but this is difficult to do, particularly when there are several explanatory variables. One approach is outlined in Krasker et al. (1983, section 5).

Finally, although the preceding remarks have dealt with regression, outliers occur and have similar consequences in many other statistical contexts, such as discrete or censored dependent variable models, stochastic parameter models, or linear structural models. The most reliable way to identify outliers in these contexts is to estimate robustly the model's underlying parameters, and check for observations that deviate greatly in an appropriate sense from the model's prediction. For example, Krasker and Welsch (1985b) have presented a qualitatively robust weighted-instrumental-variables estimator for simultaneous-equations models, analogous to their proposal for regression. In general, however, methods for dealing with outliers in models of the kind just mentioned are far less developed than those for regression.

See Also

- ▶ Estimation
- ▶ Least Squares
- ▶ Residuals

Bibliography

Andrews, D.F., P.J. Bickel, F.R. Hampel, P.J. Huber, W.H. Rogers, and J.W. Tukey. 1972. *Robust estimates of location: Survey and advances*. Princeton: Princeton University Press.

Belsley, D.A., E. Kuh, and R.E. Welsch. 1980. *Regression diagnostics*. New York: Wiley.

Donoho, D.L., and P.J. Huber. 1983. The notion of breakdown point. In *A Festschrift for Erich L. Lehmann*, ed. P. Bickel, K. Doksum, and J.L. Hodges Jr.. Belmont: Wadsworth International Group.

Hampel, F.R. 1968. Contributions to the theory of robust estimation. PhD thesis, University of California, Berkeley.

Hampel, F.R. 1971. A general qualitative definition of robustness. *Annals of Mathematical Statistics* 42: 1887–1896.

Huber, P.J. 1964. Robust estimation of a location parameter. *Annals of Mathematical Statistics* 35(1): 73–101.

Krasker, W.S. 1980. Estimation in linear regression models with disparate data points. *Econometrica* 48: 1333–1346.

Krasker, W.S., and R.E. Welsch. 1985a. Efficient bounded-influence regression estimation. *Journal of the American Statistical Association* 77(379): 595–604.

Krasker, W.S., and R.E. Welsch. 1985b. Resistant estimation for simultaneous-equations models using weighted instrumental variables. *Econometrica* 53(6): 1475–1488.

Krasker, W.S., E. Kuh, and R.E. Welsch. 1983. Estimation for dirty data and flawed models. In *Handbook of econometrics*, vol. 1, ed. Z. Griliches and M.D. Intriligator. Amsterdam: North-Holland.

de Laplace, P.S. 1818. *Deuxième supplèment à la théorie analytique des probabilités*. Paris: Courcier. Reprinted in *Oeuvres de Laplace*, vol. 7, 569–623. Paris: Imprimerie Royale, 1847. Repinted in *Oeuvres complètes de Laplace*, vol. 7, 531–580. Paris: Gauthier-Villars, 1886.

Legendre, A.M. 1805. On the method of least squares. Trans. in *A source book in mathematics*, ed. D.E. Smith. New York: Dover Publications, 1959.

Newcomb, S. 1886. A generalized theory of the combination of observations so as to obtain the best result. *American Journal of Mathematics* 8: 343–366.

Rousseeuw, P.J. 1984. Least median of squares regression. *Journal of the American Statistical Association* 79(388): 871–880.

Siegel, A.F. 1982. Robust regression using repeated medians. *Biometrika* 69: 242–244.

Stigler, S.M. 1973. Simon Newcomb. Percy Daniell, and the history of robust estimation, 1885–1920. *Journal of the American Statistical Association* 68(344): 872–879.

Taylor, L.D. 1974. Estimation by minimizing the sum of absolute errors. In *Frontiers of econometrics*, ed. P. Zarembka. New York: Academic Press.

Tukey, J.W. 1960. A survey of sampling from contaminated distributions. In *Contributions to probability and statistics*, ed. I. Olkin. Stanford: Stanford University Press.

Output and Employment

J. A. Kregel

Within his broader analysis of the 'nature and causes of the wealth of nations', Adam Smith (1776) identified the primary determinants of the growth of national output as labour productivity (given by the state of technology as determined by the division of labour), and the proportion of the total working population 'productively' employed (in modern language, producing outputs directed to the support of capital accumulation). For Smith, and other classical economists, the problem was not that commodities might remain unsold or labour unemployed, but the composition of output and employment required for capital accumulation: a high proportion of 'unproductive' labour would slow the pace of technological change by reducing the expansion of the market and thus the division of labour. When capital accumulation fell below the growth of population unemployment increased and wages would fall below subsistence, reducing population growth. The distribution of income between rent and profits was a key determinant of the composition of output: landowners expenditure on services or luxury goods being unproductive.

But whatever the rate of capital accumulation it was argued that a 'glut of commodities in the aggregate' was impossible, since 'there cannot be an aggregate supply without an equal aggregate demand' (James Mill, Mill 1844, p. 238). If individuals must produce in order to purchase

commodities, and do not want 'money but in order to lay it out, either in articles of productive, or articles of unproductive consumption' (p. 233–4) then since 'the demand and supply of every individual are always equal to one another, the demand and supply of all the individuals in the nation, taken aggregately, must be equal' (p. 232). Although it was always possible for 'miscalculation' to produce 'superabundance or defect' (p. 241) of commodities in particular markets, they would cancel in the aggregate. Mill's argument that what was true of an individual's output and employment was *a fortiori* true of the aggregate of all actions represented in the economy as a whole also formed the basis of Say's Law, and dominates modern discussion of the 'microfoundations of macroeconomics'.

Neoclassical theory did not challenge this method of approach, but shifted emphasis from the growth-maximizing composition of output and employment to analysis of individual utility and profit maximizing allocation of given resources to alternative uses. The existence of excess supply or demand in any market represented misallocation of resources and an unexploited possibility to increase total profit or utility by substitution in production or consumption until marginal utility or profit generated by the marginal purchase was equal for each commodity produced or purchased. Thus Lionel Robbins's famous definition of economics as 'the science which studies human behaviour as a relationship between ends and scarce means which have alternative uses' (1935, p. 16). The theory thus took as given the size of available output that Smith and Ricardo had tried to explain, and analysed what they had taken for granted, the decisions determining the allocation of expenditure across various commodities.

Since labour was also a scarce 'means' owned by the individual its allocation could be analysed relative to a market equilibrium wage rate which assured employment for all those willing and able to work at that wage. The profit-seeking individual acting in the market would thus insure movements in relative prices guaranteeing that excess supplies of commodities or labour were only temporary, arising from imperfections or frictions in the adjustment of competitive price.

In his *Theory of Unemployment* (1933) Pigou used the presumption that what was true of the individual market was true of the economy as a whole, to extend the analysis to the aggregate level (cf. Roncaglia and Tonveronachi 1985): if Q is aggregate real output and N the level of employment, then (1) $Q = f(N)$ $[f' > 0, f' > 0]$ given the technology embodied in existing equipment. Given the money stock M, and income velocity of circulation $(1/k)$, the Cambridge version of the quantity theory equation of exchange determines nominal income, Y: (2) $M = k(Y) = k(pQ)$. The division of Y into real output and the general price level, p, as well as the level of employment, is then given by the money wage w, and competitive profit maximization: (3) $w/p = f(Q)$, The level of aggregate real output and employment are inversely related to the real wage: employment could be increased either by reducing money wages given the money supply or increasing the money supply, given money wages. Frictions in the market adjustment process might temporarily keep money wages too high or the money stock and prices too low to produce equilibrium in the aggregate output and labour market.

Keynes (1936) directly questioned this extension of the analysis of a single market to the aggregate economy because it failed to capture the interdependence of output and expenditures, or of supply and demand, at the aggregate level. Keynes pointed out that the expansion of employment and output that Pigou's theory presumed to follow from a reduction in money wages rested on the implicit assumption 'that aggregate demand depends on the quantity of money multiplied by the income velocity of money' (1936, p. 258) so that assuming a given stock of money was equivalent to the assumption that the aggregate effective demand is fixed. . . . whilst no one would wish to deny the proposition that a reduction in money-wages *accompanied by the same aggregate effective demand as before* will be associated with an increase in employment, the precise question at issue is whether the reduction in money-wages will . . . be accompanied by the same aggregate effective demand as before . . . which is not

reduced in full proportion to the reduction in money wages (ibid., pp. 259–60).

Keynes noted that this crucial assumption had probably been overlooked because of an unwarranted extension of the argument that the horizontal average revenue curve of the competitive firm is unaffected by changes in its level of output. But if demand is exogenously given for each firm, it is exogenous for the sum of all firms. Keynes argued that if the 'theory is not allowed to extend by analogy its conclusion in respect of a particular industry to industry as a whole, it is wholly unable to answer the question what effect on employment a reduction in money-wages will have' (ibid., p. 260). The answer to this question involved the precise relationship between changes in money wages and prices, wage and non-wage real incomes and consumption and investment expenditures which the theory did not provide: what was needed was a theory of the determinants of aggregate demand.

Keynes's 'principle of effective demand' sought to provide an explicit explanation of aggregate demand in terms of the combination of the propensity to consume setting consumption expenditure, and liquidity preference and the money supply setting the rate of interest together with the efficiency of capital determining investment expenditure, the multiplier converting the two types of expenditure into aggregate income. In his theory equilibria might occur in which labour willing and able to work at going wages in competitive markets could not find employment, while lower wages would only reduce income and consumption expenditure in like or greater proportion. Instead of a unique stable equilibrium at full employment output, Keynes's analysis suggests the possibility of equilibria at any level of output and employment.

Keynes thus replaced both the quantity theory and what he called the 'second classical postulate' that labour could determine its real wage by altering its money supply price. Although profit maximization would assure the equality represented by equation (3) above, it would no longer be the relation which determined Q, nor would the quantity equation (2) determine Y.

The analysis also implicitly rejects the central proposition of Pigou's theory that flexible real wages determined in the labour market produce an automatic tendency to full employment output in all markets. This recognition that 'other things' will influence the real wage and thus the behaviour of all markets represents a criticism of Marshall's partial equilibrium analysis, but it also questions the automatic tendency to full employment output in a fully interdependent general equilibrium analysis, for in the absence of a Walrasian 'auctioneer' providing perfect, costless information, no single agent can predict the behaviour of the system as a whole without knowledge of the consequences of his actions on the behaviour of others. Without an explicit analysis of aggregate demand neither the partial equilibrium of a single market, nor a general equilibrium of all markets simultaneously, can provide the assurance that changes in wages and prices will not produce a more than offsetting change in aggregate demand. In opposition to Mill, individual or market equilibrium can only be understood relative to aggregate equilibrium; since aggregate equilibrium can occur at any level of output and employment there can be no automatic tendency to any unique level of production in individual markets.

See Also

► Effective demand

Bibliography

Keynes, J.M. 1936. *The general theory of employment, interest and money*. London: Macmillan for the Royal Economic Society. 1973.

Mill, J. 1844. *Elements of political economy*, 3rd ed. Revised and corrected. Reprinted, New York: Kelley, 1965.

Pigou, A.C. 1933. *The theory of unemployment*. London: Macmillan.

Robbins, L. 1935. *An essay on the nature and significance of economic science*, 2nd ed. London: Macmillan.

Roncaglia, A., and M. Tonveronachi. 1985. Pre-Keynesian roots of the neoclassical synthesis. *Cahiers d'Economie Politique* No.s 10–11.

Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*. Oxford: Oxford University Press. 1976.

Output Fall – Transformational Recession

Barry W. Ickes

Abstract

A significant decline in GDP has been a common feature in transition economies. This sharp drop in output has been seen as a surprise and puzzle to many observers. Understanding the nature of the output fall is crucial to understanding transition. Analysis is complicated by measurement issues associated with moving from plan to the market. Theoretical models of the output fall are examined, including those that see the output fall as a natural consequence of the legacies of the Soviet-type economic system.

Keywords

Arrears; Asset specificity; Command economy; Coordination problems; Double marginalization model; Hold-up problem; Incomplete contracts; Incomplete information; Monopoly; National income measurement; Output fall in transition economies; Over-Industrialization; Planning; Price controls; Price liberalization; Privatization; Search frictions; Second economy; Technical complementarities; Trade dependency; Uncertainty problems; Value destruction; Wage rigidity

JEL Classifications

P

A significant output decline has been a common feature in transition economies. To some extent this is a surprise: transition represents the removal of (highly significant) distortions. See, for example, Blanchard (1997, p. v): 'The fact that the transition came with an often large initial decrease in output should be seen as a puzzle. After all, the previous economic system was characterized by a myriad of distortions. One might have expected that removing most of them would lead to a large increase, not a decrease, in output.' Or as Svejnar (2000, p. 8) notes, 'The depth and length of the early transition depression was unexpected.' Similarly, Robert Mundell has written:

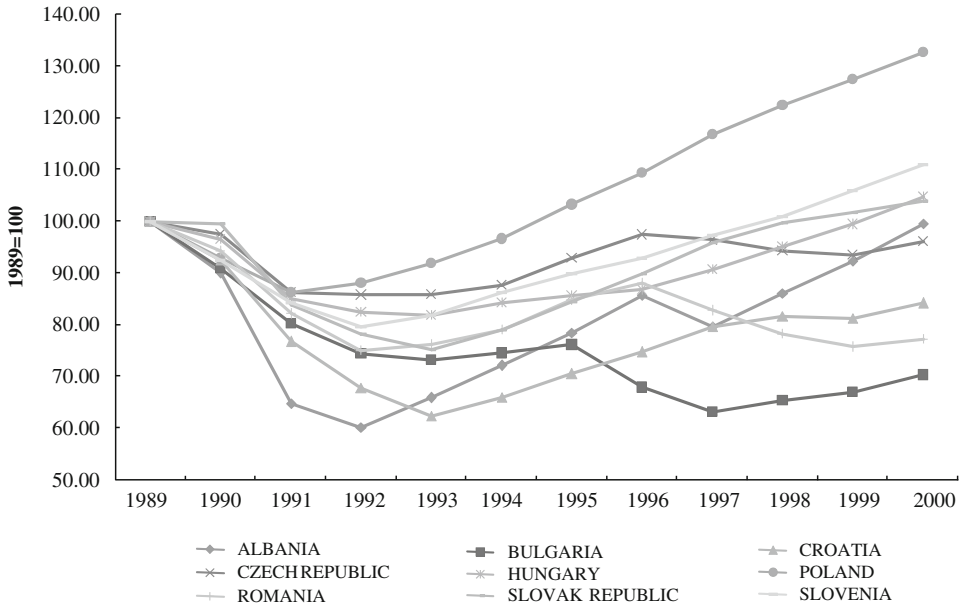
> The first and most obvious conclusion is that output contracted by a cumulative percentage never before experienced in the history of capitalist economies (at least in peacetime). Early denials that the contractions were occurring have proved to be incorrect. We observe that cumulative contractions over the 1990–4 period ranged widely, from a low of 18% to a high of more than 80%. (Mundell 1997, pp. 97–8)

Hence, a simple neoclassical argument would predict that output would rise rather than fall as the transition starts. Yet output fell in each transition economy, and quite significantly. The *officially reported* cumulative output decline for 26 transition economies from 1989 to 1995 was 41 per cent. Of this, the average decline in central Europe was 28 per cent and in the former Soviet Union it was 54 per cent (Fischer and Sahay 2000, Table 1). By comparison, output in the United States during the Great Depression declined by 34 per cent. The ubiquitous nature of the output fall thus represents an important puzzle for transition economics, and understanding the causes and nature of the output fall is crucial.

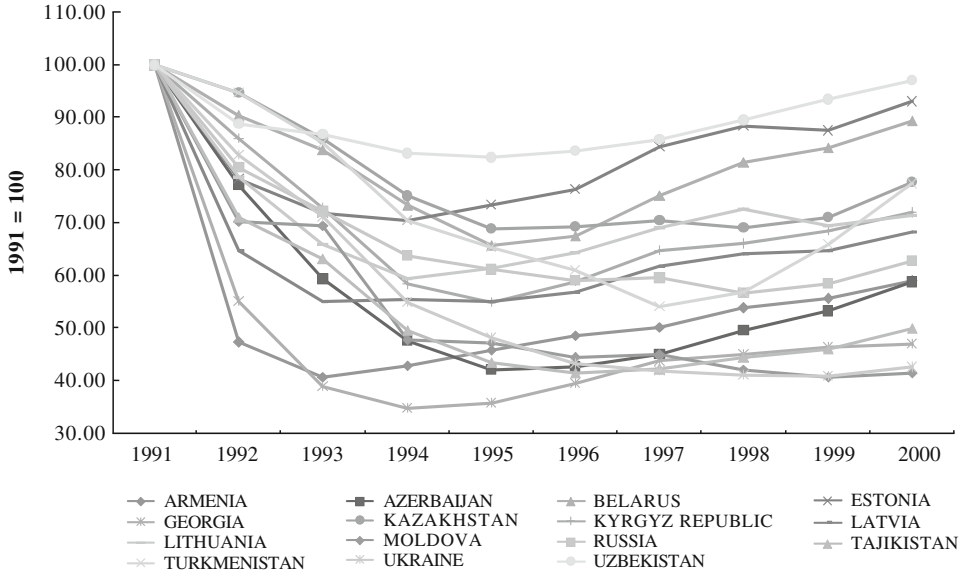
Analysis of the output fall is complicated by important measurement issues. In the change of economic systems from plan to market, the valuation of goods and services changes dramatically. This makes it important to distinguish official measures of the output fall from welfare-based measures.

Stylized Facts

It is useful to begin with some stylized facts about the output fall. Official GDP measures of output are given in Figs. 1 and 2. It is evident from these figures that in all transition economies output follows a U-shaped pattern. This represents another interesting puzzle. A theory based on the chaotic nature of the collapse of planning might predict that output would collapse at the start of



Output Fall – Transformational Recession, Fig. 1 Official GDP growth in central and eastern Europe (*Source*: International Monetary Fund Dataset)



Output Fall – Transformational Recession, Fig. 2 GDP in the former Soviet Union, 1989–2000 (*Source*: International Monetary Fund Dataset)

transition, but would rise from that point. The pattern displayed by the transition economies, on the other hand, suggests that the peak output fall occurs with a lag of several years. So an additional part of the puzzle is to explain why the output fall intensifies in the early transition.

Measured output fell in all transition economies. Generally, the declines are larger in the

former Soviet Union (FSU) than in central and eastern European economies (CEEs). For example, using 1989 as the starting point, the falls in Poland (15 per cent), Hungary (18 per cent), and the Czech Republic (21 per cent) were relatively moderate compared with Russia, where from 1991 GDP fell by 40 per cent. Later reformers appear to have larger falls: Ukraine has had a very significant fall in output. (There is, however, a puzzle concerning the output path of Uzbekistan. The output fall was smaller there than in any former Soviet republic, yet it reformed the least. For an analysis, see Zettelmeyer 1998.)

If we look at industrial output, rather than GDP, the observed declines would be even larger: about 40–50 per cent in central Europe and 50–60 per cent in the FSU. The reason, of course, is that most of the negative value added under planning was in industry, so we would expect a larger contraction there.

The decline in investment, especially in inventories and housing, was even greater than the decrease in GDP. This is especially true for defence. Hence, consumption has fallen less than GDP. In a sense, this is not a surprise as investment is more volatile than output in market economies. Yet transition as an economic process involves restructuring, and this does require investment. The fact that investment absorbed so much of the shock means that the resumption of growth was delayed even further. But it also means that living standards have not fallen as much as GDP. This is important for considering the welfare effects of the output decline.

Measurement Issues

Perhaps the output fall is overestimated. (Aslund 2002, p. 121, considers the output fall to be a myth.) There are many problems with interpreting official data in the context of transition, especially with respect to living standards: too many, indeed, to discuss here. Tracing output dynamics in the transition is complicated by the measurement issues that arise as the economic environment changes from central planning to market forces. Hence, an important issue in understanding the output fall is to gauge the extent to which it is a statistical rather than a real phenomenon.

Some observers (Aslund 2002; Campos and Coricelli 2002) argue that the size of the output fall is overstated because of the growth in the size of the shadow economy in early transition. It is argued that the hidden economy grew substantially during the transition period. Hence, actual production fell by less than measured output. The factual basis of this claim is controversial, however. It is also suspect theoretically. The biggest incentive to growth in the second economy is price controls. Hence, price liberalization should result in an immediate drop in the size of the shadow economy. The countervailing pressure could come from tax incentives, but it is hard to believe that this force is stronger than the impact of price controls.

The typical evidence cited in support of the proposition that the hidden economy grew in transition is that measured output fell by more than electricity production. Estimates based on comparing electricity consumption and GDP assume that the elasticity is close to unity. But this elasticity is well below unity in market economies *during recessions*, so employing the unit elasticity assumption amounts to assuming away the phenomenon to be measured. In Finland, for example, real GDP fell by about 11 per cent from 1990 to 1993 while electricity consumption rose by 5.5 per cent (Statistics Finland). By the logic of the advocates of the power consumption thesis we are led to conclude that the hidden economy exploded in size over these three years. For example, if the hidden economy initially was five per cent of total output, then for electricity consumption to rise with no change in intensity of use the hidden economy would have had to grow by 319 per cent! This seems hard to believe. A more likely explanation is the decline in capacity utilization that occurs in recessions causes kilowatt hours of electricity per unit of GDP to increase.

Moreover, as shown by Alexeev and Pyle (2003) the frequently cited estimates of Johnson et al. (1997) assumed no growth in the size of the shadow economy of the Soviet Union from the late 1970s to the collapse of the system. (The same error is made by Aslund 2002, p. 122.) This

assumption is rejected by all observers of the Soviet economy. Hence, these empirical estimates of the growth in the shadow economy are based on too small an estimate of its initial size.

A second measurement problem in assessing the output fall arises because of the inadequacy of the inherited statistical system to cope with a market economy. Command economies, by their nature, focused on population statistics with regard to output. This is natural in a planned economy where the output produced was the result of a central plan. Indeed, the very nature of command required the planners to coordinate output, hence the statistical system needed to record what each enterprise produced. (Of course, in practice, this was difficult, as discussed in command economy.) The demise of the planning system weakened the authority of central statistical systems. More importantly, new entry became increasingly important in market economies, and the inherited statistical systems are not organized effectively to capture this.

It is also argued that under command systems enterprises had an incentive to overstate output in order to achieve bonuses, while firms in market economies want to hide output in order to avoid taxes (for example, Shleifer and Treisman 2004). It is thus argued that much output is simply missed by the change in the incentive to report. While it is certainly the case that firms have an incentive to hide output – especially when the financial system is undeveloped so they cannot seek external finance – the incentive to over-report under planning is less clear. Enterprises in planned economies were subject to the notorious ratchet effect. Higher production today meant higher output targets in the future – essentially a highly progressive dynamic tax system. The typical response to the ratchet effect was to produce only as much as needed to satisfy the plan. Hence, it is not at all clear that enterprises over-reported output in the command system.

A more important reason to question the magnitude of the output fall is the contraction in value-destroying activities. Because prices were distorted in planned economies, a portion of economic activity actually destroyed value at market prices. The contraction in these activities represents an increase in welfare, and correctly measured represents an increase in national income as well. The problem is that at the prices that prevailed in command economies this output appeared to be valuable; hence the contraction is measured as a fall in output.

There are two aspects to this decline. First, the separation of domestic from world prices means that activities that produce value added at domestic prices could destroy value at world prices. Given the underpricing of raw materials and overpricing of industrial goods characteristic of planned economies, this was more than a theoretical possibility. External liberalization then leads to a contraction of these activities (McKinnon 1991). The second aspect is that domestic prices were similarly distorted so that domestic price liberalization has a similar effect. This is discussed below.

To the extent that a reduction of value-destroying activity occurs at the same time as output falls, it is clear that movements in measured output are not consistent with movements in welfare. Indeed, if a greater measured output fall is associated with a faster removal of value-destroying activities, then it is likely that welfare is enhanced by the output fall. In this case the output fall is associated with more reform and quicker removal of welfare destroying activities. (This also means that output recovery could mean a resurgence of value-destroying activities, in which case the upward-sloping part of the U shape is welfare decreasing. Unlikely, but it might be relevant for Belarus under President Lukashenko). Of course, for this to be the case there must a serious distortion in national income measurements. To the extent that output measurements use base-weighted prices this is possible.

It is difficult to measure the extent to which the output fall is overstated by the contraction of value destroying activities. For example Aslund (2002, p. 126) estimates that about 20 per cent of GDP was value destroying in the last years of Communism. He uses, as an indicator, the decline in the share of industry in GDP. Soviet-type economies were over-industrialized, and liberalization led to sectoral shifts as services, which were previously undersupplied, expanded. Moreover, shifts in relative prices, discussed below, also

lead to a reduction in the share of value added produced by industry. Thus, one cannot infer value destruction from the change in industrial shares. The general problem is that output may be falling for various reasons so one cannot consider all of the contraction to be previously value destroying. One valuable indicator of the importance of value destruction is given by the comparison of the contraction in industrial output with the rise in consumption that occurred in transition economies. In Russia, for example, industrial output contracted by roughly 35 per cent from January 1992 to January 1994. Real disposable income, on the other hand, increased by almost 70 per cent in the same period (albeit from depressed levels). The fact that real disposable income was growing at the same time as industrial output was contracting suggests that the cessation of value-destroying activity was an important process, and that some of the output fall may be overstated.

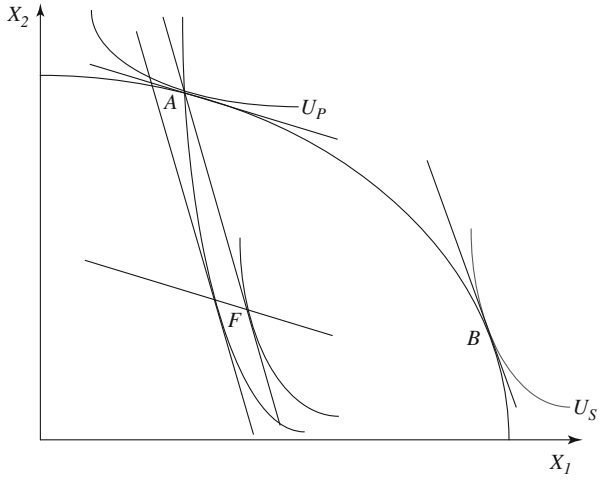
A related problem is the shift in preferences. Gaddy and Ickes (2003) argue that a specific index number problem leads to an overstatement of the output fall – the *camellia effect*. The argument is easily understood in terms of an analogy. Consider a flower shop that specializes in the sale of extremely rare camellias. Cultivating these plants is inordinately expensive, but this activity is profitable because the shop has a customer willing to pay very high prices for camellias. Now suppose this customer passes away. The shop can no longer sell rare camellias at a price that covers the cost of production. So camellia cultivation ceases. Resources that were previously devoted to camellia production will now be used for something else, say, roses. Profits at the flower shop fall because camellias were very profitable as long as their special customer lived. But given that there is no longer a market for rare camellias (while there is a market for roses, everyone is better off with rose cultivation than if they continued to cultivate camellias as if nothing had changed.) In the Soviet regime defence output was demanded despite the enormous cost. It had value as long as the Communist Party had command over resources. The special customer of Soviet times made it 'valuable' to produce defence output. When the Soviet system collapsed, so did the special customer. Output thus fell – valued at Soviet prices – because at those prices defence output was valued far above cost. After the fall this output is not valued sufficiently and production declines. This is an output fall, but welfare is certainly higher with lower defence production given that the Communist Party is no longer the measure of value.

To see this, suppose that we have two final goods, (x_1, x_2), and that the pre-transition production bundle is $\left(x_1^A, x_2^A\right)$, where good 2 is defence output, and A represents planners preferences. The post-transition allocation is $\left(x_1^B, x_2^B\right)$, and reflects social preferences. We might consider, for example, that at point A there is large military production and little civilian production, reflecting planners' preferences (U_P). The new production bundle is at point B, based on society's preferences. Note that using pre-transition prices to value output, GDP is $Y^A = \sum_i p_i^A x_i^A$.

Now suppose that liberalization causes the production bundle to move to point F in Fig. 3. This is the most pessimistic outcome – demand for x_2 declines with almost no increase in $x1$. Measured in real terms, at the old prices, output falls approximately by the distance AF in units of $x2$, or $\sum_i p_i^A x_i^F - \sum_i p_i^A x_i^A$. But this greatly overestimates the welfare change, because it places a high value on the output that has fallen in valuation.

Although output has fallen precipitously at planners' prices, measured at the new prices welfare has clearly increased. The minimum expenditure to achieve the old welfare level $e\left(P^B,\ U_A^P\right)$ is less than the cost of purchasing bundle F at the new prices. It is evident that welfare is higher at point F than at point A. Output has risen at the new prices but has fallen at the old prices.

From Fig. 3 we can also distinguish the fall in output due to coordination-type failure and that due to measurement. If resources are fully utilized we would be at point B. Hence $\sum_i p_i^B x_i^B - \sum_i p_i^B x_i^F \equiv \Omega$ measures the fall in output due to coordination-type failure. The measured fall in output, could be larger or smaller than this. The key point, however, is that the measured fall does not measure Ω at all.



Output Fall – Transformational Recession, Fig. 3 The camellia effect

Notice that, if the resources devoted to defence production are highly specialized, then there may be great inertia in response to the demand shift. It may be very hard to find alternative uses for these inputs. Output may remain depressed for quite a while. There may also be interesting behavioural issues to think about. A Russian defence enterprise director may expect that the government will soon restore orders and that cuts were temporary. This would lead to inertia in shifting to new activities. Both of these inertial forces could prolong the decline in output.

The importance of the camellia effect for thinking about the output decline is especially important in comparative terms. The camellia effect explains why transitional recessions are observed. But the size of this drop will be proportional to the share of 'camellias' in GDP, and this clearly differs across the post-Communist world. (Even for the former Soviet Union the differences are dramatic, as Russia had a much larger than average share of Soviet defence industry; see Gaddy 1996.)

In a country like Russia the size of the defence sector was especially large. This exacerbates the size of the output drop that is due to transitional factors. To measure the pure transition effect we should compare what would have been produced under central planning had planners' preferences not determined production decisions with what happened during transition. Ignoring the camellia effect mixes the two sources of output fall.

Theories of the Output Fall

Theories of the output fall in transition generally fall into one of two classes. The first class of theories treats this phenomenon as a sign of inefficiency. The output fall is thus welfare decreasing. The second class treats the output fall as a natural feature of liberalization but does not consider the fall to be welfare reducing. (One could also consider the specific negative shocks that have caused output disruptions. For central Europe there is the breakup of Council for Mutual Economic Assistance (CMEA) trade plus the end of subsidized energy from the Soviet Union. For the former Soviet Union there is the disruption in trade caused by the breakup of a common economic space into 15 independent countries. For Russia, there is the decline in oil prices. The importance of movements in the oil price for Soviet and Russian output has been emphasized by Gaddy and Ickes 2005. The power of this explanation has been fortified by the close timing of the recovery of Russian output with the increase in oil prices starting in the later 1990s.)

A basic framework for thinking about the output fall is the reallocation problem. Consider an

economy with two sectors, state (S) and private (P). Initially all labour is employed in the state sector. It is assumed that labour productivity in the private sector (β) exceeds that in the state sector (α), $\alpha < \beta$. The reallocation process occurs as labour moves from the state to the private sector. Per-capita output, y_t, is thus given by

$$y_t = \alpha \frac{L_t^S}{L_t} + \beta \frac{L_t - L_t^S}{L_t};$$

it is immediately apparent that rather than decline, output will increase monotonically in the transition. Hence, to obtain an output fall some unemployment of resources is necessary. If the private sector cannot absorb all the labour released from the state sector then labour will be unemployed, L^U. In that case per-capita output is given by

$$y_t = \alpha \frac{L_t^S}{L_t} + \beta \frac{L_t - L_t^S - L_t^U}{L_t}.$$

This simple framework suggests that to produce an output fall some rigidity or friction is required that prevents smooth reallocation of the labour released from the state sector. The essence of transition suggests that this will be likely. In addition to the normal culprits such as wage rigidity, institutional features play a critical role. For example, prior to the privatization of state sector assets, capital is immobile between sectors. This naturally limits the absorption rate of the private sector. Hence, the exit rate from unemployment will depend on the rate of growth of the private sector. What is important to understand are the determinants of the exit rates from these states. Notice that the growth of the private sector may depend on what is happening in the other sectors. This dependence can occur for several reasons. First, following Aghion and Blanchard (1994), unemployment can cause fiscal deficits which must be financed at the expense of the private sector, limiting its growth. Second, the growth of the private sector may depend on the rate at which complementary resources are released from the state sector. This is especially true for the most basic of resources for production, space. Until privatization of fixed capital takes place it is difficult for new private enterprises to obtain space for production, let alone to lease equipment.

At the most basic level, unemployment can be due to rigidity in real wages. But it is hard to understand how this can explain the output falls that were actually observed, as real wages fell in most transition economies once prices were liberalized. Hence the need for more fully developed theories.

Double Marginalization

Li (1999) develops a theory of the output fall in transition based on double marginalization. The basic idea is that the dismantling of central planning or centralized organization of production permits monopolistic and vertically interdependent enterprises to pursue their own monopoly profits by restricting output and intermediate trade to the detriment of the economy as a whole. The basic idea is that the collapse of planning institutions removes constraints on intermediate producers' activities. Intermediate producers now have monopoly power, so they raise prices. This happens all along the supply chain, and results in an increase in the cost of producing final output. So there is less final output available and government output falls. The essential reason is that the enterprises do not consider the consequences of their price increases for the profits of the other enterprises. Since there is less left over for consumers, it is equivalent to a decrease in real wages, and hence labour supply falls.

The essential idea of the double marginalization theory is that output falls because liberalization precedes the development of competition. Entry is a process that takes time. Hence, the theory would predict that output falls would be greater in economies that are less able to 'import' competition through opening the economy. This roughly fits the picture of larger output falls in the FSU than in the CEEs. But the theory also predicts that the output fall should be largest when liberalization first takes place, since that is when market power is most potent. The effect of double marginalization should wane over time. This is harder to reconcile with the paths of output in Figs. 1 and 2.

The double marginalization model also predicts that each enterprise will face a contraction in demand and an increase in input prices relative to wage rate. The contraction in demand is attributable to the following factors in this model: the decline in real wage rate, the decline in the government's real income and the decline in input demand. The increase in input prices relative to wage rate is attributable in this model to monopoly pricing by a 'web of monopolies'. The more complex is the web of inter-industry production, the greater the propagation of the price shock. Hence, complexity magnifies any intermediate price markup throughout the economy, resulting in higher input prices relative to wage rate. The sharp increase in input costs is indicative of a sharp supply contraction. This prediction is also consistent with empirical observations.

Disorganization

Blanchard and Kremer (1997) (see also Blanchard 1997) have developed a model of disorganization that has had great impact. Their argument is that the output fall is a result of the chaos that surrounds the elimination of central planning. They focus on three mechanisms (hold-up problems, coordination, and uncertainty problems) that are greatly magnified as the result of missing institutions likely to be important at the start of transition. The basic idea is that the collapse of planning causes performance to decline during the period when alternative market mechanisms have not yet developed.

The basic idea can be understood in terms of a simple example presented by Blanchard and Kremer. Consider a vertical chain of production. Assume that each step is carried out by a different enterprise. A unit of a primary good is needed at the first step. At the end of the n steps one unit of the final good results, and we normalize the price of this good to unity. The value of the intermediate output, at each step, is zero. The supplier of the primary input has an alternative use, which is c. This could be much lower than one. It is a private opportunity that could be exporting the good, or selling it for a less fabricated use. Under planning the relations in the chain were directed from above. With liberalization alternative activities may be considered.

The end of planning thus leads to n bargaining problems. Each unit must bargain with a supplier and a customer. They assume that there is Nash bargaining at each step, so that the surplus is split given the symmetry of the situation. To see what happens start with the last step. The value of the surplus in the last stage (bargaining between the final producer and the last intermediate producer) is 1. This follows because the value of the good at stage n is still zero. So the last intermediate producer gets one half of the surplus. Similar bargaining takes place at all the upstream stages. At the $n-1$ stage there is one half to split ... Continue in this fashion and it follows that the first intermediate producer gets $\left(\frac{1}{2}\right)^n$. The surplus available to split at the first stage is $\left(\frac{1}{2}\right)^n - c$, since the first producer must purchase the primary input to produce. It is thus clear that unless $c < \left(\frac{1}{2}\right)^n$ the raw material will be diverted and production will cease. Moreover, c does not need to be all that large to trigger defection that results in a fall in output that could be as large as $1 - \left(\frac{1}{2}\right)^n$. Thus rather meagre private opportunities can cause a rather large fall in output.

Blanchard and Kremer interpret n as the level of complexity of production. As n increases, the likelihood of defection increases exponentially. This is a hold-up problem. Each producer in the chain must produce before bargaining with the next in line. This suggests that the problem would go away if each of the producers could sign an enforceable contract before production takes place. As long as $c < 1$, defection could be avoided and production could take place, if the intermediate producers could sign a contract to split the $1 - c$ before production. The problem is thus one of asset specificity and incomplete contracts. Eliminating the ministry before institutions that support contracts are developed is the source of the problem. Vertical integration could help, but this requires ownership to be specified, another problem early in transition. The notion that producers in transition could suffer from this problem is not far-fetched. (It is interesting to compare this outcome with the *double marginalization* case. Notice that in that case the raw

materials producer has market power and thus a higher share of the surplus than is the case in the bargaining problem. This makes production in the state sector more likely. Of course, what is not explained is why the producer is able to extract monopoly rents in a situation of bilateral monopoly.)

Blanchard and Kremer consider other examples based on incomplete information. A state-owned enterprise must negotiate with many suppliers that may have outside options. Each of the suppliers produces a key input without which production is impossible. With uncertainty over the magnitude of outside options a state-owned enterprise must guess how much to pay for the inputs. When outside opportunities are low the possibility that the state-owned enterprise offers too low a price is negligible. But as these outside opportunities rise this probability increases. Even if it is still efficient to sell to the state-owned enterprise because of uncertainty over the size of these options, the price offered may be too low and production falls. The interesting feature of this model is that it produces a U-shaped output path. The key assumptions are technological complementarities and inefficient bargaining.

A coordination example can also be constructed. Suppose that the firm needs n workers (it could be supplying firms, but this is easier), and the technology is Leontief. If all workers stay, the firm produces one unit of output per worker. If a worker leaves, a replacement is hired with output per worker equal to $\gamma < 1$. Here again n measures the degree of complexity, while γ is an inverse measure of the specificity of the production process or job-specific human capital.

Each worker has an alternative opportunity given by c, distributed on $[0, \ \overline{c}]$, where $\overline{c}$ represents the maximum outside opportunity, which is of course a function of the state of the transition. Draws from this distribution are independent across workers. The distribution is known, but the specific realization is private information. This could be thought of as alternative employment, perhaps in a Western multinational. The firm pays a common wage, w, to all workers, equal to output per worker. This simplifies the analysis, but is probably not crucial.

The key assumption of the model is that workers must decide whether to take up the alternative before they know the decision of the other workers. This creates the coordination problem. Workers are risk neutral, so that all we need to look at is expected output. There are thus two potential outcomes: (*a*) all workers stay, output per worker and thus the wage are equal to unity, or; (*b*) one or more workers leave, output per worker and the wage are equal to γ.

The decision problem for the agents boils down to determining some threshold level of outside opportunities, c^*, such that if $c < c^*$, workers stay and vice versa. If a worker leaves he receives c. If he stays his expected earnings will depend on what the other $n - 1$ workers do. Assume symmetry so that the other workers also have the same c^*. Then the probability that they all stay is $(F(c^*))^{n-1}$, where $F(\cdot)$ is the distribution function so that $F(0) = 0$ and $F(\overline{c}) = 1$. Expected output per worker is thus equal to $(F(c^*))^{n-1} + \gamma[1 - (F(c^*))^{n-1}]$.

The key point is that there may be multiple equilibria, depending on the level of outside opportunities. If alternative opportunities are very low, workers always stay in the firm, and output equals 1. As outside opportunities increase there are two equilibria; in one of these output falls close to γ. With very high outside opportunities production in the state sector ceases. Note the problem here is coordination, not uncertainty. If the outside opportunity were common knowledge, with $\gamma < c < \overline{c}$ there would still be two equilibria.

The essential feature of the disorganization model is that central planning is replaced before the infrastructure of markets is created. The lack of central organization leads to disorganization, and the development of outside opportunities makes this problem more severe. Over time, market infrastructure develops and disorganization problems are lessened.

Roland and Verdier (1999) develop a related model of disorganization, focusing on search frictions rather than bargaining problems. In their model liberalization means that enterprises can search for new suppliers and customers. There are good matches and bad matches. If too many

bad clients are searching the productivity of potential matches may fall. What is critical in their model is that relationship-specific investments take place only after long-term matches are formed. If search continues this will not happen, investment demand will fall, and output can fall.

Investment specificity is crucial in this model. Without it output would not fall even with bad matches, since the partners could produce this period and keep on searching. It is the asset specificity that introduces the cost of bad matches.

The Roland–Verdier model is interesting from a theoretical point of view, but one may wonder how relevant it really is for explaining the output fall. The problem is that the initial output fall was associated with very little search for new suppliers. The predominant behaviour was a relationship-conservatism. Agents tried to maintain their relationships as much as possible. Networks of suppliers already had relationship-specific investments. The problem is that they had no customers who would purchase the goods at a price that covered their new costs.

Micro-distortions

A more subtle, but equally important explanation of the output fall focuses on the micro distortions due to Soviet pricing rules. Ericson (1999) has analysed this problem. His focus is on structural problems with Soviet pricing – the arbitrariness and non-uniformity of producers' prices across users of the product within standard commodity aggregates. Ericson shows that Soviet pricing rules hid inefficiency and waste, creating an illusion of capacity and output that wasn't there. The advantage of this theory is that it can explain why prices exploded when output fell. His argument is that post-Soviet 'stagflation' is, to some extent, a consequence of the irrational structure of production hidden in apparently consistent (adjusted) input–output (I–O) matrices and economic statistics.

Soviet pricing rules contained three systematic distortions: (*a*) basic factors were seriously undervalued (land was free, and capital-in-place virtually so); (*b*) raw materials and natural resources were undervalued; and (*c*) highly processed goods – in particular investment products and services – were seriously overvalued. These distortions in the principles of economic valuation used in centrally planned economies systematically hide tremendous waste, exaggerating both net outputs and net income (economic value) produced, while understating the productivity of that most seriously mismeasured factor of production, capital. This implies that the size of the apparent initial collapse in industrial production is evidently exaggerated, even if one ignores new economic activity generated in the wake of the reforms. However, the wasteful production structure can also spur a continuing and deepening collapse, as it is not economically viable in a market environment.

Ericson shows that embedding these distortions in the input–output tables that are used to create national income statistics results in lower prices for inputs than for final uses, and generates an understatement of the share of gross output used in the production process. Thus, it leads to an overstatement of the share of net output. Furthermore, these distortions cannot be revealed by any consistent input–output framework derived from the 'value' of transactions between sectors; the methodology itself imposes a consistency that hides those distortions. This means that the true nature of the system cannot be revealed until price liberalization takes place. Until then, intersectoral relationships are hidden. This is what creates the 'circus mirror' effect discussed by Gaddy and Ickes (2002). (A circus mirror distorts size and shape. Soviet pricing rules had the same effect, making value added look larger and intermediate input use look smaller). Just as an individual may look taller and thinner in a circus mirror, the Soviet-type economy appeared more productive under Soviet pricing rules. Liberalization revealed the true nature of the economy.

Ericson shows that for the case of Russia the 1991 input–output coefficients were substantially understated, hiding significant materials input use and waste, and hence obscuring much of the inherited inefficiency in the industrial structure. This inefficiency became of consequence for

producers when liberalization released them from ministerial tutelage and constraints, and made them primarily responsible for covering their own costs. Because enterprises are initially constrained by existing technological structures, the first impact of liberalization is typically seen in the move to raise prices to cover their full material costs and to compensate for any increases. This led to increases in industrial prices that far exceeded the general rate of inflation, raising the real price of industrial output and consequently real materials costs. As in the double marginalization theory, price increases in the intermediate sector propagate through the economy and result in less final output. But the impulse is different. Ericson's theory does not require any market power on the part of intermediate producers. Price increases are solely due to price liberalization itself in the context of Soviet pricing. Of course, at those increased real prices, demand for many products, now not supported by plan requirements, falls dramatically; producers find they are unable to sell at higher prices and hence unable to recover the full costs of production. Yet they continued to operate and ship output to traditional users of their product.

Ericson's theory is thus consistent with several important aspects of the output fall that are hard to explain in other models. First, his theory explains why the output fall is associated with a rise in the price level. Second, it is consistent with higher wholesale price inflation than consumer price inflation. Third, it is consistent with the explosion of inter-enterprise arrears. Supply and disorganization type theories make no prediction with regard to overall inflation and they are inconsistent with the latter two observations.

Empirical Analysis

Most empirical analyses of the output fall has been focused on assessing the role of policies (primarily, stabilization and liberalization) and initial conditions in determining the size of the fall in output. This literature is too large to summarize here (a good summary is Campos and Coricelli 2002), but a few points can be made. First, results are very dependent on how policies, especially the speed and extent of liberalization, are measured, and how initial conditions are proxied. Measures of liberalization that rely on expert evaluation are subject to performance bias: that is, the liberalization score that is assessed is often inferred from economic performance. The set of initial conditions that are important include the degree of over-industrialization, repressed inflation, dependence on CMEA trade, distance from Frankfurt, years spent under Communism, initial income, and the rate of urbanization. Depending on the set used results can differ dramatically.

One of the most comprehensive studies of the impact of policies versus initial conditions is by Berg et al. (1999). They use a sample of 26 transition economies and use a general to specific modelling approach that allows for differential effects of policies and initial conditions and for time-dependent effects of initial conditions. They find that structural reforms are more important than either policies or initial conditions in explaining the cross-country variation in performance. Initial conditions play the predominant role in explaining the output fall, while structural reforms explain the recovery. The most important initial conditions appear to be the degree of over-industrialization and trade dependency.

Conclusion

Although the size of the output fall indicated by official measures is clearly overstated, the fact that output and incomes did fall in the aftermath of liberalization is not disputed. Moreover, the fact that output followed a U-shaped pattern has had important consequences for transition. Not least of these is the negative effect it had on the political support for many economic reformers. The output decline made it politically difficult to stick with reforms. Hence, the output declines may have altered the course of policy reform in transition. Ironically, it seems that reform reversals were often associated with longer output declines.

See Also

- ▶ Command Economy
- ▶ Institutional Trap
- ▶ Second Economy (Unofficial Economy)
- ▶ Soft Budget Constraint
- ▶ Virtual Economy

Bibliography

Aghion, P., and O. Blanchard. 1994. On the speed of transition in central Europe. In *NBER macroeconomics annual*. Cambridge, MA: MIT Press.

Alexeev, M., and W. Pyle. 2003. A note on measuring the unofficial economy in the former Soviet republics. *Economics of Transition* 11: 153–175.

Aslund, A. 2002. *Building capitalism: The transformation of the former Soviet bloc*. Cambridge: Cambridge University Press.

Berg, A., E. Borenzstein, R. Sahay, and J. Zettelmeyer. 1999. The evolution of output in transition economies: explaining the differences. Working Paper No. 99/73, International Monetary Fund.

Blanchard, O. 1997. *The economics of transition in Eastern Europe*. Oxford: Oxford University Press.

Blanchard, O., and M. Kremer. 1997. Disorganization. *Quarterly Journal of Economics* 112: 1091–1126.

Campos, N., and F. Coricelli. 2002. Growth in transition: What we know, what we don't know, and what we should. *Journal of Economic Literature* 40: 793–836.

Ericson, R.E. 1999. The structural barrier to transition hidden in input–output tables of centrally planned economies. *Economic Systems* 23(3): 199–224.

Fischer, S., and R. Sahay. 2000. The transition economies after ten years. Working Paper No. 7664. Cambridge, MA: NBER.

Gaddy, C.G. 1996. *The price of the past*. Washington, DC: Brookings Institution Press.

Gaddy, C.G., and B.W. Ickes. 2002. *Russia's virtual economy*. Washington, DC: Brookings Institution Press.

Gaddy, C.G., and B.W. Ickes. 2003. How to think about the post-Soviet output fall. Working paper, Brookings Institution and Pennsylvania State University. Online. Available at http://econ.la.psu.edu/Bbickes/OutputFall.pdf. Accessed 20 May 2007.

Gaddy, C.G., and B.W. Ickes. 2005. Resource rents and the Russian economy. *Eurasian Geography and Economics* 46: 559–583.

Johnson, S., D. Kaufmann, and A. Shleifer. 1997. The unofficial economy in transition. *Brookings Papers on Economic Activity* 1997(2): 159–239.

Li, W. 1999. A tale of two reforms. *RAND Journal of Economics* 30: 120–136.

McKinnon, R. 1991. *The order of economic liberalization: Financial control in the transition to a market economy*. Baltimore: Johns Hopkins Press.

Mundell, R.A. 1997. The great contractions in transitions economies. In *Macroeconomic stabilization in transition economies*, ed. M.I. Blejer and M. Skreb. New York: Cambridge University Press.

Popov, V. 2005. Shock therapy versus gradualism reconsidered: Lessons from transition economies after 15 years of reforms. TIGER Working Paper No. 82.

Roland, G., and T. Verdier. 1999. Transition and the output fall. *Economics of Transition* 7: 1–28.

Shleifer, A., and D. Treisman. 2004. A normal country. *Foreign Affairs* 83(2): 20–38.

Statistics Finland. Online. Available at http://www.tilastokeskus.fi/index_en.html Accessed 28 June 2007.

Svejnar, J. 2002. Transition economies: Performance and challenges. *Journal of Economic Perspectives* 16(1): 3–28.

Zettelmeyer, J. 1998. The Uzbek growth puzzle. Working Paper No. 98/133, International Monetary Fund.

Overhead Costs

Basil S. Yamey

John Maurice Clark wrote in 1923 that the term overhead costs is 'variously used', although there is the central underlying concept that these costs are 'costs that cannot be traced home and attributed to particular units of business in the same direct and obvious way in which, for example, leather can be traced to the shoes that are made from it'. 'Most of the real problems' stem from the fact 'that an increase or decrease in output does not involve a proportionate increase or decrease in cost' (1923, p. 1). The notion of overhead costs is similar to that of Alfred Marshall's 'supplementary costs', that is, charges or expenditures that, unlike 'prime' or 'direct' costs, 'cannot generally be adapted quickly to changes in the amount of work there is for them to do' (1920, p. 360). Thus overhead costs, a term used infrequently in economic theory or analysis nowadays, are akin to the more familiar 'fixed costs' (as in the variable costs/fixed costs dichotomy). However, the feature that they cannot be traced directly to particular units of output or activities gives overhead costs some of the flavour of common or joint costs.

Overhead costs do not raise special questions for economic theory that do not arise in connection with fixed costs or common costs generally. They are evidently important for many issues in applied economics. Two well-known books with 'overhead costs' in their titles consist largely of studies on subjects such as transport, public utilities, two-part tariffs and competition in retailing (Clark 1923; Lewis 1949).

This essay concentrates on the treatment of overhead costs in modern cost accounting, which dates from the second half of the 19th century. It became standard practice in cost accounting for overheads or oncost to be allocated to units or batches of production or to departments or divisions within the firm. Thus the total cost of, say, a batch of production is ascertained by accumulating the direct costs of that batch and adding an 'allocation' of overheads. Allocation of overheads has been made on a variety of bases. In a surviving 15th century set of cost calculations, the costs common to a number of products were allocated according to the weights of those products. In 1890 Marshall observed two allocation bases. He wrote that in 'some branches of manufacture it is customary to make a first approximation to the total cost of producing any class of goods, by assuming that their share of the general expenses of the business is proportionate either to their prime cost, or to the special labour bill that is incurred in making them' (1920, p. 195). Several other bases have been advocated and used. Further elaboration has been achieved by subdividing overheads into categories (e.g., manufacturing and distribution overheads) and sub-categories, and by using different bases for different categories. Yet more elaboration has been introduced by calculating overhead allocations on the basis either of a 'standard' output or of the expected output in the period in question. There has been much discussion about which bases of allocation are 'fair', 'reasonable' or 'appropriate', and whether, for example, interest on capital is an element of overhead that should be allocated in the cost accounts, now generally called management accounts.

Economists have criticized the accounting treatment and have argued that the allocation of overheads costs necessarily is arbitrary; that these costs do not form part of short-run marginal costs; that costs in cost accounting refer to past costs; and that for all these reasons accounting figures purporting to measure 'total costs' are at best irrelevant for output, pricing and investment decisions, and at worst may mislead the decision-maker who is not aware of their make-up (e.g., Solomons 1952, esp. articles by R.S. Edwards, R.H. Coase, W.T. Baxter and D Solomons). The accounting treatment of these costs can be especially misleading when the efficiency or performance of a manager of a department (or division or product group) in a firm is judged on the basis of his department's recorded profit; this profit will in part depend upon allocated costs for which he has no responsibility and over which he has no control.

Recognition of the implications of the conventional treatment has caused many accountants and firms to abandon or disregard the allocation of overheads in management accounting. Instead, emphasis in the accounts is placed on the determination of the 'contribution' to fixed overheads and profits made by the particular product, division or department, namely the difference between the revenues generated and the sum of the variable costs incurred. In the same spirit, in compiling accounting information bearing on the performance of a manager, his account is not charged with allocations of those overheads over which he has no control. The various categories of overhead costs are budgeted and monitored directly, a distinction being made between those that are fixed for the period in question and those that are to some extent variable. Concentration of attention on the contributions made by (or budgeted for) particular products and activities is found in some firms in which, nevertheless, overhead costs are allocated in the traditional way in accordance with selected allocation formulae.

It is a well-known proposition, to quote Marshall again, that 'it is of course just as essential in the long run that the price obtained should cover general or supplementary costs as that it should cover prime costs' (1920, p. 420). Economists and many accountants say, in effect, that the allocation of overheads to products or activities cannot

contribute rationally to the long-run 'recovery' of overheads in pricing and output decisions.

However, an economic and business rationale for overhead cost allocations has been considered occasionally (recently again in Zimmerman 1979). The rationale concerns firms in which authority to make certain decisions is assigned by the central management to otherwise subordinate managers, such as managers of particular product groups or of geographically dispersed sales offices.

In varying degrees, managers make use of the assets owned by the firm and of services supplied by other divisions or departments within the firm. Managers compete for internallysupplied resources and services. If a manager's performance is judged wholly or in part on the basis of the accounting profits he achieves, and *a fortiori* if his remuneration is related to his profits, he has an incentive to use internally supplied inputs as if they were free goods unless his account is in some way charged for them. This gives rise to waste in the use of resources and services. Demands by one manager on the services provided by the firm's assets and by other parts of the organization may deflect these services from more profitable uses within the firm.

Overhead allocations may serve, it is argued, as a set of internal prices to be 'paid' by a manager for the use of inputs supplied to him within the organization. Provided that the prices (the overhead allocation rate or amount) for the various inputs appropriately reflect opportunity costs, the internal price system within the firm together with profit-maximizing behaviour of managers will (*ex ante*) maximize the profits for the firm as a whole from its available resources, and will dispense with the need for administrative controls and rationing.

To reflect opportunity costs properly, the amount of overhead to be allocated to users would have to be adjusted in the light of the changing level (and expected level) of internal demand for the services of the resources in question. The appropriate amount to be allocated will almost invariably not be the actual outlays incurred by the firm, nor be related in any way to those outlays. For example, the amount should be zero if the cost is fixed and the underlying resources are not expected to be used fully in the relevant period; and it should exceed actual outlays when there is excess demand. Further, the overheads allocated to a particular user department should closely reflect that department's consumption of the services in question, and should *ceteris paribus* be smaller when the usage is lower.

Systems or schemes used in practice for the allocation of overheads do not generate accounting charges that are sensitive to changes in internal and external market conditions or to variations in the rate of consumption of services. For overhead cost allocations to perform an efficient rationing function would require a different approach from that generally adopted in management accounting systems.

What is required is a set of shadow prices, revised whenever there has been (or is expected to be) a material change within the firm in the supply and demand conditions for the services in question. Methods for estimating the shadow prices might range from the exercise of judgement by experienced managers to the use of mathematical programming techniques.

It may seem that the inclusion of allocations of overhead costs in cost data would help decision-makers in a firm to assess the level of prices for its products at which new competitors might be attracted to enter the market. However, the calculation of the established firm's own average costs (variable *and* fixed costs, including sunk costs) can serve as no more than a rough guide for this purpose, since allowance has to be made, for instance, for new methods available to new entrants and for learning costs.

Allocations of overhead costs to particular products or transactions come into their own when prices are determined by formula and not by the market (as in some defence contracts), or where a government agency engages in control of maximum prices for whatever reason. Again, allocations may be crucial when a regulatory agency has to ensure that a regulated enterprise does not engage in cross-subsidizing its unprotected activities from the profits of its protected activities. They may also be critical when a regulatory agency seeks to determine whether a multiproduct firm has been charging monopoly prices for one of its several products, or whether a firm has

discriminated in the prices for its products sold to different customers. In the past, also, cooperative schemes designed to reduce the intensity of price competition have included the adoption by members of an industry of a uniform costing system. Such systems have involved the use of standard methods for the allocation of overheads to products (Solomons 1950). In all these cases, the arbitrary nature of the allocations cannot be escaped; and the bases of allocation to be adopted provide much scope for ingenuity in argument.

See Also

- ▶ Accounting and Economics
- ▶ Fixed Factors

Bibliography

Clark, J.M. 1923. *Studies in the economics of overhead costs*. Chicago: University of Chicago Press.
Lewis, W.A. 1949. *Overhead costs*. London: George Allen & Unwin.
Marshall, A. [1890] 1920. *Principles of economics*, 8th edn. London: Macmillan.
Solomons, D. 1950. Uniform cost accountancy – A survey. *Economica* 17: 237–253, 386–400.
Solomons, D. (ed.). 1952. *Studies in costing*. London: Sweet & Maxwell.
Zimmerman, J.L. 1979. The costs and benefits of cost allocations. *Accounting Review* 54: 504–521.

Over-Investment

Michael Bleaney

The term 'over-investment' is used principally in relation to a certain type of theory of the trade cycle in industrial capitalist economies. Such theories flowered into a brief prominence in the inter-war period, but disappeared virtually without trace after 1940, probably owing to the fact that they did not attach sufficient weight to the concept of effective demand. The key characteristic of these 'over-investment' theories was their stress on a disproportionate development of the producer goods industries not only as a feature of the boom but also as a cause of the subsequent relapse into depression. Since a similar pattern has been observed in some socialist economies since 1945 and has been held by many authors to be the major cause of fluctuations in the growth rates of real output, 'over-investment' theories of cycles in socialist economic systems will also be discussed.

According to Haberler (1937), whose work constitutes the best survey of the trade cycle literature of this period, one may distinguish a monetary and a non-monetary strand of over-investment theory. Prominent amongst the former were L. Mises and F.A. Hayek; amongst the latter, A. Spiethoff and G. Cassel. The monetary strand followed up Wicksell's idea that the monetary authorities could cause the money rate of interest to deviate from the equilibrium rate (or natural rate in Wicksell's terminology) which brought planned savings and investment into equality, thus causing investment intentions to get out of balance with the savings plans of the community. Mises regarded ideological and political pressures on central banks to maintain low interest rates as the main initiating cause of trade cycles. Hayek (1933, p. 150) was sceptical about this, and preferred to stress changes in the economic environment (such as new inventions) which create new investment opportunities. These developments would raise the natural rate of interest, but the ability of the commercial banks to create money means that these new demands for credit are initially met at the existing rate of interest.

Either way, a disequilibrium is set up in which demand for investment goods is out of balance with the demand for consumer goods. In the boom the investment goods industries are over-developed, and the pressure on resources will pull up production costs. In the absence of a sufficient further monetary expansion many investment projects will begin to seem ill-judged, and will not be completed. Investment falls off dramatically, and a slump ensues. A distinctive feature of these theories was their tendency to see slumps as a necessary process of purging the economy of the maladjustments created in the course of the booms. This idea is articulated most clearly by Hayek (1933,

pp. 19–22), who argues that an expansionary monetary policy, far from curing such slumps, merely prolongs them by delaying the necessary readjustments. Hayek interprets the slump of 1929–33 in this fashion, following the boom of 1927–9.

The non-monetary theorists laid greater stress on the real factors which might cause investment to rise at the start of a boom; they were generally willing to concede that a credit expansion was a necessary permissive factor in this, but they did not see monetary disturbances as the prime cause of the trade cycle. Since in Hayek's theory (but less so in that of Mises) money is a permissive as much as an initiating factor (because of the elasticity of bank credit), the monetary/non-monetary distinction is ultimately of little importance. The distinctive feature of the over-investment theories is the presumption that the boom constitutes a misdirection of productive resources towards the investment goods industries as compared with the equilibrium situation, and that the deflationary aspects of depression are necessary to cure this. In the Hayekian theory it is perfectly possible to imagine a slump caused by under-investment, where the natural rate of interest falls below the money rate; this possibility is however very much down-played in the analysis.

A major difficulty with this theory was that it never achieved a really satisfactory explanation of the *necessity* of depression in purging the maladjustments of the boom. Why could equilibrium not be re-established without the 'overshooting' effect of relapse into depression? A further set of questions was raised by the publication of Keynes's *General Theory*, which implied that depressions were pre-eminently conditions of low effective demand and unused productive capacity. If this were so, then the slump would not cure a state of over-investment; it would exacerbate it. Finally, the over-investment theory was too restrictive even on its own terms. Suppose that the natural rate of interest rose, not as a result of the opening-up of new investment opportunities, but as a result of a fall in the community's desire to save. The enhanced consumption demand would generate boom conditions in a manner similar to that analysed for increased investment. But (at least initially) this boom would not be characterized by over-expansion of the producer goods industries relative to consumer goods, but precisely the opposite. It would be an under-investment rather than an over-investment boom.

Theories of investment cycles in socialist economies have been based on experience in eastern Europe since 1950 (Bajt 1971). These theories could be termed 'over-investment' theories to the extent that they perceive these economies to be organized in such a way as to create a situation of persistent excess demand, and to give priority to investment over consumer demand in cases of shortage. This means that in periods when investment plans are unusually ambitious, the excess demand and shortage of consumer goods become exceptionally acute, resulting in delayed completion of investment projects, popular dissatisfaction etc. In reaction to this, the investment tempo is deliberately reduced; but once the problem has been solved, political pressures for more ambitious plans may build up once again.

The tendency towards over-investment results from the particular institutional circumstances rather than the mere fact that the means of production are mostly in public ownership. The important features are: the 'softness' of enterprise budget constraints, which enables them to win any competition for resources with consumers; insufficient penalties to the enterprise for unprofitable investments; deliberate understatement of the costs of investment projects in order to obtain scarce investment credits; and insufficient information in the hands of the central planners to enable them to counteract these tendencies effectively (Kornai 1980). Nevertheless there seems no reason why, by learning from experience, the planners should not be able at least to reduce the amplitude of such cycles.

See Also

► Investment (Neoclassical)
► Trade Cycle

Bibliography

Bajt, A. 1971. Investment cycles in European socialist economies. *Journal of Economic Literature* 9: 53–63.
Haberler, G. 1937. *Prosperity and depression*. Geneva: League of Nations.
Hayek, F.A. 1933. *Monetary theory and the trade cycle*. London: Jonathan Cape.
Kornai, J. 1980. *The economics of shortage*. Amsterdam: North-Holland.

Overlapping Generations Model of General Equilibrium

John Geanakoplos

Abstract

The OLG model of Allais and Samuelson retains the methodological assumptions of agent optimization and market clearing from the Arrow–Debreu model, yet its equilibrium set has different properties: Pareto inefficiency, multiplicity, positive valuation of money, and a golden rule equilibrium in which the rate of interest is equal to population growth (independent of impatience). These properties are shown to derive not from market incompleteness, but from lack of market clearing 'at infinity': they can be eliminated with land or uniform impatience. The OLG model is used to analyse bubbles, social security, demographic effects on stock returns, the foundations of monetary theory, Keynesian vs. real business cycle macromodels, and classical vs. neoclassical disputes.

Keywords

Agent optimization; Allais, M.; Animal spirits; Arrow–Debreu model of general equilibrium; Backward induction; Bubbles; Cobb–Douglas functions; Comparative statics; Consumption loan model; Continuum of equilibria; Cores; Demography; Double coincidence of wants; Equilibrium; Existence of equilibrium; Expectations sensitivity hypothesis; Impatience; Incomplete markets; Indeterminacy of equilibrium; Infinite horizons; Involuntary unemployment; Keynesianism; Marginal utility of money; Market clearing; Money; Multiple equilibria; New classical macroeconomics; Numeraire; Overlapping generations models of general equilibrium; Pareto efficiency; Pareto inefficiency; Perfect foresight; Price normalization; Samuelson, P. A.; Sequential equilibrium; Social security; Sraffa, P.; Sunspots; Uncertainty; Uniform impatience; Uniqueness of equilibrium

JEL Classifications

E1

The consumption loan model that Paul Samuelson introduced in 1958 to analyse the rate of interest, with or without the social contrivance of money, has developed into what is without doubt the most important and influential paradigm in neoclassical general equilibrium theory outside of the Arrow–Debreu economy. Earlier Maurice Allais (1947) had presented similar ideas which unfortunately did not then receive the attention they deserved. A vast literature in public finance and macroeconomics is based on the model, including studies of the national debt, social security, the incidence of taxation and bequests on the accumulation of capital, the Phillips curve, the business cycle, and the foundations of monetary theory. In this article I give a hint of these myriad applications only in so far as they illuminate the general theory. My main concern is with the relationship between the Samuelson model and the Arrow–Debreu model.

Allais's and Samuelson's innovation was in postulating a demographic structure in which generations overlap, indefinitely into the future; up until then it had been customary to regard all agents as contemporaneous. In the simplest possible example, in which each generation lives for two periods, endowed with a perishable commodity when young and nothing when old, Samuelson noticed a great surprise. Although each agent could be made better off if he gave half his youthful birthright to his predecessor, receiving in turn half

from his successor, in the marketplace there would be no trade at all. A father can benefit from his son's resources, but has nothing to offer in return.

This failure of the market stirred a long and confused controversy. Samuelson himself attributed the suboptimality to a lack of double coincidence of wants. He suggested the social contrivance of money as a solution. Abba Lerner suggested changing the definition of optimality. Others, following Samuelson's hints about the financial intermediation role of money, sought to explain the consumption loan model by the incompleteness of markets. It has only gradually become clear that the 'Samuelson suboptimality paradox' has nothing to do with the absence of markets or financial intermediation. Exactly the same equilibrium allocation would be reached if all the agents, dead and unborn, met (in spirit) before the beginning of time and traded all consumption goods, dated from all time periods, simultaneously under the usual conditions of perfect intermediation. Indeed, in the early 20th century Irving Fisher (1907, 1930) implicitly argued that any sequential economy without uncertainty, but with a functioning loan market, could be equivalently described as if all markets met once with trade conducted at present value prices.

Over the years Samuelson's consumption loan example, infused with Arrow–Debreu methods, has been developed into a full-blown general equilibrium model with many agents, and multiple kinds of commodities and production. It is equally faithful to the neoclassical methodological assumptions of agent optimization, market clearing, price taking, and rational expectations as the Arrow–Debreu model. This more comprehensive version of Samuelson's original idea is known as the overlapping generations (OLG) model of general equilibrium.

Despite the methodological similarities between the OLG model and the Arrow–Debreu model, there is a profound difference in their equilibria. The OLG equilibria may be Pareto suboptimal. Money may have positive value. There are robust OLG economies with a continuum of equilibria. Indeed, the more commodities per period, the higher the dimension of multiplicity may be. Finally, the core of an OLG economy may be empty. None of this could happen in any Arrow–Debreu economy.

The puzzle is: why? One looks in vain for an externality, or one of the other conventional pathologies of an Arrow–Debreu economy. It is evident that the simple fact that generations overlap cannot be an explanation, since by judicious choice of utility functions one can build that into the Arrow–Debreu model. It cannot be simply that the time horizon is infinite, as we shall see, since there are classes of infinite horizon economies whose equilibria behave very much like Arrow–Debreu equilibria. It is the combination, that generations overlap indefinitely, which is somehow crucial. In Section 4, "Understanding OLG Economies as Lack of Market Clearing at Infinity" I explain how.

Note that in the Arrow–Debreu economy the number of commodities, and hence of time periods, is finite. One is tempted to think that, if the end of the world is put far enough off into the future, it could hardly matter to behaviour today. But recalling the extreme rationality hypotheses of the Arrow–Debreu model, it should not be surprising that such a cataclysmic event, no matter how long delayed, could exercise a strong influence on behaviour. Indeed, the OLG model proves that it does. One can think of other examples. Social security, based on the pay-as-you-go principle in the United States in which the young make payments directly to the old, depends crucially on people thinking that there might always be a future generation. Otherwise the last generation of young will not contribute; foreseeing that, neither will the second-to-last generation of young contribute, nor, working backward, will any generation contribute. Another similar example comes from game theory, in which cooperation depends on an infinite horizon. On the whole, it seems at least as realistic to suppose that everyone believes the world is immortal as to suppose that everyone believes in a definite date by which it will end. (In fact, it is enough that people believe, for every T, that there is positive probability the world lasts past T.)

In Section 1, "Indeterminacy and Suboptimality in a Simple OLG Model", I analyse a simple one-commodity OLG model from the

present value general equilibrium perspective. This illustrates the paradoxical nature of OLG equilibria in the most orthodox setting. These paradoxical properties can hold equally for economies with many commodities, as pointed out in Section 4, "Understanding OLG Economies as Lack of Market Clearing at Infinity". Section 2, "Endogenous Cycles" discusses the possibility of equilibrium cycles in a one-commodity, stationary, OLG economy. In Section 3, "Money and the Sequential Economy", I describe OLG equilibria from a sequential markets point of view, and show that money can have positive value.

In the simple OLG economy of Section 1, "Indeterminacy and Suboptimality in a Simple OLG Model" there are two steady-state equilibria, and a continuum of non-stationary equilibria. Out of all of these, only one is Pareto efficient, and it has the property that the real rate of interest is always zero, just equal to the rate of population growth, independent of the impatience of the consumers or the distribution of endowments between youth and old age. This 'golden rule' equilibrium seems to violate Fisher's impatience theory of interest.

In Section 5, "Land, the Real Rate of Interest, and Pareto Efficiency" I add land to the one-commodity model of Section 1, "Indeterminacy and Suboptimality in a Simple OLG Model". It turns out that now there is a unique steady-state equilibrium that is Pareto efficient and that has a positive rate of interest, greater than the population growth rate, that increases if consumers become more impatient. Land restores Fisher's view of interest. In this setting it is also possible to analyse the effects of social security.

In Section 6, "Demography in OLG" I briefly introduce variations in demography. It is well known that birth rates in the United States oscillated every 20 years over the 20th century. Stock prices have curiously moved in parallel, rising rapidly from 1945 to 1965, falling from 1965 to 1985, and rising ever since. One might therefore expect stock prices to fall as the post-war baby boom generation retires. But some authors have claimed that these parallel fluctuations of stock prices must be coincidental. Otherwise, since demographic changes are known long in advance, rational investors would have anticipated the price fluctuations and changed them. In Section 6, "Demography in OLG" I allow the size of the generations to alternate and confirm that in OLG equilibrium land prices rise and fall with demography, even though the changes are perfectly anticipated.

In Section 7, "Impatience and Uniform Impatience" I show that not just land but also uniform impatience restores the properties of infinite horizon economies to those found in finite Arrow–Debreu economies.

Section 8, "Comparative Statics for OLG Economies" takes up the question of comparative statics. If there is a multiplicity of OLG equilibria, what sense can be made of comparative statics? Section 8, "Comparative Statics for OLG Economies" summarizes the work showing that, for perfectly anticipated changes, there is only one equilibrium in the multiplicity that is 'near' an original 'regular' equilibrium. For unanticipated changes, there may be a multidimensional multiplicity. But it is parameterizable. Hence, by always fixing the same variables, a unique prediction can be made for changes in the equilibrium in response to perturbations. In Section 9, "Keynesian Macroeconomics" we see how this could be used to understand some of the New Classical–Keynesian disputes about macroeconomic policy. Different theories hold different variables fixed in making predictions.

Section 10, "Neoclassical Equilibrium Versus Classical Equilibrium" considers a neoclassical-classical controversy. Recall the classical economists' conception of the economic process as a never-ending cycle of reproduction in which the state of physical commodities is always renewed, and in which the rate of interest is determined outside the system of supply and demand. Samuelson attempted to give a completely neoclassical explanation of the rate of interest in just such a setting. It now appears that the market forces of supply and demand are not sufficient to determine the rate of interest in the standard OLG model. In other infinite-horizon models they do.

Section 11, "Sunspots" summarizes some work on sunspots in the OLG model. Uncertainty in dynamic models seems likely to be very important in the future.

An explanation of the puzzles of OLG equilibria without land is given in Section 4, "Understanding OLG Economies as Lack of Market Clearing at Infinity": lack of market clearing 'at infinity'. By appealing to non-standard analysis, the mathematics of infinite and infinitesimal numbers, it can be shown that there is a 'finite-like' Arrow–Debreu economy whose 'classical equilibria', those price sequences which need not clear the markets in the last period, are isomorphic to the OLG equilibria. Lack of market clearing is also used to explain the suboptimality and the positive valuation of money.

Indeterminacy and Suboptimality in a Simple OLG Model

In this section we analyse the equilibrium set of a one-commodity per period, overlapping generations (OLG) economy, assuming that all agents meet simultaneously in all markets before time begins, just as in the Arrow–Debreu model. Prices are all quoted in present value terms; that is, p_t is the price an agent would pay when the markets meet (at time $-\infty$) in order to receive one unit of the good at time t. Although this definition of equilibrium is firmly in the Walrasian tradition of agent optimization and market clearing, we discover three surprises. There are robust examples of OLG economies that possess an uncountable multiplicity of equilibria, that are not in the core, or even Pareto optimal. This lack of optimality (in a slightly different model, as we shall see) was pointed out by Samuelson in his seminal (1958) paper. The indeterminacy of equilibrium in the one-commodity case is usually associated first with Gale (1973). In later sections we shall show that these puzzles are robust to an extension of the model to multiple commodities and agents per period, and to a non-stationary environment. We shall add still another puzzle in Section 3, "Money and the Sequential Economy", the positive valuation of money, which is also due to Samuelson.

A large part of this section is devoted to developing the notation and price normalization that we shall use throughout. In any Walrasian model the problem of price normalization (the 'numeraire problem') arises. Here the most convenient solution in the long run is not at first glance the most transparent.

Consider an overlapping generation (OLG) economy $E = E_{-\infty,\infty}$ in which discrete time periods t extend indefinitely into the past and into the future, $t \in \mathbf{Z}$. Corresponding to each time period there is a single, perishable consumption good x_t. Suppose furthermore that at each date t one agent is 'born' and lives for two periods, with utility

$$u^t(\ldots, x_t, x_{t+1}, \ldots) = a^t \log x_t + (1 - a^t) \log x_{t+1}$$

defined over all vectors

$$x = (\ldots, x_{-1}, x_0, x_1, \ldots) \in L = R_+^{\mathbf{Z}}.$$

Thus we identify the set of agents A with the time periods $\mathbf{Z}$. Let each agent $t \in A$ have endowment

$$e^t = \left(\ldots, e_t^t, e_{t+1}^t, \ldots\right) \in L$$

which is positive only during the two periods of his life. Note that

$$\sum_{t \in A} e_s^t = e_s^{s-1} + e_s^s \text{ for all } s \in \mathbf{Z}.$$

An equilibrium is defined as a (present value) price vector

$$p = (\ldots, p_{-1}, p_0, p_1, \ldots) \in L$$

and allocation

$$\bar{x} = \left[x^t = \left(\ldots, x_t^t, x_{t+1}^t, \ldots\right); t \in A\right]$$

satisfying $\bar{x}$ is feasible, that is,

$$\sum_{t \in A} x_s^t = \sum_{t \in A} e_s^t, \text{ for all } s \in \mathbf{Z} \tag{1}$$

and

$$\sum_{s \in \mathbf{Z}} p_s e_s^t < \infty \text{ for all } t \in A \tag{2}$$

O

and

$$x^t \in \arg\max_{x \in L} \left\{ u^t(x) | \sum_{s \in \mathbf{Z}} p_s x_s \leq \sum_{s \in \mathbf{Z}} p_s e_s^t \right\}. \tag{3}$$

The above definition of equilibrium is precisely in the Walrasian tradition, except that it allows for both an infinite number of traders and commodities. All prices are finite, and consumers treat them as parametric in calculating their budgets. The fact that the definition leads to robust examples with a continuum of Pareto-suboptimal equilibria calls for an explanation. We shall give two of them, one at the end of this section, and one in Section 4, "Understanding OLG Economies as Lack of Market Clearing at Infinity". Note that condition (2) becomes necessary only when we consider models in which agents have positive endowments in an infinite number of time periods.

As usual, the set of (present value) equilibrium price sequences displays a trivial dimension of multiplicity (indeterminacy), since, if p is an equilibrium, so is kp for all scalars $k > 0$. We can remove this ambiguity by choosing a price normalization $q_t = p_{t+1}/p_t$, for all $t \in \mathbf{Z}$. The sequence $q = (\ldots, q_{-1}, q_0, \ldots)$ and allocations $(x^t; t \in A)$ form an equilibrium if (1) above holds together with

$$x^t \in \arg\max_{x \in L} \left\{ u^t(x) | x_t + q_t x_{t+1} \leq e_t^t + q_t e_{t+1}^t \right\}. \tag{4}$$

Notice that we have taken advantage of the finite lifetimes of the agents to combine (2) and (3) into a single condition (4). We could have normalized prices by choosing a numeraire commodity, and setting its price equal to one, say $p_0 = 1$. The normalization we have chosen instead has three advantages as compared with this more obvious system. First, the q system is time invariant. It does not single out a special period in which a price must be 1; if we relabelled calendar time, then the corresponding relabelling of the q_t would preserve the equilibrium. In the numeraire normalization, after the calendar shift, prices would have to be renormalized to maintain $p_0 = 1$. Second, on account of the monotonicity of preferences, we know that, if the preferences and endowments are uniformly bounded

$$0 < \underline{a} \leq a^t \leq \overline{a} < 1, \; 0 < \underline{e} \leq e_t^t, \; e_{t+1}^t \leq \overline{e} \leq 1 \text{ for all } t \in A,$$

then we can specify uniform a priori bounds $\underline{k}$ and $\overline{k}$ such that any equilibrium price vector q must satisfy $\underline{k} \leq q_t \leq \overline{k}$ for all $t \in \mathbf{Z}$. Third, it is sometimes convenient to note that each generation's excess demand depends on its own price. We define

$$\left[Z_t^t(q_t), Z_{t+1}^t(q_t)\right] = \left(x_t^t - e_t^t, x_{t+1}^t - e_{t+1}^t\right)$$

for x^t satisfying (4), as the excess demand of generation t, when young and when old. We can accordingly rewrite equilibrium condition (1) as

$$Z_t^{t-1}(q_{t-1}) + Z_t^t(q_t) = 0 \text{ for all } t \in \mathbf{Z}. \tag{5}$$

Let us now investigate the equilibria of the above economy when preferences and endowments are perfectly stationary. To be concrete, let

$$a^t = a \text{ for all } t \in \mathrm{A},$$

and let

$$e_t^t = e, \text{ and } e_{t+1}^t = 1 - e, \text{ for all } t \in A,$$

where $e > a \geq 1/2$. Agents are born with a larger endowment when young than when old, but the aggregate endowment of the economy is constant at 1 in every time period. Furthermore, each agent regards consumption when young as at least as important as consumption when old ($a \geq 1/2$), but on account of the skewed endowment the marginal utility of consumption at the endowment allocation when young is lower than when old:

$$\frac{a}{e} < \frac{1-a}{1-e}.$$

If we choose

$$q_t = \overline{q} = \frac{(1-a)e}{(1-e)a} > 1$$

for all $t \in \mathbf{Z}$, then we see clearly that at these prices each agent will just consume his endowment; $q = (\ldots, \overline{q}, \overline{q}, \ldots)$ is an equilibrium price vector, with $x^t = e^t$ for all $t \in A$. Note that if we had used the price normalization $p_0 = 1$, the equilibrium prices would be described by

$$(\ldots, p_0, p_1, p_2, \ldots) = (\ldots, 1, \overline{q}, \overline{q}^2, \ldots)$$

where $p_t \to \infty$ as $t \to \infty$. With $a = 1/2$ and $e = 3/4$, we get $\overline{q} = 3$ and $p_t = 3^t$.

But there are other equilibria as well. Take $q = (\ldots, 1, 1, 1, \ldots)$, and

$$(x_t^t, x_{t+1}^t) = (a, 1-a) \text{ for all } t \in A.$$

This 'golden rule' Pareto equilibrium dominates the autarkic equilibrium previously calculated. With $a = 1/2$ and $e = 3/4$, we see that (1/2, 1/2) is much better for everyone than (3/4, 1/4). This raises the most important puzzle of overlapping generations economies: why is it that equilibria can fail to be Pareto optimal? We shall discuss this question at length in Section 4, "Understanding OLG Economies as Lack of Market Clearing at Infinity".

For now, let us observe one more curious fact. We can define the *core* of our economy in a manner exactly analogous to the finite commodity and consumer case. We say that a feasible allocation $x = (x^t; t \in A)$ is in the core of the economy E if there is no subset of traders $A' \subset A$, and an allocation $y = (y^t; t \in A')$ for A' such that

$$\sum_{t \in A'} y^t = \sum_{t \in A'} e^t,$$

and

$$u^t(y^t) > u^t(x^t) \text{ for all } t \in A'.$$

A simple argument can be given to show that the core of this economy is empty. For example, the golden rule equilibrium allocation is Pareto optimal, but not in the core. Since $a < e$, every agent is consuming less when young than his initial endowment. Thus for any $t_0 \in A$, the coalition $A' = \{t \in A | t \geq t_0\}$ consisting of all agents born at time t_0 or later can block the golden rule allocation.

Let us continue to investigate the set of equilibria of our simple, stationary economy. Gale (1973) showed that for any $\overline{q}_0$, with $1 < \overline{q}_0 < \overline{q}$, there is an equilibrium price sequence

$$q = (\ldots, q_{-1}, q_0, q_1, \ldots)$$

with $q_0 = \overline{q}_0$. In other words, there is a whole continuum of equilibria, containing a nontrivial interval of values. Incidentally, it can also be shown that for all such equilibria q, $q_t \to \overline{q}$ as $t \to \infty$, and $q_t \to 1$ as $t \to -\infty$. Moreover, these equilibria, together with the two steady state equilibria, constitute the entire equilibrium set.

This raises the second great puzzle of overlapping generations economies. There can be a non-degenerate continuum of equilibria, while in finite commodity and finite agent economies there is typically only a finite number. Thus if we considered the finite truncated economy $E_{-T,T}$ consisting of those agents born between $-T$ and T, and no others, then it can easily be seen that there is only a unique equilibrium $(q_{-T}, \ldots, q_T) = (\overline{q}, \ldots, \overline{q})$, no matter how large T is taken. On the other hand, in the overlapping generations economy, there is a continuum of equilibria. Moreover, the differences in these equilibria are not to be seen only at the tails. In the OLG economy, as $\overline{q}_0$ varies from 1 to $\overline{q}$, the consumption of the young agent at time zero varies from a to e, and his utility from $a \log e + (1-a) \log (1-e)$ (which for e near 1 is close to $-\infty$), all the way to $a \log a + (1-a) \log (1-a)$. By pushing the 'end of the world' further into the future, one does not approximate the world which does not end. We shall take up this theme again in Section 4,

O

"Understanding OLG Economies as Lack of Market Clearing at Infinity".

It is very important to understand that the multiplicity of equilibria is not due to the stationarity of the economy. If we imagined a non-stationary economy with each a^t near a and each (e_t^t, e_{t+1}^t) near $(e, 1-e)$, we would find the same multiplicity. One might hold the opinion that in a steady-state economy one should only pay attention to steady-state equilibria, that is, only to the autarkic and golden rule equilibria. In non-steady-state economies, there is no steady-state equilibrium to stand out among the continuum. One must face up to the multiplicity.

Let us reconsider how one might demonstrate the multiplicity of equilibria, even in a non-stationary economy. This will lead to a first economic explanation of indeterminacy similar to the one originally proposed by Gale. Suppose that in our non-stationary example we find one equilibrium $\hat{q} = (\ldots, \hat{q}_{-1}, \hat{q}_0, \hat{q}_1, \ldots)$ satisfying:

$$Z_t^{t-1}(\hat{q}_{t-1}) + Z_t^t(\hat{q}_t) = 0 \text{ for all } t \in \mathbf{Z}. \qquad (6)$$

Let us look for 'nearby' equilibria.

We shall say that generation t is expectations sensitive at $\hat{q}_t$ if both $[\partial Z_t^t(\hat{q}_t)/\partial q_t] \neq 0$ and $[\partial Z_{t+1}^t(\hat{q}_t)/\partial q_t] \neq 0$. If the first inequality holds, then the young's behaviour at time t can be influenced by what they expect to happen at time $t+1$. Similarly, if the second inequality holds, then the behaviour of the old agent at time $t+1$ depends on the price he faced when he was young, at time t. Recalling the logarithmic preferences of our example, it is easy to calculate that the derivatives of excess demands, for any $q_t > 0$, satisfy

$$\frac{\partial Z_t^t(q_t)}{\partial q_t} = a^t e_{t+1}^t \neq 0$$

and

$$\frac{\partial Z_{t+1}^t(q_t)}{\partial q_t} = \frac{-1(1-a^t)e_t^t}{q_t^2} \neq 0.$$

Hence, by applying the implicit function theorem to (1) we know that there is a nontrivial interval I_{t-1}^F containing $\hat{q}_{t-1}$ and a function F_t with domain I_{t-1}^F such that $F_t(\hat{q}_{t-1}) = \hat{q}_t$, and more generally,

$$Z_t^{t-1}(q_{t-1}) + Z_t^t[F_t(q_{t-1})] = 0 \text{ for all } q_{t-1} \in I_{t-1}^F.$$

Similarly there is a non-trivial interval I_t^B containing $\hat{q}_t$, and a function B_t with domain I_t^B such that $B_t(\hat{q}_t) = \hat{q}_{t-1}$, and more generally, $Z_t^{t-1}[B_t(q_t)] + Z_t^t(q_t) = 0$, for all $q_t \in I_t^B$. Of course, if $F_t(q_{t-1}) = q_t \in I_t^B$, then $B_t(q_t) = q_{t-1}$.

These forward and backward functions F_t and B_t, respectively, hold the key to one understanding of indeterminacy. Choose any relative price $q_0 \in I_0^F \cap I_0^B$ between periods 0 and 1. The behaviour of the generation born at 0 is determined, including its behaviour when old at period 1. If $q_0 \neq \hat{q}_0$, and generation 1 continues to expect relative prices $\hat{q}_1$ between 1 and 2, then the period 1 market will not clear. However, it will clear if relative prices q_1 adjust so that $q_1 = F_1(q_0)$. Of course, changing relative prices between period 1 and 2 from $\hat{q}_1$ to q_1 will upset market clearing at time 2, if generation 2 continues to expect $\hat{q}_2$. But if expectations change to $q_2 = F_2(q_1)$, then again the market at time 2 will clear. In general, once we have chosen $q_t \in I_t^F$, we can take $q_{t+1} = F_{t+1}(q_t)$ to clear the $(t+1)$ market. Similarly, we can work backwards. The change in q_0 will cause the period 0 market not to clear, unless the previous relative prices between period -1 and 0 were changed from $\hat{q}_{-1}$ to $q_{-1} = B_0(q_0)$. More generally, if we have already chosen $q_t \in I_t^B$, we can set $q_{t-1} = B_t(q_t)$ and still clear the period t market.

Thus we see that it is possible that an arbitrary choice of $q_0 \in I_0^F \cap I_0^B$ could lead to an equilibrium price sequence q. What happens at time 0 is undetermined because it depends on expectations concerning period 1, and also the past. But what can rationally be expected to happen at time 1 depends on what in turn is expected to happen at time 2, and so on.

There is one essential element missing in the above story. Even if $q_t \in I_t^F$, there is no guarantee that $q_{t+1} = F_{t+1}(q_t)$ is an element of I_{t+1}^F. Similarly, $q_t \in I_t^B$ does not necessarily imply that $q_{t-1} = B_t(q_t) \in I_{t-1}^B$. In our steady state example,

this can easily be remedied. Since all generations are alike,

$$F_t = F_1,\ B_t = B_0,\ I_t^F = I_0^F \text{ and } I_t^B = I_0^B \text{ for all } t \in \mathbf{Z}.$$

One can show that the interval $(1, \overline{q}) \subset I_0^F \cap I_0^B$, and that if $q_0 \in (1, \overline{q})$, then $F_1(q_0) \in (1, \overline{q})$, and $B_0(q_0) \in (1, \overline{q})$. This establishes the indeterminacy we claimed.

In the general case, when there are several commodities and agents per period, and when the economy is non-stationary, a more elaborate argument is needed. Indeed, one wonders, given one equilibrium $\hat{q}$ for such an economy, whether after a small perturbation to the agents there is any equilibrium at all of the perturbed economy near $\hat{q}$. We shall take this up in Section 8, "Comparative Statics for OLG Economies".

It is worth noting that we can define two more complete markets OLG economies with present value prices. In the economy $E_{0,\infty}$ only agents born at time $t \geq 1$ participate. The definition of OLG equilibrium is the same as before, except that now the set of agents is restricted to the participants, and market clearing is only required for $t \geq 1$. In the q-normalized form, equilibrium is defined by $q = (q_1, q_2, \ldots)$ such that

$$Z_1^1(q_1) = 0\ Z_t^{t-1}(q_{t-1}) + Z_t^t(q_t) = 0\ \forall t \geq 2.$$

It is immediately apparent (with one agent born per period and one good) that $E_{0,\infty}$ has a unique equilibrium, at which no agent trades and which is Pareto inefficient.

We could also define an economy $E_{0,\infty}^M$ in which only agents $t \geq 1$ participate, but where we require (in the normalized price version) that

$$Z_1^1(q_1) = -M\ \ Z_t^{t-1}(q_{t-1}) + Z_t^t(q_t) = 0\ \ \forall t \geq 2.$$

Equilibrium in $E_{0,\infty}^M$ is as if we gave an outside agent who had no endowment the purchasing power of M at time 1, and still managed to clear all markets $t \geq 1$. As long as $0 \leq M \leq Z_1^0(\overline{q})$, $E_{0,\infty}^M$ has an equilibrium. Take q_0 solving $M = Z_1^0(q_0)$, and $q_1 = F_1(q_0)$ and $q_t = F_t(q_{t-1})$ for $t \geq 2$.

We examine these two models more closely in Section 3, "Money and the Sequential Economy".

Endogenous Cycles

Let us consider another remarkable and suggestive property that one-commodity, stationary OLG economies can exhibit. We shall call the equilibrium $q = (\ldots, q_{-1}, q_0, q_1, \ldots)$ periodic of period n if $q_0, q_1, \ldots, q_{n-1}$ are all distinct, and if for all integers i and j, $q_i = q_{i+jn}$. The possibility that a perfectly stationary economy can exhibit cyclical ups and downs, even without any exogenous shocks or uncertainty, is reminiscent of 1930s–1950s business cycle theories. In fact, it is possible to construct a robust one-commodity per period economy which has equilibrium cycles of every order n. Let us see how.

As before, let each generation t consist of one agent, with endowment $e^t = (\ldots, 0, e, 1-e, 0, \ldots)$ positive only in period t and $t+1$, and utility $u^t(x) = u_1(x_t) + u_2(x_{t+1})$. Again, suppose that $\overline{q} = u_2'(1-e)/u_1'(e) > 1$. It is an immediate consequence of the separability of u^t, that for $q_t \leq \overline{q}$,

$$Z_t^t(q_t) \leq 0, Z_{t+1}^t(q_t) \geq 0, \frac{\partial Z_{t+1}^t(q_t)}{\partial q_t} < 0.$$

From monotonicity, we know that $Z_{t+1}^t(q_t) \to \infty$ as $q_t \to 0$. Hence it follows that for any $0 < q_0 < \overline{q}$, there is a unique $q_{-1} = B_0(q_0)$ with

$$Z_0^{-1}[B_0(q_0)] + Z_0^0(q_0) = 0.$$

From the fact that $Z_0^0(q_0) \geq -e$ for all q_0, it also follows that there is some $\underline{q} \leq 1$ such that if $q_0 \in \left[\underline{q}, \overline{q}\right]$ then $B_0(q_0) \in \left[\underline{q}, \overline{q}\right]$.

Now consider the following theorem due to the Russian mathematician Sarkovsky and to the mathematicians Li and Yorke (1975).

Sarkovsky–Li–Yorke Theorem *Let* $B: \left[\underline{q}, \overline{q}\right] \to \left[\underline{q}, \overline{q}\right]$ *be a continuous function from a non-trivial closed interval into itself. Suppose that*

O

there exist a three-cycle for B, that is, distinct points q_0, q_1, q_2, *in* $[\underline{q}, \overline{q}]$ *with* $q_1 = B(q_0)$, $q_2 = B(q_1)$, $q_0 = B(q_2)$. *Then there are cycles for B of every order n.*

Grandmont (1985), following related work of Benhabib and Day (1982) and Benhabib and Nishimura (1985), gave a robust example of a one-commodity, stationary economy (u_1, u_2, e) giving rise to a three-cycle for the function B_0. Of course a cycle for B_0 is also a cyclical equilibrium for the economy, hence there are robust examples of economies with cycles of all orders.

Theorem (Benhabib–Day 1982; Benhabib–Nishimura1985; Grandmont 1985). *There exist robust examples of stationary, one-commodity OLG economies with cyclical equilibria of every order n.*

This result is extremely suggestive of macroeconomic fluctuations arising for endogenous reasons, even in the absence of any fundamental fluctuations. Note first, however, that all of the cyclical equilibria, except the autarkic one-cycle $(\ldots, \overline{q}, \overline{q}, \overline{q}, \ldots)$, can be shown to be Pareto optimal (see Section 4, "Understanding OLG Economies as Lack of Market Clearing at Infinity"), while the theory of macroeconomic business cycles is concerned with the welfare losses from cyclical fluctuations. (On the other hand, the fact that cyclical behaviour is not incompatible with optimality is perhaps an important observation for macroeconomics.) More significantly, it must be pointed out that Sarkovsky's theorem is a bit of a mathematical curiosity, depending crucially on one dimension. And of course nonstationary economies, even with one commodity, will typically not have any periodic cycles. By contrast, the multiplicity and suboptimality of non-periodic equilibria that we saw in Section 1, "Indeterminacy and Suboptimality in a Simple OLG Model" are robust properties that are maintained in OLG economies with multiple commodities and heterogeneity across time. The main contribution of the endogenous business cycle literature is that it establishes the extremely important, suggestive principle that very simple dynamic models can have very complicated ('chaotic') dynamic equilibrium behaviour.

In the next section we turn to another phenomenon that can generally occur in overlapping generations economies, but never in finite horizon models.

Money and the Sequential Economy

Money very often has value in an overlapping generations model, but it never does in a finite horizon Arrow–Debreu model. The reason for its absence in the latter model is familiar: money would enable some agents to spend more on goods than they received from sales of their goods. But that would mean in the aggregate that spending on goods would exceed revenue from the sale of goods, contradicting market clearing in goods.

This argument can be given another form. Without uncertainty, Arrow–Debreu equilibrium can be reinterpreted as a sequential equilibrium with contemporaneous prices. But if the number of periods is finite, then in the last period the marginal utility of money to every consumer is zero, hence so is its price. In the second-to-last period nobody will pay to end up holding any money, because in the last period it will be worthless. By induction it will have no value even in the first period.

Evidently both these arguments fail in an infinite horizon setting. There is no last period, so the backward induction argument has no place to begin. And with an infinite number of consumers, aggregate spending and revenue might both be infinite, preventing us from comparing their sizes. On the other hand, there are infinite horizon models where money cannot have value. The difference between the OLG model and these other infinite horizon models will be discussed in Section 7, "Impatience and Uniform Impatience".

Strictly speaking, the overlapping generations model we have discussed so far has been modelled along the lines of Arrow–Debreu: each agent faced only one budget constraint and equilibrium was defined as if all markets met simultaneously at the beginning of time ($-\infty$). In such a model money has no function. However, we can define another model, similar to the

first considered by Samuelson, in which agents face a sequence of budget constraints and markets meet sequentially, and where money does have a store-of-value role. Surprisingly, this model turns out to have formally the same properties as the OLG model we have so far considered. To distinguish the two models we shall refer to this latter monetary model as the Samuelson model.

Suppose that we imagine a one-good per period economy in which the markets meet sequentially, according to their dates, and not simultaneously at the beginning of time. Suppose also that there are no assets or promises to trade. In such a setting it is easy to see that there could be no trade at all, since, as Samuelson put it, there is no double coincidence of wants. The old and the young at any date t both have the same kind of commodity, so they have no mutually advantageous deal to strike. But as Samuelson pointed out, introducing a durable good called money, which affects no agent's utility, might allow for much beneficial trade. The old at date t could sell their money to the young for commodities, who in turn could sell their money when old to the next period's young. In this manner new and more efficient equilibria might be created. The 'social contrivance of money' is thus connected to both the indeterminacy of equilibrium and the Pareto suboptimality of equilibrium, at least near autarkic equilibria. The puzzle, we have said, is how to explain the positive price of money when it has no marginal utility.

A closer examination of the equilibrium conditions of Samuelson's sequential monetary equilibrium reveals that, although it appears much more complicated, it reduces to the timeless OLG model we have defined above, but with one difference, namely, that the budget constraint of the generation endowed with money is increased by the value of money. The introduction of the asset money thus 'completes the markets' in the sense of Arrow (1953), by which we mean that the equilibrium of the sequential economy can be understood as if it were an economy in which money did not appear and all the markets cleared at the beginning of time (except, as we said, that the incomes of several agents are increased beyond the value of their endowments). The puzzle of how money can have positive value in the Samuelson model can thus be reinterpreted in the OLG model as follows. How is it possible that we can increase the purchasing power of one agent beyond the value of his endowment, without decreasing the purchasing power of any other agent below his, and yet continue to clear all the markets? Before giving a more formal treatment of the foregoing, let me re-emphasize an important point. It has often been said that the paradoxical properties of equilibrium in the sequential Samuelson consumption loan model can be explained on the basis of incomplete markets. Adding money to the model, however, completes the markets, in the precise sense of Arrow–Debreu, but the result is the OLG model in which the puzzles remain.

Let us now formally define the sequential one-commodity Samuelson model with money, $E^{M,S}_{0,\infty}$. Consider a truncated economy in which there is a new agent 'born' at each date $t \geq 0$, whose utility depends only on the two goods dated during his lifetime, and whose endowment is positive only in those same commodities. At each date $t \geq 1$ there will be two agents alive, a young one and an old one. Let us suppose that trade does not begin until period 1, so that the date 0 generation must consume its endowment when it is young. To this truncation of our earlier model we now add one extra commodity, which we call money. Money is a perfectly durable commodity that affects no agent's utility. Agents are endowed with money $\left(M^t_t, M^t_{t+1}\right)$, in addition to their commodity endowments.

A (contemporaneous) price system is defined as a sequence

$$(\pi; p) = (\pi_1, \pi_2, \ldots; p_1, p_2, \ldots)$$

of contemporaneous money prices π_t and contemporaneous commodity prices p_t for each $t \geq 1$. The budget set for any agent $t \geq 1$ is defined by

$$\left\{(m_t, m_{t+1}, x_t, x_{t+1}) \geq 0 \middle| \pi_t m_t + p_t x_t \leq \pi_t M^t_t + p_t e^t_t \text{ and } \pi_{t+1} m_{t+1} + p_{t+1} x_{t+1} \leq \pi_{t+1} M^t_{t+1} + p_{t+1} e^t_{t+1} + \pi_{t+1} m_t\right\}.$$

For agent 0 the budget constraint is

O

$$\{(m_0, m_1, x_0, x_1) \geq 0 | \, m_0 = M_0^0, x_0 = e_0^0, \text{ and}$$
$$\pi_1 m_1 + p_1 x_1 \leq \pi_1 M_1^0 + p_1 e_1^0 + \pi_1 m_0\}.$$

The budget constraints express the principle that in the Samuelson model agents cannot borrow at all, and cannot save, that is, purchase more when old than the value of their old endowment, except by holding over money m_t from when they were young. Let $m_t^t(\pi, p)$ and $m_{t+1}^t(\pi, p)$ be the utility maximizing choices of money holdings by generation t when young and when old. As before, the excess commodity demand is defined by $Z_t^t(\pi, p)$ and $Z_{t+1}^t(\pi, p)$.

To keep things simple, we suppose that agent 0 is endowed with $M_1^0 = M$ units of money when he is old, but all other endowments M_s^t are zero. Since money is perfectly durable, total money supply in every period is equal to M. Equilibrium is defined by a price sequence (π, p) such that for all $t \geq 1$,

$$m_t^{t-1}(\pi, p) + m_t^t(\pi, p) = M \text{ and } Z_t^{t-1}(\pi, p) + Z_t^t(\pi, p) = 0.$$

At first glance this seems a much more complicated system than before.

But elementary arguments show that in equilibrium either $\pi_t = 0$ for all t, and there is no intergenerational trade of commodities, or $\pi_t > 0$ for all t, or $\pi_t < 0$ for all t. In the case where $\pi_t > 0$, no generation will choose to be left with unspent cash when it dies, hence $m_{t+1}^t(\pi, p) = 0$ for all t, hence money market clearing is reduced to

$$m_t^t(\pi, p) = M \text{ for all } t \geq 1.$$

By homogeneity of the budget sets, if $\pi_t > 0$, we might as well assume $\pi_t = 1$ for all t. But then the prices p_t become the same as the present value prices from Section 1, "Indeterminacy and Suboptimality in a Simple OLG Model". From period by period Walras' Law, we deduce that, if the goods market clears at date t, so must the money market. So we never have to mention money market clearing.

Moreover, by taking $q_t = (\pi_t p_{t+1})/(\pi_{t+1} p_t)$ we can write the commodity excess demands for agent $t \geq 1$ just as in Section 1, "Indeterminacy and Suboptimality in a Simple OLG Model", by

$$\left[Z_t^t(q_t), Z_{t+1}^t(q_t)\right]$$

and they are the same as

$$\left[Z_t^t(\pi, p), Z_{t+1}^t(\pi, p)\right].$$

The only agent who behaves differently is agent 0, whose budget set must now be written

$$B^0(\mu, M) = \{(x_0, x_1) | \, x_0 = e_0^0, x_1 \leq e_1^0 + \mu M\},$$

where

$$\mu = \frac{\pi_1}{p_1}.$$

We can then write agent 0's excess demand for goods at time 1 as

$$Z_1^0(\mu, q, M) = Z_1^0(\mu M) = \mu M.$$

Thus any sequential Samuelson monetary equilibrium can be described by (μ, q), $\mu \geq 0$, satisfying

$$Z_1^0(\mu M) + Z_1^1(q_1) = 0,$$

and

$$Z_t^{t-1}(q_{t-1}) + Z_t^t(q_t) = 0 \text{ for all } t \geq 2.$$

But of course that is precisely the same as the definition of an OLG equilibrium for $E_{0,\infty}^{\mu M}$ given in Section 1, "Indeterminacy and Suboptimality in a Simple OLG Model".

Understanding OLG Economies as Lack of Market Clearing at Infinity

In this section we point out that the suboptimality of competitive equilibria, the indeterminacy of

non-stationary equilibria, the non-existence of the core, and the positive valuation of money can all occur robustly in possibly non-stationary OLG economies with multiple consumers and $L > 1$ commodities per period. We also note the important principle that the potential dimension of indeterminacy is related to L. In the two-way infinity model, it is $2L - 1$. In the one-way infinite model without money it is $L - 1$; in the one-way infinity model with money the potential dimension of indeterminacy is L.

None of these properties can occur (robustly) in a finite consumer, finite horizon, Arrow–Debreu model. In what follows we shall suggest that a proper understanding of these phenomena lies in the fact that the OLG model is isomorphic, in a precise sense, to a '*-finite' model in which not all the markets are required to clear.

One of the first explanations offered to account for the differences between the Arrow–Debreu model and the sequential Samuelson model with money centred on the finite lifetimes of the agents and the multiple budget constraints each faced. These impediments to intergenerational trade (for example, the fact that an agent who is 'old' at time t logically cannot trade with an agent who will not be 'born' until time $t + s$) were held responsible. But as we saw in the last section, without uncertainty the presence of a single asset like money is enough to connect all the markets. Formally, as we saw, the model is identical to what we called the OLG model in which we could imagine all trade taking place simultaneously at the beginning of time, with each agent facing a single budget constraint involving all the commodities. What prevents trade between the old and the unborn is not any defect in the market, but a lack of compatible desires and resources.

Another common explanation for the surprising properties of the OLG model centres on the 'paradoxes' of infinity, as suggested by Shell (1971). In finite models, one proves the generic local uniqueness of equilibrium by counting the number of unknown prices, less 1 for homogeneity, and the number of market clearing conditions, less 1 for Walras' Law, and notes that they are equal. In the OLG model there is an infinity of prices and markets, and who is to say that one infinity is greater than another? We already saw that the backward induction argument against money fails in an infinite horizon setting, where there is no last period. Surely it is right that infinity is at the heart of the problem. But this explanation does not go far enough. In the model considered by Bewley (1972) there is also an infinite number of time periods (but a finite number of consumers). In that model all equilibria are Pareto optimal, and money never has value, even though there is no last time period. The problem of infinity shows that there may be a difference between the Arrow–Debreu model and the OLG model. In itself, however, it does not predict the qualitative features (like the potential dimension of indeterminacy) that characterize OLG equilibria.

Consider now a general OLG model with many consumers and commodities per period. We index utilities $u^{t,h}$ by the time of birth t, and the household $h \in H$, a finite set. Household (t, h) owns initial resources $e_t^{t,h}$ when young, an L-dimensional vector, and resources $e_{t+1}^{t,h}$ when old, also an L-dimensional vector, and nothing else. As before utility $u^{t,h}$ depends only on commodities dated either at time t or $t + 1$. Given prices

$$q_t = (q_{ta}, q_{tb}) \in \Delta_{++}^{2L-1} = \left\{ q \in R_{++}^{2L} \mid \sum_{\ell=1}^{L} (q_\ell + q_{L+\ell}) = 2 \right\}$$

consisting of all the $2L$ prices at date t and $t + 1$, each household in generation t has enough information to calculate the relevant part of its budget set

$$B^{t,h}(q_t) = \left\{ (x_t, x_{t+1}) \in R_+^{2L} \mid q_{ta} \cdot x_t + q_{tb} \cdot x_{t+1} \leq q_{ta} \cdot e_t^{t,h} + q_{tb} \cdot e_{t+1}^{t,h} \right\}.$$

Hence we can write household excess demand $\left[Z_t^{t,h}(q_t), Z_{t+1}^{t,h}(q_t)\right]$ and the aggregate excess demand of generation t as $\left[Z_t^t(q_t), Z_{t+1}^t(q_t)\right]$, where

$$Z_{t+s}^t(q_t) = \sum_{h \in H} Z_{t+s}^{t,h}(q_t), s = 0, 1.$$

Of course we need to put restrictions on the q_t to ensure their compatibility, since q_{tb} and $q_{t+1,a}$ refer to the same period $t+1$ prices. But this is easily done by supposing that

$q_{tb} = \lambda_t q_{t+1a}$ for some $\lambda_t > 0, \forall t \in \mathbf{Z}$.

Present value OLG prices p can always be recovered from the normalized prices q via the recursion

$$p_1 = q_{1a}p_t = q_{ta}(\lambda_1\lambda_2 \ldots \lambda_{t-1}) \text{ for } t \geq 2 p_t$$
$$= q_{ta}\left(\lambda_0^{-1}\lambda_{-1}^{-1} \ldots \lambda_t^{-1}\right) \text{ for } t \leq 0.$$

We shall now define three variations of the OLG model and equilibrium, depending on when time starts, and whether or not there is money.

Suppose first that time goes from $-\infty$ to ∞. We can write the market clearing condition for equilibrium exactly as we did in the one-commodity, one-consumer case, as

$$Z_t^{t-1}(q_{t-1}) + Z_t^t(q_t) = 0, t \in \mathbb{Z}. \qquad \text{(A)}$$

Similarly we can define the one-way infinity economy $E_{0,\infty}$, in which time begins in period 0, but trade begins in time 1. We simply retain the same market clearing conditions for $t \geq 2$,

$$Z_t^{t-1}(q_{t-1}) + Z_t^t(q_t) = 0, t \geq 2 \qquad \text{(A}'\text{)}$$

$$\sum_{h \in H} \tilde{Z}_1^{0,h}(q_{1a}) + Z_1^1(q_1) = 0, \qquad (7)$$

it being understood that $Z_1^{0,h}$ has been modified to $\tilde{Z}_1^{0,h}(q_{1a})$ because every agent $(0, h)$ is forced to consume his own endowment at time 0, so that he maximizes over his budget set

$$B^{0,h}(q_{1a})$$
$$= \left\{(x_0, x_1) \in R_+^{2L} \,|\, x_0 = e_0^{0,h}, q_{1a} \cdot x_1 \leq q_{1a} \cdot e_1^{0,h}\right\}.$$

Finally, let us define equilibrium in a one-way infinity model with money, $E_{0,\infty}^M$, when agents $(0, h)$ are endowed with money M^h, in addition to their commodities, by (μ, q), $\mu \geq 0$, satisfying

$$\sum_{h \in H} \tilde{Z}_1^{0,h}(q_{1a}, \mu M^h) + Z_1^1(q_1) = 0, \qquad \text{(A}''\text{)}$$

and

$$Z_t^{t-1}(q_{t-1}) + Z_t^t(q_t) = 0, \text{ for } t \geq 2.$$

Again it is understood that the agents $(0, h)$ born in time 0 cannot trade in time 0, and they maximize over the budget set

$$B^{0,h}(q_{1a}, \mu M^h)$$
$$= \left\{(x_0, x_1) \in R_+^{2L} \,|\, x_0 = e_0^{0,h}, q_{1a} \cdot x_1 \leq q_{1a} \cdot e_1^{0,h} + \mu M^h\right\}.$$

These are the natural generalizations of the one-good economies defined in Section 1, "Indeterminacy and Suboptimality in a Simple OLG Model". (There is one small difference. With many agents born per period we can no longer conclude that if one agent holds a positive amount of money when young, then so must every other agent – no matter when he is born. We shall ignore this complication and allow some agents to hold negative money.)

We must now try to understand very generally why there may be many dimensions of OLG equilibria, why they might not be Pareto efficient, and how it is possible that some agents can spend beyond their budgets without upsetting market clearing.

Our explanation amounts to 'lack of market clearing at infinity'. We illustrate this for the case $E_{0,\infty}$.

Consider the truncated economy $E_{0,T}$ consisting of all the agents born between periods 0 and T. Market clearing in $E_{0,T}$ is defined to be identical to that in $E_{0,\infty}$ for $t = 1$ to $t = T$. But at $t = T+1$, we require $Z_{T+1}^T(q_t) = 0$ in $E_{0,T}$. This is a perfectly conventional Arrow–Debreu economy, and so necessarily has some competitive equilibria, all of which are Pareto efficient; generically its equilibrium set is a 0-dimensional manifold.

We have already seen in Section 1, "Indeterminacy and Suboptimality in a Simple OLG Model" what a great deal of difference there is between the economies $E_{0,T}$ (no matter how large T is) and $E_{0,\infty}$. The interesting point is that, by

appealing to non-standard analysis, which makes rigorous the mathematics of infinite and infinitesimal numbers, one can easily show that the economy $E_{0,\,T}$, for T an infinite number, inherits any property that holds for all finite $E_{0,\,T}$. Thus the paradoxical properties of the economy $E_{0,\,\infty}$ do not stem from infinity alone, since the infinite economy $E_{0,\,T}$ does not have them. We shall need to modify $E_{0,\,T}$ before it corresponds to $E_{0,\,\infty}$. Nevertheless, the economies $E_{0,\,T}$ do provide some information about $E_{0,\,\infty}$.

Theorem (Balasko–Cass–Shell 1980; Wilson 1981). *Under mild conditions, at least one equilibrium for $E_{0,\,\infty}$ always exists.*

To see why this is so, note that $E_{0,\,T}$ is well-defined for any finite T. From non-standard analysis we know that the sequence $E_{0,\,T}$ for $T \in \mathbf{N}$ has a unique extension to the infinite integers. Now fix T at an infinite integer. We know that $E_{0,\,T}$ has at least one equilibrium, since $E_{0,\,s}$ does for all finite s. But if T is infinite, $E_{0,\,T}$ includes all the finite markets $t = 1,2,\ldots$, so all those must clear at an equilibrium q^* of $E_{0,\,T}$. Taking the standard parts of the prices q_t^* for the finite t (and ignoring the infinite t) gives an equilibrium q for $E_{0,\infty}$.

To properly appreciate the force of this proof, we shall consider it again, when it might fail, in Section 7, "Impatience and Uniform Impatience", where we deal with infinite lived consumers.

In terms of the existence of equilibrium, $E_{0,\infty}$ (and similarly $E^M_{0,\infty}$ and $E_{-\infty,\infty}$) behaves no differently from an Arrow–Debreu economy. But the indeterminacy is a different story.

Definition A classical equilibrium for the economy $E_{0,\,T}$ is a price sequence $q^* = (q_1, \ldots, q_T)$ that clears the markets for $1 \leq t \leq T$, but at $t = T + 1$, market clearing $Z^T_{T+1}(q_T) = 0$ is replaced by

$$Z^T_{T+1}(q_T) \leq \sum_{h \in H} e^{T+1,h}_{T+1}.$$

Thus in a classical equilibrium there is lack of market clearing at the last period. The aggregate excess demand in that period, however, must be less than the endowment the young of period $T+1$ would have had, were they part of the economy. Economies in which market clearing is not required in every market are well understood in economic theory. Note that in a classical equilibrium the agents born at time T are not rationed at $T+1$; their full Walrasian (notional) demands are met, out of the dispossessed endowment of the young. But we do not worry about how this gift from the $T+1$ young is obtained. The significance of our classical equilibrium for the OLG models can be summarized in the following theorem from Geanakoplos and Brown (1982):

Theorem (Geanakoplos–Brown 1982) *Fix T at an infinite integer. The equilibria q for $E_{0,\infty}$ correspond exactly to the standard parts of classical equilibria q^* of $E_{0,\,T}$.*

The Walrasian equilibria of the economy $E_{0,\infty}$, which apparently is built on the usual foundations of agent optimization and market clearing, correspond to the 'classical equilibria' of another finite-like economy $E_{0,\,T}$ in which the markets at $T+1$ ('at infinity') need not clear. The existence of a classical equilibrium in $E_{0,\,T}$, and thus an equilibrium in $E_{0,\infty}$, is not a problem, because market clearing is a special case of possible non-market clearing, and $E_{0,\,T}$, being finite-like, always has market clearing equilibria.

Thus even though the number of prices and the number of markets in $E_{0,\infty}$ are both infinite, by looking at $E_{0,\,T}$ it is possible to say which is bigger, and by how much. There are exactly L more prices than there are markets to clear. From Walras' Law we know that if all the markets but one clear, that must clear as well. Hence having L markets that need not clear provides for $L-1$ potential dimensions of indeterminacy.

Corollary (Geanakoplos–Brown 1982). *For a generic economy $E_{0,\infty}$, there are at most $L-1$ dimensions of indeterminacy in the equilibrium set.*

Though the classical equilibria of $E_{0,\,T}$ generically have $L-1$ dimensions of indeterminacy, it is by no means true that there must be $L-1$ dimensions of *visible* indeterminacy. If we consider any classical equilibrium q^* for a generic economy $E_{0,\,T}$, then we will be able to arbitrarily perturb some set of $L-1$ prices near their q^*

values, and then choose the rest of the prices to clear all the markets up through time T. But which $L-1$ prices these are depends on which square submatrix N (of derivatives of excess demands with respect to prices) is invertible. For example, call the economy $E_{0,\infty}$ intertemporally separable if each generation t consists of a single agent whose utility for consumption at date t is separable from his utility for consumption at date $t+1$. Then the $L-1$ free parameters must all be chosen at date $T+1$ (as part of $q_{T,b}$), that is, way off at infinity.

Corollary (Geanakoplos–Polemarchakis 1984). *Intertemporally separable economies $E_{0,\infty}$ generically have locally unique equilibria (in the product topology).*

For example, a natural generalization of the example in Section 1, "Indeterminacy and Suboptimality in a Simple OLG Model" would be to generations consisting of a single Cobb–Douglas consumer of $L > 1$ goods when young and when old. The corollary shows that this economy has no indeterminacy of equilibrium. Since Cobb–Douglas economies seem so central, one might guess that multi-good OLG economies $E_{0,\infty}$ do not generate indeterminacy. But that is incorrect. Separability with one agent drastically reduces the effect expectations about future prices can have on the present, because changes in future consumption do not change marginal utilities today. In the separable case, changing all L prices tomorrow only affects today through the one dimension of income.

Even when the $L-1$ degrees of freedom may be chosen at time $t=1$, there still may be no visible indeterminacy, if the matrix N has an inverse (in the non-standard sense) with infinite norm. But when the free $L-1$ parameters may be chosen at $t=1$ and also the matrix N has an inverse with finite norm, then all nearby economies must also display $L-1$ dimensions of indeterminacys.

Theorem (Kehoe–Levine 1984; Geanakoplos–Brown 1982). *In the $E_{0,\infty}$ OLG model there are robust examples of economies with $L-1$ dimensions of indeterminacy. In the monetary economy, $E^{M}_{0,\infty}$ there are robust examples of economies with L dimensions of indeterminacy.*

Let us now turn our attention to the question of Pareto optimality.

Definition An allocation $\bar{x}=\left(x^{t,h};0\le t\le T\right)$ is classically feasible for the economy $E_{0,T}$ if $\sum_{(t,h)\in A}x_s^{t,h}\le\sum_{(t,h)\in A}e_s^{t,h}$, for $0\le s\le T+1$. The classically feasible allocation $\bar{x}$ for $E_{0,T}$ is a classic Pareto optimum if there is no other classically feasible allocation $\bar{y}$ for $E_{0,T}$ with $u^t(y^{t,h})>u^t(x^{t,h})$ for all $(t,h)\in A$ with $0\le t\le T$, with at least one inequality $(0,h)$ representing a non-infinitesimal difference.

Theorem (Geanakoplos–Brown 1982). *The Pareto-optimal allocations $\bar{x}$ for the OLG economy $E_{0,\infty}$ are precisely the standard parts of classical Pareto-optimal allocations $\bar{x}^*$ for $E_{0,T}$, if T is fixed at an infinite integer.*

The upshot of this theorem is that the effective social endowment includes the commodities e_{T+1}^{T+1} of the generation born at time $s=T+1$, even though they are not part of the economy $E_{0,T}$. Since the socially available resources exceed the aggregate of private endowments, it is no longer a surprise that a Walrasian equilibrium, in which the value of aggregate spending every period must equal the value of aggregate private endowments, is not Pareto optimal.

On the other hand, this does not mean that all equilibria are Pareto suboptimal. If the (present value) equilibrium prices $p_t\to 0$, as $t\to\infty$ (or, more generally, if p_{T+1} is infinitesimal), then the value of the extra social endowment is infinitesimal, and there are no possible non-infinitesimal improvements. To see this, let $(p,\bar{x})$ be an equilibrium in present value prices for the OLG economy $E_{0,\infty}$. Consider the concave–convex programming problem of maximizing the utility of agent $(0,\bar{h})$, holding all other utilities of agents (t,h) with $0\le t\le T$ at the levels $u^{t,h}(x^t)$ they get with $\bar{x}$, over all possible allocations in $E_{0,T}$ that do not use more resources, even at time $T+1$, than $\bar{x}$. Clearly $\bar{x}$ itself is a solution to this problem. But now let us imagine raising the constraints at time $T+1$ from

$$\sum_{h\in H} x_{T+1}^{T,h} \text{ to } \sum_{h\in H}\left(e_{T+1}^{T,h}+e_{T+1}^{T+1,h}\right).$$

What is the rate of change of the utility $u^{0,\overline{h}}$? From standard concave programming theorems, for the first infinitesimal additions to period $T+1$ resources, the rate of change of $u^{0,\overline{h}}$ is on the order of p_{T+1}, assuming p_1 is normalized to equal the marginal utility of consumption for agent $(0,\overline{h})$ at date 1. Additional resources bring decreasing benefits. This shows that if p_{T+1} is infinitesimal, then there are no possible non-infinitesimal improvements with a finite amount of extra resources.

An important example of $p_t \to 0$ occurs when the prices are summable, as they are when they decline geometrically to zero. Thus in a stationary equilibrium with a positive real interest rate, equilibrium must be Pareto efficient. Another proof of efficiency in the case of geometric present value prices is to observe that then the present value of the aggregate endowment must be finite, so the standard proof of Pareto efficiency in a finite horizon model goes through.

If p_t increases geometrically to infinity, then it is evident that equilibrium cannot be Pareto efficient. Thus, in a stationary equilibrium with a negative real interest rate, equilibrium must be Pareto inefficient.

When $p_t \not\to 0$ but also does not increase exponentially to infinity, the calculation becomes much more delicate. An infinitesimal increase ε in resources at time $T+1$ can be used to increase utility of $(0,\overline{h})$ on the order of $p_{T+1}\,\varepsilon$, which is still infinitesimal if p_{T+1} is non-infinitesimal but finite. As the increases ε get larger, this rate of change could drop quickly, as higher derivatives come into play (assuming that agents have strictly concave utilities), leaving infinitesimal (and thus invisible) increases in utility even with a finite increase in resources. Second derivatives, and their uniformity, come into play. But this subtle case has been brilliantly dealt with:

Theorem (Cass 1972; Benveniste–Gale 1975; Balasko–Shell 1980; Okuno–Zilcha 1980). *If agents have uniformly strictly concave utilities, and if the aggregate endowment is uniformly bounded away from 0 and ∞, then the equilibrium $(p, \overline{x})$ with present value prices p for an OLG economy $E_{0,\infty}$ is Pareto optimal if and only if*

$$\sum_{t=0}^{\infty} 1/||p_t|| = \infty.$$

Note that in this theorem it is the present value prices that play the crucial role. It follows immediately from this theorem that the golden rule equilibrium $q = (..., 1, 1, 1, ...)$. for the simple one good, stationary economy of Section 1, "Indeterminacy and Suboptimality in a Simple OLG Model" is Pareto optimal, since the corresponding present value price sequence is also (..., 1, 1, 1, ...). In fact, a moment's reflection shows that any periodic, non-autarkic equilibrium must also be periodic in the present value prices p. Hence, as we have said before, but without a proof, the cyclical equilibria of Section 2, "Endogenous Cycles" are all Pareto optimal.

Having explained the indeterminacy and Pareto suboptimality of equilibria for $E_{0,\infty}$ in terms of lack of market clearing at infinity, let us re-examine the monetary equilibria of OLG economies $E_{0,\infty}^{M}$, where $M = (M^h; h \in H)$ is the stock of money holdings by the agents $(0, h)$ at time 0.

The next theorem shows that any monetary equilibrium allocation of $E_{0,\infty}^{M}$, corresponds to the standard part of a non-monetary economy $E_{0,T}(z)$ obtained from $E_{0,t}$ by augmenting the endowments of the first generation $(0, h)_{h\in H}$ by a vector of goods z at time $T+1$.

Definition Let $z \in R^L$ be a vector of commodities for time $T+1$. Suppose that $-\sum_{h\in H} e_{T+1}^{T,h} \le \sum_{h\in H} M^h z \le \sum_{h\in H} e_{T+1}^{T+1,h}$. Let the augmented non-monetary economy $E_{0,T}(z, M)$ be identical to the non-monetary economy $E_{0,T}$, except that the endowment of each agent $(0, h)$ is augmented by $M^h \cdot z$ units of commodities at time $T+1$.

Theorem (Geanakoplos–Brown 1982). *Fix an infinite integer T. The equilibria q of the monetary economy $E_{0,\infty}^{M}$ are precisely obtained by taking standard parts of full market clearing equilibria q* of all the augmented non-monetary economies $E_{0,T}(z, M)$.*

O

The above theorem explains how it is possible to give agents $(0, h)$ extra purchasing power without disturbing market clearing in the economy $E^M_{0,\infty}$. The answer is that the purchasing power comes from owning extra commodities at date $T + 1$, and equilibrium in $E^M_{0,\infty}$. does not require market clearing in date $T + 1$ commodities.

The above theorem gives another view of why there are potentially L dimensions of monetary equilibria: the augmenting endowment vector z can be chosen from a set of dimension L. It also explains how money can have positive value: it corresponds to the holding of extra physical commodities. The theorem also explains how the 'social contrivance of money' can lead to Pareto-improving equilibria, even in OLG economies where there is already perfect financial intermediation. The holding of money can effectively bring more commodities into the aggregate private endowment. The manifestation of the 'real money balances' is the physical commodity bundle z at date $T + 1$. Money plays more than just an intermediation role.

Before concluding this section let us consider a simple generalization. Suppose that agents live for three periods. What plays the analogous role to $E_{0,T}$? The answer is that prices need to be specified through time $T + 2$, but markets are only required to clear through time T. There are therefore $2L - 1$ potential dimensions of indeterminacy, even in the one-sided economy. In general, we must specify the price vector up until some time s, and then require market clearing only in those commodities whose excess demands are fully determined by those prices.

This reasoning has an important generalization to production. Suppose that capital invested at time t can combine with labour at time $t + 1$ to produce output at time $t + 1$, and suppose that all agents live two periods. Is there any difference between the case where labour is inelastically supplied, and the case where leisure enters the utility? In both cases the number of commodities is the same, but in the latter case the potential dimension of indeterminacy is one higher, since the supply of labour at any time might depend on further prices.

Land, the Real Rate of Interest, and Pareto Efficiency

Allais and Samuelson argued that the infinity of both time periods and agents radically changed the nature of equilibrium. Samuelson suggested that equilibrium might not be Pareto efficient, and that the real rate of interest might be negative, even if the economy did not shrink over time. In our one-good example from Section 1, "Indeterminacy and Suboptimality in a Simple OLG Model", the autarkic equilibrium has a negative real interest rate since each $q_t < 1$, and the real interest rate is $1/q_t - 1$.

They also thought that a second, new kind of equilibrium would emerge in which the real rate of interest is divorced from any of the considerations like impatience that Irving Fisher had stressed. They thought that in this new kind of equilibrium the real rate of interest would turn out to be equal to the rate of population growth, irrespective of the impatience of the consumers or the distribution of their endowments. Indeed, in the example from Section 1, "Indeterminacy and Suboptimality in a Simple OLG Model", the 'golden rule' equilibrium had real interest rate $1/q_t - 1 = 0$ in every period, irrespective of the utilities or the endowments, but equal to the population growth rate.

Furthermore, as we saw in Section 3, "Money and the Sequential Economy", Samuelson argued that it might not be necessary for an asset to be valued according to the present value of its dividends, contradicting yet another one of Fisher's central concepts. Samuelson suggested that a piece of green paper might be worth a lot, even though it pays no dividends, because the holder might think he could sell it to somebody later, who would buy it on the expectation that he could sell it to somebody else later, ad infinitum. Later authors called this a rational bubble.

It turns out that these views are incorrect if one includes in the model infinitely lived assets like land, that do pay dividends in every period.

Imagine an OLG economy as before with

$$u^t(x_t, x_{t+1}) = \frac{1}{2}\log x_t + \frac{1}{2}\log x_{t+1}\left(e_t^t, e_{t+1}^t\right) = (3, 1).$$

But let us also suppose there is one acre of *land* in the economy that produces a dividend $D_t = 1$ apple every period for ever. Suppose the economy begins in period 1, with an old agent who owns the land and has an endowment of one apple, and a newly born agent as above. We suppose that buying the land at time t gives ownership of all dividends from time $t + 1$ up to and including the dividends in the period in which the asset is sold. The apple dividend from the land at time 1 is owned by the old agent at time 1 (who presumably acquired the land at time 0 and hence has the claim on the apple).

At every period t we need to find the contemporaneous price q_t of the commodity and the price Π_t of the land.

Every agent in the economy must decide how much to consume when young, and what assets to hold when young, and how much to consume when old. The decision in old age is trivial, since the agent cannot do better than selling every asset he has and using the proceeds to buy consumption goods.

Thus for every $t \geq 1$ we can describe the decision problem of generation t by

$$\max_{y,z,\theta} u^t(y, z) = \frac{1}{2}\log y + \frac{1}{2}\log z \text{ such that } q_t y + \Pi_t \theta = q_t e_t^t = q_t 3\, q_{t+1} z = q_{t+1} e_{t+1}^t + \theta D_{t+1} + \Pi_{t+1}\theta = q_{t+1} 1 + \theta 1 + \Pi_{t+1}\theta.$$

For the original old generation, he optimizes simply by setting

$$x_1^0 = e_1^0 + D_1 + \Pi_1 = 1 + 1 = 2 + \Pi_1.$$

Denote the optimal choice of agents $t \geq 1$ by $(x^t_t, x^t_{t+1}, \theta^t)$. Market clearing requires for each $t \geq 2$ that consumption of the old plus consumption of the young is equal to total output of goods, and also that demand equals the supply of land

$$x_t^{t-1} + x_t^t = e_t^{t-1} + e_t^t + D_t = 1 + 3 + 1 = 5 \quad \theta^t = 1.$$

In period $t = 1$ we must have

$$x_1^0 + x_1^1 = e_1^0 + e_1^1 + D_1 = 1 + 3 + 1 = 5 \quad \theta^1 = 1.$$

Sequential equilibrium is thus a vector $\left(x_1^0, \left(q_t, \Pi_t, \left(x_t^t, x_{t+1}^t, \theta^t\right)\right)_{t=1}^{\infty}\right)$ satisfying the above conditions on agent maximization and market clearing.

Fisher's recipe for computing equilibrium with assets is to put the asset dividends into the endowments of their owners, and then find the usual general equilibrium with present value prices ignoring the assets. In this example that means giving agent 0 an endowment $e^0 = (2,1, \ldots)$ of two apples in period 1 and one apple every period thereafter, and ignoring the land. Equilibrium with present value prices is then described exactly as in Section 1, "Indeterminacy and Suboptimality in a Simple OLG Model".

To solve for the present value prices $(p_1, p_2, \ldots)$ we can guess that since the economy is stationary, there will be stationary equilibrium $(p_1, p_2, \ldots) = (1, p, p^2, \ldots)$. For each $t \geq 2$, we must solve

$$\frac{1}{2}\frac{[3 + p1]}{p} + \frac{1}{2}[3 + p1] = 1 + 3 + 1,$$

which gives a quadratic equation

$$p^2 - 6p + 3 = 0$$

which is solved by

$$p = \frac{6 \pm (36 - 12)^{.5}}{2} = .55, r = 1/p - 1 = 81.7\%.$$

The other root is greater than one, and could not be right, because it would give a real interest rate less than zero, which would make the present value of land infinite. Hence consumption when young and old is

O

$$(y, z) = (1.775, 3.225).$$

Clearly these values clear the consumption market for all $t \geq 2$. We know by Walras's Law that, if all markets but one clears, then the last will as well, so we don't really have to check the period 1 market. But we will check it anyway. The present value of agent 0's endowment is

$$2 + p1 + p^2 1 + \ldots = 2 + p/(1-p) = 3.225$$

and so indeed the period $t = 1$ market clears.

We can now translate this general equilibrium back into a sequential equilibrium. Taking $q_t = 1$ for every period and the real interest rate solving $p = 1/(1 + r)$, the present value of land is

$$\begin{aligned} PV\ Land &= p1 + p^2 1 + \ldots = p/(1-p) \\ &= \frac{1}{1+r}1 + \frac{1}{(1+r)^2}1 + \ldots = 1.225. \end{aligned}$$

In every period the old will consume their endowment of 1 plus the dividend of 1 plus the value of the land they will sell, which gives exactly 3.225. The sequential equilibrium is $\left(x_1^0, (q_t, \Pi_t, (x_t^t, x_{t+1}^t, \theta^t))_{t=1}^{\infty}\right) = (3.225, (1, 1.225, (1.775, 3.225, 1))_{t=1}^{\infty})$.

Despite what Allais and Samuelson said, the rate of interest at the unique steady state is positive, higher than the growth rate of population. Moreover, as noted in Geanakoplos (2005), the real interest rate does respond to shocks in exactly the way Fisher argued. Consider the same model as before, but make all the consumers more impatient

$$U(y, z) = \frac{2}{3}\log y + \frac{1}{3}\log z.$$

Then our master equation would become

$$\frac{1}{3}\frac{[3 + p1]}{p} + \frac{2}{3}[3 + p1] = 1 + 3 + 1$$

giving

$$p = .419,\ r = 139\%,\ PV\ Land = .721.$$

As Fisher would have predicted, the real rate of interest does indeed increase, and the price of land decreases.

Pareto Efficiency and Bubbles

Observe that in our example the dividends of land represent 20 per cent of all endowments every period. Since the price of land must be finite, that means in any equilibrium the present value of all endowments must be finite. We know that implies equilibrium must be Pareto efficient.

Furthermore, if the value of aggregate endowments is finite, then money cannot have value and there can be no bubbles, because the old argument is correct that markets cannot clear if some agents are spending more than the value of their commodity endowments and nobody is spending less. Land makes the OLG economy look much more like an Arrow–Debreu economy.

Social Security

The overlapping generations model is the workhorse model for examining social security. There is not space here to describe these studies. Observe simply that a pay-as-you-go system amounts to a simple transfer of endowments from each young person to each old person. We can immediately calculate the effects of such a transfer on our steady state interest rate and land value by recomputing the equilibrium for the OLG economy in which endowments are adjusted to (2,2) for every generation $t \geq 1$, and assuming the old generation 0 has an endowment of 2 apples at time 1 plus the land, which pays 1 apple every period. We get

$$p = .38,\ r = 161\%,\ PV\ Land = .62.$$

This also confirms Fisher's contention that decreasing early endowments and increasing later endowments should raise the rate of interest and lower land values.

Notice that the pay-go system gives each agent the same number of apples when he is old that he gave up when he is young, which is a below market return on his original contribution. Social security lowers the utility of every agent except the first generation. Samuelson had argued that

social security could make every agent better off. But his conclusion is false in the model with land.

It is often said that if only every generation had more children, social security would give better returns, since the young would be able to share the burden of helping the old. The trouble with that reasoning is that it ignores the fact that higher population and output growth would mean higher real interest rates, which tend to make the social security rate of return as bad as before relative to market interest rates. There is no space to discuss this here.

Demography in OLG

In America since the early 20th century, the generations have alternated in size between big and small. Everybody knows about the baby boom and echo baby boom, but the same pattern happened before. Recently many authors have suggested that the retiring of the baby boom generation will force stock prices to fall. This has been criticized on the grounds that demography is easy to predict. If agents knew that stock prices would fall when the baby boomers retired, they would fall now. These two opposing views can be analysed in the OLG model by allowing generation sizes to fluctuate.

Suppose the small generation is exactly as before, but now we alternate that small generation with a large generation that is identical in every respect, except that it is twice as big

$$u^b(y,z) = \frac{1}{2}\log y + \frac{1}{2}\log z\left(e_y^b, e_z^b\right) = (6,2).$$

As before, suppose that land produces 1 unit of output each period. Begin at time 1 with a small generation of young, and suppose the old owns the land.

We investigate whether the price of land and the real interest rate alternate between periods.

Let r_b be the interest rate that prevails when the big generation b is young, and r_a prevail when the small generation a is young. Equilibrium can be reduced to two equations. The first describes market clearing for goods in odd periods when the small generation is young and the big generation is old, and the second equation describes market clearing in even periods, when the big generation is young and the small generation is old. As before, we let $p_a = 1/(1 + r_a)$ and $p_b = 1/(1 + r_b)$. Then

$$\begin{aligned}&\frac{1}{2}\frac{[6+p_b 2]}{p_b} + \frac{1}{2}[3+p_a 1] \\ &\quad = 2+3+1 \qquad \frac{1}{2}\frac{[3+p_a 1]}{p_a} + \frac{1}{2}[6+p_b 2] \\ &\quad = 1+6+1.\end{aligned}$$

These can be simultaneously solved to get

$$\begin{aligned}p_a &= .418, r_a = 139\%, PV_{Landa} = 1.29 \\ p_b &= .912, r_b = 9.6\%, PV_{Landb} = 2.09.\end{aligned}$$

It is evident that the price of land is higher in the periods when b is young, since the interest rate is lower. Even though it is perfectly anticipated that when the big generation gets old, the price of land will fall, the price does not fall earlier because the interest rate is so low. (This point has been made by Geanakoplos et al. 2004.)

Impatience and Uniform Impatience

We have already suggested that it is useful in understanding the OLG model to consider variations, for example in which consumers live for ever. By doing so we shall also gain an important perspective on what view of consumers is needed to restore the usual properties of neoclassical equilibrium to an infinite horizon setting, a subject to which we return in Section 8, "Comparative Statics for OLG Economies".

Let us now allow for consumers $t \in A$ who have endowments e^t that may be positive in all time periods, and also for arbitrary utilities u^t defined on uniformly bounded vectors $x \in L = \mathbf{R}_+^{\mathbf{N}}$. For ease of notation we assume one good per period. A minimal assumption we need about utilities u^t is continuity on finite segments,

that is, fixing x_s for all $s > n$, $u^t(x)$ should be continuous in $(x_1,\ldots, x_n)$. We also need continuity on L, in some topology, but we will not go into these details. We also assume $\sum_{t \in A} e^t$ is uniformly bounded. In short, we suppose consumers may live for ever.

We shall find that in order to have Walrasian equilibria the consumers must be impatient. Suppose we try to form the truncated economy $E_{0,T}$ as before, say for T finite. Since utility potentially depends on every commodity, we could not define excess demands in $E_{0,T}$ unless we knew all the prices. To make it into a finite economy, let us call $E'_{0,T}$ the version of $E_{0,T}$ in which every agent is obliged to consume his initial endowment during periods $t \leq 0$ and $t > T$. Clearly $E'_{0,T}$ has an equilibrium. For this to give information about the original economy $E_{0,\infty}$, we need that consumers do not care very much about what happens to them after T, as T gets very far away. This requires a notion of impatience.

For any vector x, let ${}_n\hat{x}$ be the vector which is zero for $t > n$, and equal to x up until n. Thus ${}_n\hat{x}$ is the initial n-segment of x. To say that agent $t \in A$ is impatient means that for any two uniformly bounded consumption streams x and y, if $u^t(x) > u^t(y)$, then for all big enough n, $u^t({}_n\hat{x}) > u^t(y)$. Let us suppose that all consumers are impatient. If these segments can be taken uniformly across agents, then we say the economy is uniformly impatient. Any finite economy with impatient consumers is uniformly impatient.

Note that the OLG agents are all impatient, since none of them cares about consumption after he dies, but the economy is not uniformly impatient.

Even with an economy consisting of all impatient consumers, the truncation argument, applied at an infinite $E'_{0,T}$, does not guarantee the existence of an equilibrium. For, once we take standard parts, ignoring the infinitely dated commodities, it may turn out that the income from the sale of an agent's endowed commodities at infinite t, which he used to finance his purchase of commodities at finite t, is lost to the agent. It must also be guaranteed that the equilibria of $E'_{0,T}$ give infinitesimal total value to the infinitely dated commodities. Wilson (1981) has given an example of an economy, composed entirely of impatient agents, that does not have an equilibrium precisely for this reason.

On the other hand, if there are only finitely many agents, even if they are infinitely lived, then we have:

Theorem (Bewley 1972). *Let the economy E be composed of finitely many, impatient consumers. Then there exists an equilibrium, and all equilibria are Pareto optimal.*

The Pareto efficiency of equilibria in these Bewley economies can be derived from the standard proof of efficiency: since there is a finite number of agents, the value of the aggregate endowments is a finite sum of finite numbers, and therefore finite itself.

In the special case with separable, commonly discounted utilities of the form $u^h(x) = \sum_{t=0}^{\infty} \delta^t v^h(x_t)$, with $\delta < 1$, we have:

Theorem (Kehoe–Levine 1985)**.** *In finite agent, separable commonly discounted utility economies, there is generically* a *finite number of equilibria.*

This theorem has been extended by Shannon (1999) and Shannon and Zame (2002).

Returning to the case of an infinite number of consumers, Pareto efficiency of equilibria, if they exist, can be guaranteed as long as a finite number of the agents collectively hold a non-negligible fraction of total endowment. But that also would guarantee the existence of equilibrium, since in the economy $E'_{0,T}$ we would then get the summability of the prices, meaning the endowments at infinity would have zero value, as Wilson (1981) pointed out.

It is extremely interesting to investigate the change in behaviour of an economy that evolves from individually impatient to uniformly impatient. Wilson (1981) considered an example with one infinitely lived agent, and infinitely many, overlapping, finite-lived agents, and showed that equilibria must exist, and all must be Pareto efficient. By the foregoing remarks, no matter what the proportion of sizes of the two kinds of consumers, equilibria must exist and be Pareto efficient. Muller and Woodford (1988) showed in a

particular case that, when the single agent's proportion of the aggregate endowment is low enough, there is a continuum of equilibria, but if it is high enough there is no local indeterminacy.

Comparative Statics for OLG Economies

A celebrated theorem of Debreu assets that almost any Arrow–Debreu economy is regular, in the sense that it has a finite number of equilibria, each of which is locally unique. Small changes to the underlying structure of the economy (tastes, endowments, and so on) produce small, unique changes in each of the equilibria.

We have already seen that there are robust OLG economies with a continuum of equilibria. If attention is focused on one of them, how can one predict to which of the continuum of new equilibria the economy will move if there is a small change in the underlying structure of the economy, perhaps caused by deliberate government intervention? In what sense is any one of the new equilibria near the original one? In short, is comparative statics possible?

It is helpful at this point to recall that the OLG model is, in spirit, meant to represent a dynamic economy. Trade may occur as if all the markets cleared simultaneously at the beginning of time, but the economy is equally well described as if trade took place sequentially, under perfect foresight or rational expectations. Indeed, this is surely what Samuelson envisaged when he introduced money as an asset into his model. Accordingly, when a change occurs in the underlying structure of the economy, we can interpret it as if it came announced at the beginning of time, or as if it appeared at the date on which it actually affects the economy.

We distinguish two kinds of changes to the underlying structure of an economy $\overline{E}_{-\infty,\infty}$ starting from an equilibrium $\overline{q}$. Perfectly anticipated changes, after which we would look for a new equilibrium that cleared all the markets from the beginning of time, represent one polar case, directly analogous to the comparative statics experiments of the Arrow–Debreu economy. At the other extreme we consider perfectly unanticipated changes, say at date $t = 1$. Beginning at the original economy and equilibrium $\overline{q} = (\ldots, \overline{q}_{-1}, \overline{q}_0, \overline{q}_1, \ldots)$, we would look, after the change from $\overline{E}_{-\infty,\infty}$ to $E_{-\infty,\infty}$ at time $t = 1$ (say to the endowment or preferences of the generation born at time 1), for a price sequence $q = (\ldots, q_{-1}, q_0, q_1, \ldots)$ in which $q_t = \overline{q}_t$ for $t \leq 0$, and $Z_t^{t-1}(q_{t-1}) + Z_t^t(q_t) = 0$ for $t \geq 2$. But at date $t = 1$ we would require q_1 to satisfy $Z_1^0(q_{1a} | \overline{q}_0) + Z_1^1(q_1) = 0$, where $Z_1^0(q_{1a} | \overline{q}_0)$ represents the excess demand of the old at time 1, given that when they were young they purchased commodities on the strength of the conviction that they could surely anticipate prices $\overline{q}_{0b}$ when they got old, only to discover prices q_{1a} instead.

To study these two kinds of comparative statics, we must describe what we mean by saying that two price sequences are nearby. Our definition is based on the view that a change at time $t = 1$ ought to have a progressively smaller impact the further away in time from $t = 1$ we move. We say that q is near $\overline{q}$ if the difference $|q_t - \overline{q}_t|$ declines geometrically to zero, both as $t \to \infty$ and as $t \to -\infty$.

We have already noted in Section 1, "Indeterminacy and Suboptimality in a Simple OLG Model" that the multiplicity of OLG equilibria is due to the fact that at any time t the aggregate behaviour of the young generation is influenced by their expectations of future prices, which (under the rational expectations hypothesis) depends on the next generation's expectations, and so on. Accordingly we restrict our attention to generations whose aggregate behaviour Z^t satisfies the expectations sensitivity hypothesis:

$$\operatorname{rank} \frac{\partial Z_t^t(p_t, p_{t+1})}{\partial p_{t+1}} = \operatorname{rank} \frac{\partial Z_{t+1}^t(p_t, p_{t+1})}{\partial p_t} = L.$$

For economies composed of such generations we can apply the implicit function theorem, exactly as in Section 1, "Indeterminacy and Suboptimality in a Simple OLG Model", around any equilibrium q to deduce the existence of the forward and backward functions F_t and B_t. We write their derivatives at $\overline{q}$ as D_t and D_t^{-1} respectively.

O

For finite Arrow–Debreu economies, Debreu gave a definition of regular equilibrium based on the derivative of excess demand at the equilibrium. He showed that comparative statics is sensible at a regular equilibrium, and then he showed that a 'generic' economy has regular equilibria. We follow the same program.

We say that the equilibrium $\overline{q}$ for the expectations sensitive OLG economy $\overline{E}$ is Lyapunov regular if the long-run geometric mean of the products $D_t^* D_t D_{t-1}^* D_{t-1} \cdots D_1^* D_1$ and $D_{-t}^{-1} \ldots D_{-1}^{-1} * D_{-1}^{-1}$ converge and if to these products we can associate $2L - 1$ eigenvalues, called Lyapunov exponents. The equilibrium is also non-degenerate if in addition none of these Lyapunov exponents is equal to 1.

Theorem (Geanakoplos–Brown 1985). *Let* $\overline{E} = \overline{E}_{-\infty,\infty}$ *be an expectations-sensitive economy with a regular non-degenerate equilibrium* $\overline{q}$. *Then for all sufficiently small perfectly anticipated perturbations E of* $\overline{E}$ (*including* $\overline{E}$ *itself*) *E has a unique equilibrium q near* $\overline{q}$.

Thus the comparative statics of perfectly anticipated changes in the structure of $\overline{E}$, around a regular, non-degenerate equilibrium, is directly analogous to the Arrow–Debreu model. The explanation for the theorem is that a perfectly anticipated change at time 0 gives rise to price changes that have a forward stable manifold (on which prices converge exponentially back to where they started) and a backward stable manifold, and that there is only one price at time 0 that is on both the forward and the backward stable manifolds. Note, incidentally, that one implication of the above theorem is that neutral policy changes, like jawboning or changing animal spirits, that is, those for which $\overline{q}$ itself remains an equilibrium, cannot have any effect if they are perfectly anticipated and move the economy to nearby equilibria.

Theorem (Geanakoplos–Brown 1985). *Let* $\overline{E}$ *be an expectations-sensitive economy with a regular equilibrium* $\overline{q}$. *Then, for all sufficiently small perfectly unanticipated perturbations E of* $\overline{E}$ (*including* $\overline{E}$ *itself*), *the set of unanticipated equilibria q of E near* $\overline{q}$ *is either empty, or a manifold of dimension r*, $0 \leq r \leq L$ ($L - 1$ *if there is no money in the economy*), *where r is independent of the perturbation.*

The above theorem allows for the possibility that an unanticipated change may force the economy onto a path that diverges from the original equilibrium; the disturbance could be propagated and magnified through time. And if there are nearby equilibria, then there may be many of them. (Indeed, that is basically what was shown in Section 4, "Understanding OLG Economies as Lack of Market Clearing at Infinity".) In particular, an unanticipated neutral policy change could be compatible with a continuum of different equilibrium continuations. The content of the theorem is that, if there is a multiplicity of equilibrium continuations, it is parameterizable. In other words, the same r variables can be held fixed, and for any sufficiently small perturbation, there is exactly one nearby equilibrium which also leaves these r variables fixed. We shall discuss the significance of this in the next section.

This last theorem was proved first, in the special case of steady-state economies, by Kehoe–Levine, in the same excellent paper to which we have referred already several times. The theorem quoted here, together with the previous theorem on the comparative statics of perfectly anticipated policy changes, refers to economies in which the generations may be heterogeneous across time.

Let us suppose that A is a compact collection of generational characteristics, all of which obey the expectations-sensitive hypothesis. Let us suppose that each generation's characteristics are drawn at random from A, according to some Borel probability measure. If the choices are made independently across time, then the product measure describes the selection of economies. Almost any such collection will have a complex demographic structure, changing over time. The equilibrium set is then endogenously determined, and will be correspondingly complicated. It can be shown, however, that

Theorem (Geanakoplos–Brown 1985). *If the economy E is randomly selected, as described*

above, then with probability 1, *E has at least one Lyapunov regular equilibrium.*

Note that the regularity theory for infinite economies stops short of Arrow–Debreu regularity. In the finite economies, with probability one all the equilibria are regular.

Keynesian Macroeconomics

Keynesian macroeconomics is based in part on the fundamental idea that changes in expectations, or animal spirits, can affect equilibrium economic activity, including the level of output and employment. It asserts, moreover, that publicly announced government policy also has predictable and significant consequences for economic activity, and that therefore the government should intervene actively in the marketplace if investor optimism is not sufficient to maintain full employment.

The Keynesian view of the indeterminacy of equilibrium and the efficacy of public policy has met a long and steady resistance, culminating, in the sharpest attack of all, from the so-called new classicals, who have argued that the time-honoured microeconomic methodological premises of agent optimization and market clearing, considered together with rational expectations, are logically inconsistent with animal spirits and the non-neutrality of public monetary and bond-financed fiscal policy.

The foundation of the new classical paradigm is the Walrasian equilibrium model of Arrow–Debreu, in which it is typically possible to prove that all equilibria are Pareto optimal and that the equilibrium set is finite; at least locally, the hypothesis of market clearing fixes the expectations of rational investors. In that model, however, economic activity has a definite beginning and end. Our point of view is that for some purposes economic activity is better described as a process without end. In a world without a definite end, there is the possibility that what happens today is underdetermined, because it depends on what people expect to happen tomorrow, which in turn depends on what people tomorrow expect to happen the day after tomorrow, and so on.

Consider the simple one-good per period overlapping generations economy with money $E_0^{M,S}$, which we discussed in Section 3, "Money and the Sequential Economy". Generation 0 is endowed with money when old, and equilibrium can be described with the contemporaneous commodity prices $\overline{p} = (\overline{p}_1, \overline{p}_2, \ldots)$ where we take the price of money to be fixed at 1. (In this case, as we saw in Section 3, "Money and the Sequential Economy", contemporaneous prices are also present value prices.) It is helpful to reinterpret the model as a simple production economy. Imagine that the endowment e_t^t in the first period of life is actually labour, which can be transformed into output, y_t according to the production function, $y_t = e_t$ We would then think of any purchases of goods by the old generation as demand for real output to be produced by the young. The young in turn now derive utility from leisure in their youth and consumption in their old age. Equilibrium in which consumption of the old is higher can be interpreted as an equilibrium with less leisure and higher output.

The indeterminacy of rational expectations equilibrium has the direct interpretation that optimistic expectations *by themselves* can cause the economy's *output* to expand or contract. In short the economy has an inherent volatility. The Keynesian story of animal spirits causing economic growth or decline can be told without invoking irrationality or non-market clearing.

In fact, the indeterminacy of equilibrium expectations is especially striking when seen as a response to public (but unanticipated) policy changes. Suppose the economy is in a long-term rational expectations equilibrium $\overline{p}$, when at time 1 the government undertakes some expenditures, financed, say, by printing money. How should rational agents respond? The environment has been changed, and there is no reason for them to anticipate that $(\overline{p}_2, \overline{p}_3, \ldots)$ will still occur in the future. Indeed, in models with more than one commodity (such as we will shortly consider) there may be no equilibrium $(p_1, p_2, p_3, \ldots)$ in the new environment with $p_2 = \overline{p}_2, p_3 = \overline{p}_3$, and so on. There is an ambiguity in what can be rationally anticipated.

We argue that it is possible to explain the differences between Keynesian and monetarist policy predictions by the assumptions each makes about expectational responses to policy, and not by the one's supposed adherence to optimization, market clearing, and rational expectations, and the other's supposed denial of all three.

Consider now the government policy of printing a small amount of money, ΔM, to be spent on its own consumption of real output – or equivalently to be given to generation $t = 0$ (when old) to spend on its consumption. Imagine first that agents are convinced that this policy is not inflationary, that is, that $\overline{p}_1$ will remain the equilibrium price level during the initial period of the new equilibrium. This will give generation $t = 0$ consumption level $(M + \Delta M)/\overline{p}_1$. As long as ΔM is sufficiently small and the initial equilibrium was one of the Pareto-suboptimal equilibria described in Section 1, "Indeterminacy and Suboptimality in a Simple OLG Model", there is indeed a new equilibrium price path p beginning with $p_1 = \overline{p}_1$ Output at time 1 rises by $\Delta M/\overline{p}_1$, and in fact this policy is Pareto improving. On the other hand, imagine instead that agents are convinced that the path of real interest rates $p_t/p_{t+1} - 1$ will remain unchanged. In this economy, price expectations are a function of p_1. Recalling the initial period market-clearing equation, it is clear that p_1 and all future prices rise proportionally to the growth in the money stock. The result is that output is unchanged and the old at $t = 1$ must pay for the government's consumption. If the government's consumption gives no agent utility, the policy is Pareto worsening.

This model is only a crude approximation of the differences between Keynesian and monetarist assumptions about expectations and policy. It is quite possible to argue, for example, that holding $p_2/p_1 = \overline{p}_2/\overline{p}_1$ (the future inflation rate) fixed is the natural *Keynesian* assumption to make. This ambiguity is unavoidable when there is only one asset into which the young can place their savings. We are thereby prevented from distinguishing between the inflation rate and the interest rate. Our model must be enriched before we can perform satisfactory policy analysis. Nevertheless, the model conveys the general principle that expected price paths are not locally unique. There is consequently no natural assumption to make about how expectations are affected by policy. A sensible analysis is therefore impossible without externally given hypotheses about expectations. These can be Keynesian, monetarist, or perhaps some combination of the two.

Geanakoplos and Polemarchakis (1985) build just such a richer model of macroeconomic equilibrium by adding commodities, including a capital good, and a neoclassical production function. With elastically supplied labour, there are two dimensions of indeterminacy. It is therefore possible to fix both the nominal wage, and the firm's expectations ('animal spirits'), and still solve for equilibrium as a function of policy perturbations to the economy. These institutional rigidities are more convincingly Keynesian, and they lead to Keynesian policy predictions. Moreover, taking advantage of the simplicity of the two-period lived agents, the analysis can be conducted entirely through the standard Keynesian (Hicksian) IS–LM diagram.

Keynesians themselves often postulate that the labour market does not clear. For Keynesians, lack of labour market clearing has at least a threefold significance, which it is perhaps important to sort out. First, since labour is usually taken to be inelastically supplied, it makes it possible to conceive of (Keynesian) equilibria with different levels of output and employment. Second, it makes the system of demand and supply underdetermined, so that endogenous variables like animal spirits (that is, expectations) which are normally fixed by the equilibrium conditions can be volatile. Third, it creates unemployment that is involuntary. By replacing lack of labour market clearing at time 1 with elastic labour supply and lack of market clearing 'at infinity' one can drop what seems to many an ad hoc postulate, yet retain at least the first two desiderata of Keynesian analysis.

Neoclassical Equilibrium Versus Classical Equilibrium

The Arrow–Debreu model of general equilibrium, based on agent optimization, rational expectations,

and market clearing, is universally regarded as the central paradigm of the neoclassical approach to economic theory. In the Arrow–Debreu model, consumers and producers, acting on the basis of individual self-interest, combine, through the aggregate market forces of demand and supply, to determine (at least locally) the equilibrium distribution of income, relative prices, and the rate of growth of capital stocks (when there are durable goods). The resulting allocations are always Pareto optimal.

Classical economists at one time or another have rejected all of the methodological principles of the Arrow–Debreu model. They replace individual interest with class interest, ignore (marginal) utility, especially for waiting, doubt the existence of marginal product, and question whether the labour market clears. But by far the most important difference between the two schools of thought is the classical emphasis on the long-run reproduction of the means of production, in a never-ending cycle.

Thus the celebrated classical economist Sraffa writes in Appendix D to his book:

> It is of course in Quesnay's *Tableau Economique* that is found the original picture of the system of production and consumption as a circular process, and it stands in striking contrast to the view presented by modern theory, of a one-way avenue that leads from 'Factors of Production' to 'Consumption Goods.'

The title of his book, *Production of Commodities by Means of Commodities,* itself suggests a world that has no definite beginning, and what is circular can have no end.

In the Arrow–Debreu model time has a definite end. As we have seen, that has strong implications. With universal agreement about when the world will end, there can be no reproduction of the capital stock. In equilibrium it will be run down to zero. Money, for example, can never have positive value. Rational expectations will fix, at each moment, and for each kind of investment, the expected rate of profit.

In the classical system, by contrast, the market does not determine the distribution of income. Sraffa (1960, p. 33) writes:

> The rate of profits, as a ratio, has a significance which is independent of any prices, and can well be 'given' before the prices are fixed. It is accordingly susceptible of being determined from outside the system of production, in particular by the money rates of interest. In the following sections the rate of profits will therefore be treated as the independent variable.

Other classical writers concentrate instead on the real wage as determined outside the market forces of supply and demand, for example by the level of subsistence or the struggle between capital and labour. Indeterminacy of equilibrium seems at least as central to classical economists as it is to Keynesians.

Like Keynesians, classicals often achieve indeterminacy in their formal models by allowing certain markets not to clear in the Walrasian sense. (Again like Keynesians, the labour market is usually among them.) Thus we have called the equilibrium in Section 4, "Understanding OLG Economies as Lack of Market Clearing at Infinity" in which some of the markets were allowed not to clear a 'classical equilibrium'.

What the OLG model shows is that, by incorporating the classical view of the world without definite beginning or end, it is possible to maintain all the neoclassical methodological premises and yet still leave room for the indeterminacy which is the hallmark of both classical and Keynesian economics. In particular this can be achieved while maintaining labour market clearing. The explanation for this surprising conclusion is that the OLG model is isomorphic to a finite-like model in which indeed not all the markets need to clear. But far from being the labour markets, under pressure to move towards equilibrium from the unemployed clamouring for jobs, these markets are off 'at infinity', under no pressure towards equilibrating.

We have speculated that, once one has agreed to the postulate that the resources of the economy are potentially as great at any future date as they are today, then uniform impatience of consumers is the decisive factor, according to Walrasian principles, which may influence whether the market forces of supply and demand determine a locally unique, Pareto-optimal equilibrium, or leave room for extramarket forces to choose among the continuum of inefficient equilibria. In these terms, the

Arrow–Debreu model supposes a short-run impatient economy, and OLG a long-run patient economy.

Sunspots

So far we have not allowed uncertainty into the OLG model. As a result we found no difference in interpreting trade sequentially, with each agent facing two budget constraints, or 'as if' the markets all cleared simultaneously at the beginning of time, with each agent facing one budget constraint. Once uncertainty is introduced these inpts become radically different. In either case, however, there is a vast increase in the number of commodities, and hence in the potential for indeterminacy.

If we do not permit agents to make trades conditional on moves of nature that occur before they are born, then agents will have different access to asset markets. Even in finite horizon economies, differing access to asset markets has been shown by Cass and Shell (1983) to lead to 'sunspot effects'.

A 'sunspot' is a visible move of nature which has no real effect on consumers, on account of preferences, or endowments, or through production. In the Arrow–Debreu model it also could have no effect on equilibrium trade; this is no longer true when access to asset markets differs.

The sunspot effect is intensified when combined with the indeterminacy that can already arise in an OLG economy. Consider the simple one good, steady state OLG economy of Section 2, "Endogenous Cycles". Suppose that there is an equilibrium two cycle in present value prices $p = (\ldots, p_{-1}, p_0, p_1, \ldots)$ with $p_{2t} = p^S$ and $p_{2t+1} = p^R$, for all $t \in \mathcal{T}$. Now suppose that the sun is known to shine on even periods, and hide behind rain on odd periods. The above equilibrium is perfectly correlated with the sun, even though no agent's preferences or endowments are. As usual, the same prices for $t \geq 0$ support an equilibrium, given the right amount of money, in $E^{M,S}_{0,\infty}$.

More generally, suppose that the probability of rain or shine, given the previous period's weather, is given by the Markov matrix $\pi = (\pi_{SS}, \pi_{SR}, \pi_{RS}, \pi_{RR})$. A steady state equilibrium for $E^{M,S}_{0,\infty}$, given π, is an assignment of a money price for the commodity, depending only on that period's weather, such that, if all agents maximize their expected utility with respect to π, then in each period the commodity market and money market clears. Azariadis (1981) essentially showed that, if there is a two-cycle of the certainty economy, then there is a continuum of steady state sunspot equilibria.

The sunspot equilibria, unlike the cyclical equilibria of Section 2, "Endogenous Cycles", are Pareto suboptimal whenever the matrix π is non-degenerate.

The combination of the dynamic effects of the infinite horizon OLG model with the burgeoning theory of incomplete markets under real uncertainty, is already on the agenda for the next generation's research.

See Also

► Arrow–Debreu Model of General Equilibrium

Bibliography

Allais, M. 1947. *Economie et intérêt*. Paris: Imprimerie Nationale.

Arrow, K.J. 1953. The role of securities in the optimal allocation of risk-bearing. Repr. in K.J. Arrow, *Essays in the theory of risk bearing*. Chicago: Markham, 1971.

Azariadis, C. 1981. Self-fulfilling prophecies. *Journal of Economic Theory* 25: 380–396.

Balasko, Y., and K. Shell. 1980. The overlapping-generations model, I: The case of pure exchange without money. *Journal of Economic Theory* 23: 281–306.

Balasko, Y., D. Cass, and K. Shell. 1980. Existence of competitive equilibrium in a general overlapping-generations model. *Journal of Economic Theory* 23: 307–322.

Benhabib, J., and R.H. Day. 1982. A characterization of erratic dynamics in the overlapping generations model. *Journal of Economic Dynamics and Control* 4: 37–44.

Benhabib, J., and N. Nishimura. 1985. Competitive equilibrium cycles. *Journal of Economic Theory* 35: 284–306.

Benveniste, L., and D. Gale. 1975. An extension of Cass' characterization of infinite efficient production programs. *Journal of Economic Theory* 10: 229–238.
Bewley, T. 1972. Existence of equilibria in economies with infinitely many commodities. *Journal of Economic Theory* 4: 514–540.
Cass, D. 1972. On capital overaccumulation in the aggregative, nonclassical model of economic growth. *Journal of Economic Theory* 4: 200–223.
Cass, D., and K. Shell. 1983. Do sunspots matter? *Journal of Political Economy* 91: 183–227.
Cass, D., and M.E. Yaari. 1966. A re-examination of the pure consumption loan model. *Journal of Political Economy* 74: 200–223.
Debreu, G. 1970. Economies with a finite set of equilibria. *Econometrica* 38: 387–392.
Diamond, P.A. 1965. National debt in a neo-classical growth model. *American Economic Review* 55: 1126–1150.
Fisher, I. 1907. *The rate of interest*. New York: Macmillan.
Fisher, I. 1930. *The theory of interest*. New York: Macmillan.
Gale, D. 1973. Pure exchange equilibrium of dynamic economic models. *Journal of Economic Theory* 6: 12–36.
Geanakoplos, J. 1978. Sraffa's *Production of Commodities by Means of Commodities*: Indeterminacy and suboptimality in neoclassical economics. Working Paper, RIAS. Also Chapter 3 in 'Four essays on the model of Arrow–Debreu'. Ph. D. thesis, Harvard University, 1980.
Geanakoplos, J. 2005. The ideal inflation indexed bond and Irving Fisher's theory of interest with overlapping generations. In *Celebrating Irving Fisher: The legacy of a great economist,* ed. R. Dimand and J. Geanakoplos. Oxford: Blackwell. *Journal of Economics and Sociology* 64: 257–305.
Geanakoplos, J., and D. Brown. 1982. *Understanding overlapping generations economies as lack of market clearing at infinity.* Mimeo, Yale University, revised 1985, 1986.
Geanakoplos, J., and D. Brown. 1985. Comparative statics and local indeterminacy in OLG economies: An application of the multiplicative ergodic theorem. Discussion Paper No. 773, Cowles Foundation.
Geanakoplos, J., and H.M. Polemarchakis. 1984. Intertemporally separable overlapping generations economies. *Journal of Economic Theory* 34: 207–215.
Geanakoplos, J., and H.M. Polemarchakis. 1985. Walrasian indeterminacy and Keynesian macroeconomics. *Review of Economic Studies* 53: 755–799.
Geanakoplos, J., M. Magill, and M. Quinzi. 2004. Demography and the long-run predictability of the stock market. *Brookings Papers on Economic Activity* 2004 (1): 241–325.
Grandmont, J.M. 1985. Endogenous, competitive business cycles. *Econometrica* 53: 995–1046.
Kehoe, T.J., and D.K. Levine. 1984. Regularity in overlapping generations exchange economies. *Journal of Mathematical Economics* 13: 69–93.
Kehoe, T.J., and D.K. Levine. 1985. Comparative statics and perfect foresight in infinite horizon models. *Econometrica* 53: 433–453.
Kehoe, T.J., D.K. Levine, and P.M. Romer. 1990. Determinacy of equilibrium in dynamic models with finitely many consumers. *Journal of Economic Theory* 50: 1–21.
Keynes, J.M. 1936. *The general theory of employment, interest and money.* London: Macmillan.
Li, T.Y., and J.A. Yorke. 1975. Period three implies chaos. *American Mathematical Monthly* 8: 985–992.
Lucas, R.E. 1972. Expectations and the neutrality of money. *Journal of Economic Theory* 4: 102–121.
Muller, W.J., and M. Woodford. 1988. Determinacy of equilibrium in stationary economies with both finite and infinite lived consumers. *Journal of Economic Theory* 46: 255–290.
Okuno, M., and I. Zilcha. 1980. On the efficiency of a competitive equilibrium in infinite horizon monetary economies. *Review of Economic Studies* 47: 797–807.
Samuelson, P.A. 1958. An exact consumption loan model of interest, with or without the social contrivance of money. *Journal of Political Economy* 66: 467–482.
Shannon, C. 1999. Determinacy of competitive equilibria in economies with many commodities. *Economic Theory* 14: 29–87.
Shannon, C., and B. Zame. 2002. Quadratic concavity and determinacy of equilibrium. *Econometrica* 70: 631–662.
Shell, K. 1971. Notes on the economics of infinity. *Journal of Political Economy* 79: 1002–1011.
Sraffa, P. 1960. *Production of commodities by means of commodities*. Cambridge: Cambridge University Press.
Wilson, C. 1981. Equilibrium in dynamic models with an infinity of agents. *Journal of Economic Theory* 24: 95–111.

Overproduction

B. A. Corry

The term overproduction, or general gluts as it was earlier called, and its allied, if not synonymous term, underconsumption, are rarely, if at all, to be found today in standard, orthodox textbooks of economics. Yet for past generations of writers on economics they were familiar concepts about which must ink was spilt and heated debate ranged. Discussions of overproduction were closely bound up with discussions of underconsumption because the latter has tended

to be one of the main explanations of overproduction. Hence our analysis of the concept of overproduction and a brief survey of the development of theories to explain it, has to include some discussion of underconsumption.

The absence of these terms from orthodox economics has occurred for several reasons; first they are somewhat ambiguous as regards their meaning, and as we shall illustrate in the course of our discussion, it was never clear what particular writers had in mind when they wrote of general gluts, overproduction or underconsumption. A second reason for the decline in use of the terms has been the growth of formal model construction in macroeconomics where emphasis is placed on the econometric estimation of the model rather than on an intuitive vision of the way capitalism does, or does not, operate.

The common presumption underlying all theories of overproduction, and by implication theories of underconsumption, is that capitalist market economies have a built-in flaw, which may or may not be correctable. This flaw is that the level of spending may not be sufficient to sustain the volume of output produced at full employment, so that there will be output not sold, or at least not saleable at prices that make it profitable to produce, hence overproduction will occur and eventually production will contract and general unemployment occur. Underconsumptionists argue that this overproduction arises primarily because consumption is too low and this has been overwhelmingly the main line of approach in the overproduction school.

We have already referred to ambiguities which has surrounded the use of the terms. We must now state these ambiguities and look on them in some detail. Then we shall give a brief account of the histories of the doctrines.

The first ambiguity is whether overproduction is inevitable, i.e., must always occur, or whether is occurs at certain times, e.g., at the peak and downturn phase of the trade cycle. It has been argued by some writers, for example, that the incomes earned in the production of output even if spent in their entirity on output, that is on both consumption and capital goods, could never enable producers to recoup their costs including a normal rate of return on capital. This was the basis of the famous A + B theorem associated with Major Douglas and the Social Credit Movement. The A + B theorem failed to recognize the accounting identity that the sum of income earned in production, must in a closed system, equal the sum of expenditures – so that total output *could* be bought at prices that would recoup production costs.

Another form of this inevitability argument was the view that the act of saving or rather any attempt to save was a leakage from the circular flow of income and would result in unsold output. The reasoning here was that saving is 'unspent' income, hence income earned in the production process is not returned to the income stream so production costs – including a necessary profit element – *cannot* be recouped. Put another way this view of underconsumption or overproduction assumes that consumption is the only form of expenditure. This view – what we may term crude underconsumption – is quite common in the earlier popular literature. A good example from the early 18th century is Bernard de Mandeville's famous *Fable of the Bees*, where 'Knaves turn honest', decrease their consumption and increase: thriftiness' and cause the economy to slump. In the early 19th century similar views are to be found in a group of English writers such as W. Spence who found the basis for their underconsumption views within a Physiocratic framework.

The Classical economists, in the main, refused to consider overproduction as a possibility and their main weapon of defence was Say's Law (after J.B. Say, although the basic texts of it are to be found in Adam Smith). There are several varieties of Say's Law ranging from a mere tautology to Keynesian-like versions, but the basic idea is that decisions to save are automatically translated into decisions to spend on capital accumulation, so that the circular flow of income is maintained and overproduction cannot occur. The automatic regulator of this mechanism was assumed to be the rate of interest, equating on the one hand to desire of entrepreneurs to use resources for investment purposes with, on the other hand, the desire of households to save income.

Another ambiguity concerns the problem as to whether those writers who spoke of overproduction had in mind what we may term *actual* overproduction or *potential* overproduction. Those who believed in actual overproduction envisaged situations where output would actually be produced that could not be sold at prices that covered necessary costs of production. Now this is clearly a most unlikely situation; decision takers are assumed to act rationally given the information available to them; they presumably can estimate fairly accurately the market opportunities open to them so that they simply will not produce unprofitable output. The sequence of events is rather as follows; if there is an unexpected fall in demand producers may be unsure whether the fall is temporary or permanent hence, in the short run they may continue output at its current level and allow stocks to accumulate. Once the decline in demand is seen as a longer term phenomenon they will cut back production and hence we observe not actual overproduction but potential overproduction in the sense that the capital and labour available for output could produce more than is actually being produced.

The other major theme is the underconsumptionist approach to overproduction was the unequal distribution of income, especially the relative shares of wages and profits in national income. The argument here, elements of which can be found in Sismondi and Malthus, for example, was most strongly made by J.A. Hobson. He argued that the low level of average wages leads to a high level of saving and capital accumulation, which could not receive a satisfactory rate of return because of the low consumption that the unequal distribution of incomes entailed.

We have already mentioned that fears of overproduction and/or underconsumption go back into the history of economic discussion but there are three contributions in particular that merit special discussion and of which we now give a brief account. They are (i) the Malthus–Ricardo debate; (ii) Marx's treatment of overproduction; and (iii) Keynes's treatment of overproduction.

In the early 19th century, at the end of the Napoleonic wars, there was an important economic controversy which has at its heart the very question of the possibility of overproduction, or, as it was then called 'general glut'. This was the famous controversy between Malthus and Ricardo; it was the first major debate on the subject of overproduction where, for the first time, Say's Law was paraded against the overproductionists and came out victorious. This victory really lasted in official, academic circles until Keynes's *General Theory* of the mid-1930s.

A brief look at this debate is instructive for any understanding of the disappearance of worries about macro-performance from the mainstream of economics until Keynes, although it remained central to the 'underworld' of Marxist and other dissenting branches of the subject.

The economic depression of the United Kingdom that followed the Napoleonic Wars had various contemporary explanations. The standard explanation, very much espoused by David Ricardo, was that with the transition from war to peace the structure of production had to be realigned to the demands of a peacetime economy so that there were 'sudden changes in the channels of trade' with some markets in excess demand and some in excess supply, but general excess supply was impossible.

Other writers, amongst whom perhaps Malthus was the most prominent, argued that the depression was due to the failure of effective demand, which had been boosted during the war by government expenditure, and that overproduction had occurred and caused the stagnation of output and employment. In the event Ricardo, using what we earlier termed Say's Law, won the day, and arguments that positive government macro-policy were needed to ensure stability in the volume of employment disappeared from orthodox reasoning until the victory of Keynesian economics. It is of interest to note that the current version of the New Classical Economics has reverted to the pre-Keynesian way of thinking.

Marx's attitude towards overproduction and underconsumption as its major explanatory cause is complicated and has caused much debate among Marxist scholars, so that a brief summary of his position is fraught with difficulty. He certainly did not accept the inevitable argument of the former that production could *never* be sold at

profitable prices. The analysis he gives of the conditions for sectoral balance are sufficient to demonstrate that equilibrium was possible. His refutation of Say's Law really comes from his analysis of capitalist monetary economics and its contrast with a pre-capitalist, barter-type economy. Once commodities no longer exchange for commodities but for money, which may or may not then exchange immediately for further commodities, the possibility arises that a particular volume of output may not be sold at profitable prices hence contractions of output and employment will occur.

It is not clear that Marx's analysis of possible crises can be classified as underconsumptionist in the sense depressions that are thought to be due to insufficient demand for consumer goods. Marx's theory of the cycle is basically constructed around a theory of fluctuating investment, as have practically all subsequent theories of the cycle, and the sequence he proposed was that an initial increase in demand for output would lead to an increase in the demand for labour which, in turn, would tend to force up the average real wage. This latter effect would reduce the rate of profit and check capital accumulation, hence output and employment would decline until real wages fell sufficiently to restore profitability. This is hardly an analysis that it is appropriate to level as underconsumptionalist.

Keynes, in his *General Theory*, set himself as one of his major tasks the overthrow of Say's Law. He accepted the idea that there was a tendency towards macro-equilibrium under capitalist organization. But, and it is an important but, this equilibrium was where the forces of investment and saving were balanced which was not necessarily at a level of output that would ensure full employment of labour. He further argued that this macro-equilibrium was stable so that any higher level of output produced, unless there were changes in the structural parameters of the system, would result in losses and so output and employment would contract back to its equilibrium level. In this sense Keynes accepted overproduction as a basic feature of capitalism. Was Keynes also an underconsumptionist? Not in the sense that he thought a failure of consumption expenditure was an initiating cause of a downturn in economic activity. He regarded consumption as reacting passively to income, so that a fall in income has to precede the fall in consumption in Keynesian analysis. However, his theory of the multiplier does suggest that a rise in the average or marginal propensity to consume would, for any given level of investment, lead to a higher level of output and employment.

We have seen that the term overproduction is not without its ambiguities and it is perhaps for this, and other reasons that we have given, that it is no longer in common use in orthodox economies. However the very fact that we currently observe capitalist economics producing well below their full-employment potential should make us take seriously those earlier fears of overproduction and the analyses that seek to demonstrate why it occurs. Moreover, any macrotheorizing that is prepared to acknowledge that total output may be demand constrained is accepting (implicitly or explicitly) the possibility of potential overproduction in capitalist economies.

See Also

- ▶ Malthus and Classical Economics
- ▶ Say's Law

Bibliography

Bleaney, M. 1976. *Underconsumption theories*. London: Lawrence and Wishart.

Douglas, C.H. 1933. *Social credit*. London: Eyre and Spottiswoode.

Hobson, J.A., and A.F. Mummery. 1889. *The physiology of industry*, 1956. New York: Augustus Kelley.

Keynes, J.M. 1936. *The general theory of employment, interest and money*, Vol. VIII. London: Macmillan. Reprinted in Collected Writings of John Maynard Keynes, 1973.

Mandeville, B. de. 1724. *The fable of the bees*, ed. F.-B. Kaye. Oxford: Clarendon Press, 1924

Malthus, T.R. 1836. *The principles of political economy*, 2nd edn. London: William Pickering. Reprinted, New York: Augustus Kelley, 1951.

Ricardo, D. 1951. Notes on Malthus. In *The works and correspondence of David Ricardo*, vol. II, ed. P. Sraffa. Cambridge: Cambridge University Press.

Spence, W. 1808. *Britain independent of commerce*, 4th ed. London: Cadell and Davies.

Over-Saving

Michael Bleaney

The possibility that saving could disrupt the circulation of commodities through a lack of demand was recognized at least as early as the Physiocrats. A similar argument was presented by Adam Smith. However, in both cases the analysis referred only to hoarding (i.e., the accumulation of a stock of money outside the banking system) and not to savings which were lent at interest to finance investment. Indeed both Smith and the Physiocrats regarded saving in order to transfer resources to investment with favour and were anxious to promote it. In neither case was the possibility that planned savings and planned investment could diverge examined (other than by hoarding, which was dismissed by Smith as irrational because of the loss of interest involved and therefore by implication insignificant), so that there was an implicit assumption that a decision to save was also a decision to invest.

Nevertheless the first series of writers to present over-saving as a serious problem to the economic system proved unable to pinpoint the fallacy in the equation of saving and investment intentions, which was undoubtedly a major reason for their defeat. Prominent amongst them was Thomas Malthus, but some of the elements of Malthus's arguments are to be found in earlier work by the Earl of Lauderdale (1804) and William Spence, whose pamphlet *Britain Independent of Commerce* (1808) stimulated James Mill to his well-known restatement of Say's Law in *Commerce Defended* (1808). Malthus claimed that if savings were at too high a level they would cause a deficiency of effective demand, which would make investment unprofitable. Recent attempts to formalize his ideas along these lines include Costabile and Rowthorn (1985) and Eltis (1984). These authors cite passages from Malthus which they take to indicate that he did not assume equality of planned saving with planned investment. However, many commentators (for references see those just cited) have held the contrary opinion, for the reason that at many points Malthus seems to be arguing something quite different. Thus in his *Principles of Political Economy* the burden of the argument appears to be that the transfer of resources from consumption to investment will inevitably lead to an increase in supply whilst simultaneously depressing demand. This does not imply a distinction between planned saving and planned investment (indeed quite the opposite) but rather reflects confusion over different time periods in the analysis: investment expenditure represents demand during the gestation period, and only on completion results in increased (potential) supply – but at this point the project ceases to absorb current savings.

A slightly different strand of argument, initiated by Sismondi, emphasized the impoverishment of the masses in the factory system and argued that this created a problem of lack of markets. As a comment on industrialism this line of argument was taken up by certain writers within the Russian populist movement in the later 19th century, notably V. Vorontsov and N.F. Danielson (Bleaney 1976). As a theory of crises it became absorbed into labour movement culture as theoretical support for demands for higher wages. The essential idea was that inequality in the distribution of income created too high a propensity to save. This notion was developed most cogently by J.A. Hobson in a number of books (Hobson 1902, 1909, 1922).

In general these theories lacked an adequate discussion of the determinants of investment, or exhibited a tendency to believe that there was a stringent upper limit to the rate of investment which was compatible with a balanced economy; the existence of alternative growth paths, characterized by different rates of investment, was implicitly denied (or held to be true only within strict limits). There was usually no explicit discussion of the loanable funds theory, according to which the problem would be resolved by a sufficiently low interest rate discouraging saving and encouraging investment. But one exception is Hobson (1922), who defends his position mainly on the grounds that saving and investment are relatively insensitive to the rate of interest.

It is interesting to consider these over-saving theories in the light of Kaldor's (1955–6) theory of income distribution. In this theory redistribution of income between wages and profits at full employment is the mechanism by which planned savings are adjusted to planned investment. Over-saving theorists such as Hobson could be interpreted as stating that aggregate planned savings were too high relative to planned investment because the real wage was persistently too low and profits too high; the required redistribution from profits to wages (i.e., from those with a higher to those with a lower marginal propensity to save) was prevented because the weak bargaining position of labour ensured that any fall in prices would be matched by a compensating fall in money wages. Hence the fall in savings tended to come about through a contraction of output rather than a redistribution of income at full employment.

There was some tendency to link this idea in the years around 1900 to the growth of trusts and cartels and the increasing concentrating of industry. This can be discerned in the works of Hobson and Hilferding (1910); the latter argues that capitalism is entering a new stage in which its original dynamism has given way to a tendency to stagnation. In Hobson's view the increasing concentrating of industry exacerbates the over-saving problem by raising profits at the expense of wages; the activities of trade unions, therefore, help to resolve it. Hilferding, by contrast, puts more emphasis on the redistribution of profits from capital outside to capital inside the cartels; he argues that this reduces the aggregate volume of investment because cartels obtain their high rate of profit only by restricting output.

In the post-war period it has sometimes been argued that a tendency to over-saving continues to exist but has been counteracted by the growth of military expenditures. This idea goes back at least a century, to Vorontsov, and assumes that the tax revenue raised to finance military expenditure successfully absorbs saving rather than reducing consumption. Modern theories of over-saving have tended to rest on the Hobson–Hilferding argument that an increasing concentration of industry redistributes income towards profits; Cowling (1982) represents the most coherent attempt to develop this theme, based on a model developed from the work of Kalecki, but including a much more detailed discussion of the theory of oligopoly. According to his data the degree of monopoly (the ratio of price to prime cost) increased significantly in the UK from 1945 to 1975. However this was entirely accounted for by the rise in the proportion of salaried workers (allocated by Cowling to fixed costs) and may simply reflect technical developments. Even though industry has become more concentrated at a national level since 1945 (measured by output), many observers would argue that industrial markets have become less concentrated because of reductions in tariff barriers and transport costs.

See Also

► Underconsumptionism

Bibliography

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

Costabile, L., and R.E. Rowthorn. 1985. Malthus's theory of wages and growth. *Economic Journal* 95: 418–437.

Cowling, K. 1982. *Monopoly capitalism*. London: Macmillan.

Eltis, W.A. 1984. *The classical theory of economic growth*. London: Macmillan.

Hilferding, R. 1910. *Finance capital*. Trans. M. Wathnick and S. Gordon, ed. T. Bottomore. London: Routledge, 1981.

Hobson, J.A. 1902. *Imperialism*. London: Nisbet.

Hobson, J.A. 1909. *The industrial system*. London: Longman.

Hobson, J.A. 1922. *The economics of unemployment*. London: Allen & Unwin.

Kaldor, N. 1955–6. Alternative theories of distribution. *Review of Economic Studies* 23: 83–100.

Lauderdale, Earl of. 1804. *An inquiry into the nature and origin of public wealth*. Edinburgh: Constable.

Mill, J. 1808. *Commerce defended*. London/Edinburgh: Oliver & Boyd, 1966.

Spence, W. 1808. *Britain independent of commerce*, 4th ed. London: Cadell & Davies.

Overshooting

Jürg Niehans

An economic variable may 'overshoot' its steady-state value in many different contexts. In recent economic theory the term has assumed a more specific meaning, describing a characteristic relationship between current returns and capital gains on financial assets.

The total yield of an asset, i, consists of the current return, r (rental, dividend, coupon), and the capital gain, $\dot{p}$ both expressed as a proportion of its market price, p:

$$i = \frac{r}{p} + \frac{\dot{p}}{p}.$$

In the steady state, $\dot{p}/p = 0$ and $i = r/p$. Suppose arbitrage sees to it that i is always equal to the yield on other assets, j, regarded as given and constant. Suppose further that there is some new information indicating that r will be above its steady-state level for a limited period. As a consequence, there will be an instantaneous increase, or 'jump', in the asset price reflecting the present value of the extra returns. From then on, the temporary gain in r will be continuously matched by a capital loss, so that $\dot{p}/p < 0$. The expectation of a limited period of extra returns will thus produce an instantaneous appreciation of the asset followed by gradual depreciation. This saw-tooth pattern of the asset price is what is called overshooting. While overshooting in a general sense may well be due to speculative excesses, 'bubbles' and mistaken expectations, it is important to note that in the more specific sense described here it is not only consistent with, but an implication of, the correct anticipation of the consequences of unexpected disturbances.

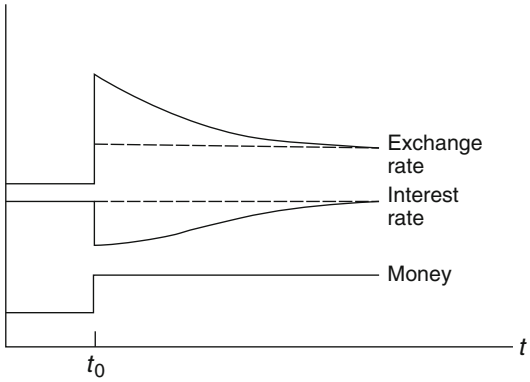
While overshooting, as such, is a commonplace feature of asset markets, it is particularly important in foreign exchange markets, where it was observed by Gustav Cassel around 1920. The excess supply of German marks, he argued, had depressed their foreign exchange value so far below its longer-term equilibrium that the expectation of future appreciation attracted speculators, even at relatively low interest rates. Unfortunately, since Cassel did not bother to provide an analytical elaboration, his insight was lost for more than half a century, to be rediscovered around 1974. It was first developed into a theoretical model by Dornbusch (1976). A compact, up-to-date survey of overshooting theory is provided in Obstfeld and Stockman (1985). A somewhat less technical overall perspective is given in Niehans (1984).

Overshooting has to be defined with reference to an equilibrium exchange rate. For a purely monetary disturbance (and in the absence of government debt), the appropriate reference point is purchasing-power parity. PPP relates to the parallel effects of an exogenous increase in the supply of fiat money on exchange rates and prices. It postulates, specifically, that these effects are proportionally equal, which implies that exchange rates and prices move in step. The proposition clearly relates to the comparison of steady states. In the short run, the effects of money on exchange rates can deviate very considerably from those on commodity prices. These deviations are the main subject of overshooting theory.

The difference between the change in the exchange rate and the contemporaneous change in the international commodity price ratio is often called the change in the real exchange rate. In the steady state, an exogenous increase in the money supply has no effect on the real exchange rate. Overshooting implies, however, that there may be sharp fluctuations in real exchange rates in the short run.

In Dornbusch's model, overshooting is essentially due to the view, rooted in the tradition of macroeconomics, that asset prices are highly flexible whereas output prices are inert. Suppose there is an unexpected increase in the money supply at time t_0 (see Fig. 1).

With sticky prices, this will be reflected in an immediate increase in real balances and thus a decline in the rate of interest. As prices gradually



Overshooting, Fig. 1

move upward, the interest rate will rise again toward its equilibrium level. Since international arbitrage equalizes foreign and domestic yields, low domestic yields must be accompanied by a declining price of foreign exchange (with the slope of the exchange rate curve reflecting the interest differential). Under perfect foresight, a gradual decline during the adjustment process is achieved by an instantaneous overshooting of the exchange rate relative to its steady-state level derived from purchasing-power parity. The dynamic properties of such a system were analysed by Gray and Turnovsky (1979); they generally involve saddle-point instability. The size of the initial overshooting has to be determined by reckoning 'backward' from the steady state.

Instead of a step-like increase in the money supply, the underlying disturbance may be an increase in the rate of monetary expansion (Frankel 1979). In this case, because of continuing inflation, overshooting cannot be defined with reference to a steady state of the nominal exchange rate. Nor is it certain that the nominal exchange rate will temporarily decline after the instantaneous increase. However, there will still be overshooting of the nominal exchange rate relative to PPP and thus in the real exchange rate. A reversal of direction may also be absent if the market takes time to recognize a change in monetary policy (Moser 1983). Econometric estimates by Driskill (1981) indicate overshooting by a factor of 2.3 in the dollar price of the Swiss franc. Moser's work suggests that this estimate may be too high because not all changes in the Swiss money supply during the period in question could legitimately be regarded as exogenous. Generally, overshooting seems to be quite sensitive to variations in conditions and model specification. It would not be surprising, therefore, if econometric estimates differed widely.

Besides interest arbitrage, there are other causes of overshooting exchange rates. Of particular importance is the fact that an economy cannot acquire (net) foreign assets overnight, but only over a period of current-account surpluses. This type of portfolio mechanism was first investigated by Kouri (1976) and further developed by Calvo and Rodriguez (1977) and Branson (1979). An exogenous increase in the money supply, with sticky prices, results in an increased demand for international assets. Since the stock of such assets cannot be immediately increased, there is an instantaneous depreciation of the domestic currency, implying overshooting relative to PPP. As domestic prices creep upward, again reducing real balances, the composition of portfolios gradually returns to the initial situation, overshooting subsides, and the exchange rate approaches purchasing-power parity. The nominal exchange rate may also overshoot its equilibrium level, but this is not certain. The sequence of current-account surpluses and deficits during the adjustment process is even more uncertain (Frenkel and Rodriguez 1982; Niehans 1984). Even for small open economies, the overshooting mechanism thus turns out to be quite complicated. In interdependent economies, the taxonomy of dynamic patterns becomes yet more complex (Niehans 1977).

After other than purely monetary disturbances, overshooting of exchange rates may occur even with perfectly flexible prices (Dornbusch and Fischer 1980; Kouri, 1983). In the case of a spontaneous increase in the domestic demand for foreign assets, the additional assets can only be provided through current-account surpluses. These, in turn, require a temporary rise of the exchange rate above its equilibrium level, which means overshooting. Asset arbitrage will see to it that the gradual re-appreciation of the domestic

currency following the instantaneous depreciation is associated with an interest differential in favour of foreign rates. But low domestic interest rates result in an increased demand for real balances, which can only be satisfied at domestic prices below their equilibrium level. The overshooting of the exchange rate is thus associated with an undershooting of interest rates and prices, followed by a gradual return to equilibrium. During this process, a depressed, but gradually appreciating, currency is accompanied by a current account surplus and thus a capital outflow. In general, however, there is no clearcut correspondence between exchange overshooting and capital flows.

Since the collapse of the gold-exchange standard in 1973, exchange rates seem to have fluctuated more than their underlying determinants (like money supplies or incomes), and also more than leading monetary theorists had expected. The theory of overshooting suggests that this may be due to the way asset markets work even under perfect foresight. As already realized by Cassel, the resulting fluctuations in real exchange rates may be the source of potentially serious disturbances in the trade, output and employment of the countries concerned. Indeed, overshooting turned out to be the principal policy problem of floating rates.

This raises the question whether overshooting could be dampened or even eliminated by suitable monetary and foreign exchange policies. Various methods have been proposed or debated. The most radical is the return to fixed exchange rates. This would eliminate overshooting by suppressing any movements in exchange rates, thus depriving countries of their monetary autonomy. Other proposals would limit exchange rate movements to a slow 'crawl' (Williamson, 1981), but it is doubtful that they would be workable without exchange control and international policy coordination. The so-called OPTICA proposal (CEC, 1977) postulated that foreign exchange interventions be used to keep exchange rates at purchasing-power parity even in the short run. This may make it impossible for central banks to follow a non-inflationary course and also raises serious stability problems. At the present time there seem to be no tested techniques whereby central banks could confidently expect to dampen overshooting without compromising other objectives. It may be better, therefore, not to rely on automatic schemes and to meet each case of serious overshooting on its merits. The most basic policy rule surely is to avoid abrupt shifts in the course of monetary policy.

See Also

► International Finance

Bibliography

Branson, W.H. 1979. Exchange rate dynamics and monetary policy. In *Inflation and employment in open economies*, ed. A. Lindbeck. Amsterdam: North-Holland.

Calvo, G.A., and C.A. Rodriguez. 1977. A model of exchange rate determination under currency substitution and rational expectations. *Journal of Political Economy* 85(3): 617–625.

Cassel, G. 1921. *The world's monetary problems: Two memoranda to the league of nations*. London: Constable.

Commission of the European Communities (CEC). 1977. *Inflation and exchange rates: Evidence and policy guidelines for the European community.* OPTICA Report 1976, Brussels.

Dornbusch, R. 1976. Expectations and exchange rate dynamics. *Journal of Political Economy* 84(6): 1161–1176.

Dornbusch, R., and S. Fischer. 1980. Exchange rates and the current account. *American Economic Review* 70(5): 960–971.

Driskill, R.A. 1981. Exchange-rate dynamics: An empirical investigation. *Journal of Political Economy* 89(2): 357–371.

Frankel, J.A. 1979. On the mark: A theory of floating exchange rates based on real interest differentials. *American Economic Review* 69(4): 610–622.

Frenkel, J.A., and C.A. Rodriguez. 1982. Exchange rate dynamics and the overshooting hypothesis. *IMF Staff Papers* 29(1): 1–30.

Gray, M.R., and S.J. Turnovsky. 1979. The stability of exchange rate dynamics under perfect myopic foresight. *International Economic Review* 20(3): 643–660.

Kouri, P.J.K. 1976. The exchange rate and the balance of payments in the short run and in the long run: A monetary approach. *Scandinavian Journal of Economics* 78(2): 280–304.

Kouri, P.J.K. 1983. Balance of payments and the foreign exchange market: A dynamic partial equilibrium model. In *Economic interdependence and flexible exchange rates*, ed. J.S. Bhandari and B.H. Putnam. Cambridge, MA: MIT Press.

Moser, B. 1983. *Der frankenkurs des dollars 1973–1980: Ein test der kaufkraftparitätentheorie*, Berner Beiträge zur Nationalökonomie, vol. 43. Bern: Haupt.
Niehans, J. 1977. Exchange rate dynamics with stock/flow interaction. *Journal of Political Economy* 85(6): 1245–1257.
Niehans, J. 1984. *International monetary economics*. Baltimore: Johns Hopkins University Press.
Obstfeld, M., and A.C. Stockman. 1985. Exchange-rate dynamics. In *Handbook of international economics*, vol. II, ed. R.W. Jones and P.B. Kenen. Amsterdam: North-Holland.
Williamson, J. 1981. *Exchange rate rules: The theory, performance and prospects of the crawling peg*. New York: St Martin's Press.

Overstone, Lord [Samuel Jones Loyd] (1796–1883)

D. P. O'Brien

Keywords

Balance of payments; Bank Charter Act (1844); Banking principle; Currency principle; Endogenous trade cycle; Liquidity preference; Monetary base; Overstone, Lord; Palmer rule; Specie-flow mechanism

JEL Classifications

B31

Samuel Loyd (the single '1' seems to have been a device adopted by his father to shake off Welsh relatives), Lord Overstone, was born on 25 September 1796, the son of Lewis Loyd, a Unitarian minister turned banker, and Sarah Loyd (née Jones), the daughter of a Manchester banker. Lewis Loyd's drive and ability transformed an obscure provincial bank into a major concern. An MP from 1819 to 1826, Overstone only began to devote himself seriously to banking after the death of his mother in 1821. Though perhaps lacking his father's flair, he was a shrewd and successful banker, influential with his contemporaries. He retired from business only in 1850, on his elevation to the peerage by Lord John Russell.

In 1837 he entered, with considerable effectiveness, the arena of monetary controversy with his *Reflections suggested by a perusal of Mr. J. Horsley Palmer's pamphlet on the Causes and Consequences of the pressure on the Money Market.* This was not his first statement on the matter; he had been a witness before the 1832 committee on the renewal of the Bank Charter. But it was the start of his preeminence as a monetary writer, a pre-eminence which was to prove decisive in the debates leading up to the renewal of the Bank Charter in 1844 and which shaped the institutional framework of British monetary policy from that time until the First World War. Overstone's monetary thought starts from a position that the economy contains an endogenous trade cycle – he was indeed one of the first people to identify the stages of the cycle. Monetary policy could then be procyclic, responding to the needs of customers (the Banking principle) or it could act counter-cyclically so as to stabilize the level of prices and activity (the Currency principle). The theoretical position underlying the latter was as follows. In the upswing of the cycle money income rose, exports were less competitive, and a balance of payments deficit developed. Counter-cyclical contraction of the currency, in line with the loss of specie through the balance of payments deficit, would then moderate the upswing and prevent it getting out of hand. Conversely, in the lower half of the cycle, with a balance of payments surplus, the money supply would be increased. (O'Brien 1971; O'Brien 1975; Wood 1939)

The origins of this position were threefold: Hume's theory of the balance of payments, positing a direct link between the money supply, the price level, exports, and imports: the Ricardian theory of the equilibrium distribution of the precious metals (that when countries were in relative money income equilibria, there would be no net flows of precious metal) deriving from Hume; and the Ricardian definition of 'excess'. The last is particularly crucial. If specie was flowing out then, by definition, there was excess currency. This idea leads in turn to the principle, formulated in 1826 by several writers, of 'metallic fluctuation': a paper currency should fluctuate in amount

exactly as an identically circumstanced metallic one would do.

On this basis Overstone emerged as a critic of the 'Palmer Rule' under which the Bank allowed drains of specie to fall on deposits equally with notes: unless deposits were as important as notes in correcting the price level in relation to the balance of payments, the drain might exhaust the specie without correcting the balance of payments. Overstone's emphasis was on control of currency as the high-powered money base, with deposits as part of an inverted credit pyramid lacking any independent effect of their own, and dependent upon the currency base if banks behaved properly with respect to reserve ratios.

Thus, fundamental to monetary control was separation of departments in the Bank: the Banking department followed Banking principles, but the Issue department must follow Currency principles and thus stabilize economic activity, following automatic rather than discretionary procedures.

The role of the rate of interest in all this was twofold: short-run balance of payments correction, although this could only be a palliative if relative money incomes were out of line, and the production of an effect on confidence which in turn affected liquidity preference through increasing precautionary reserve holdings when the rate was raised, thus reducing the effectiveness of a given money supply. This variation in liquidity preference with confidence was an important part of the analysis, and was built into the 1844 Act with weekly publication of the Bank reserves, which were supposed to cause prudent adjustment of other reserves. This in turn would avoid the Bank of England's having to act as lender of last resort, a role which Overstone opposed as incompatible both with inducing the rest of the system to respond counter-cyclically and with the necessary limitation of the high-powered base.

Overstone was a many-sided man. But it is as a monetary theorist that he is chiefly remembered.

See Also

► Banking School, Currency School, Free Banking School

Selected Works

1857. *Tracts and other publications on metallic and paper currency*. London: privately printed. London: Longmans, 1858.

1858. *The evidence given by Lord Overstone, before the Select Committee of the House of Commons of 1857, on Bank Acts, with additions*. London: Longmans.

Bibliography

O'Brien, D.P., ed. 1971. *The correspondence of Lord Overstone*, 3 vols. Cambridge: Cambridge University Press.

O'Brien, D.P. 1975. *The classical economists,* ch. 6. Oxford: Clarendon Press.

Wood, E. 1939. *English theories of central banking control, 1819–1858*. Cambridge, MA: Harvard University Press.

Owen, Robert (1771–1858)

N. W. Thompson

Keywords

Autarky; Class conflict; Labour time; Owen, R.; Pecuniary and non-pecuniary penalties; Poverty alleviation; Socialism

JEL Classifications

B31

Born in Newtown, Montgomeryshire (Powys), in 1771, Robert Owen was in many ways both the child and the victim of his age, making his fortune as a cotton manufacture involved in the industrial transformation of Britain and dissipating it in his efforts to eliminate its evils. With the purchase of the New Lanark cotton mills in 1797 Owen did, for a time, successfully combine the roles of factory owner and social reformer, showing how a humanized working environment might effect a reformation in human character. For the modern social scientist, one interesting innovation Owen

implemented was the silent monitor, a four-sided block that was hung next to each worker's machine; a supervisor would turn the block to a colour that reflected the worker's effort during the day; colours were recorded in a 'book of character'. (See Podmore 1906, based on Owen's autobiography.) The silent monitor was meant to substitute for corporal punishment as a discipline device; it resonates with recent thinking on social sanctions; see pecuniary versus non-pecuniary penalties.

Owen's success in the New Lanark venture encouraged him to devote his life to the regeneration of mankind and it also provided him with the funds necessary to attempt this. However, further practical experiments proved disastrous. The cooperative communities he established, such as those at New Harmony Indiana in 1824 and Queenwood in Hampshire in 1839, soon collapsed, while his efforts in 1832 to socialize money through a National Equitable Labour Exchange proved equally disastrous. However, such failures never inspired self-doubt and Owen remained to the end of his long life a living embodiment of hope's capacity to triumph over experience.

As a cotton manufacturer Owen grasped the potential for material abundance which industrialization was creating in early 19th-century Britain; yet as an acute observer of economic life he was equally aware of the existence of widespread material impoverishment. His chief concern in his economic writings was, therefore, to investigate this paradox of poverty in the midst of abundance and show how it might be resolved.

For Owen the realization of economic prosperity for all was obstructed by the tendency, in a competitive market economy, for rapid mechanization to create 'a most unfavourable disproportion between the demand for and supply of labour'. This resulted in its progressive devaluation which in turn caused a diminution in consumption and a general economic crisis as manufacturers responded to a deficiency of effective demand by reducing output and laying off labour. As Owen phrased it, 'It is want of a profitable market that alone checks the successful and otherwise beneficial industry of the labouring-classes'.

To remove this constraint upon production and to realize the potentialities of industrial development, Owen believed that 'Human labour [should] acquire its natural and intrinsic value, which would increase as science advanced', and to secure this Owen argued in such works as his *Report to the County of Lanark* (1821) that goods should be valued according to the labour time that they embodied and exchanged against labour notes rather than conventional money. Such a socialization of exchange, Owen believed, would give labour its whole product and further ensure that aggregate supply and aggregate demand expanded *pari passu.*

It was these ideas which bore practical fruit in the National Equitable Labour Exchange, where attempts were made to value goods and reward labour in terms of time. As might be expected this institution suffered a speedy demise. However, it was never seen by Owen as more than a stepping stone to his ideal of a 'new moral world' of neo-autarkic cooperative communities, where each would contribute to the common stock according to ability and consume according to need. Insulated thus against the exploitation and vagaries of a competitive market economy, material well-being could be assured and the character of man created anew.

Owen's economic writing was only one facet of a more general attempt to construct a science of society – a science which would have both an explanatory and prescriptive power and which could be used to determine the means necessary to transform man from an egotistical, competitive atom into a truly social being. It was this broader intellectual enterprise which enthused and interested British socialist thinkers in the first half of the 19th century, as can be seen, for example, in their redefinition of 'political' as 'social' or 'moral' economy.

Engels in *The Condition of the Working Class in England* (1844) remarked that, 'English socialism arose with Owen, a manufacturer, and proceeds therefore with great consideration towards the bourgeoisie', and, undoubtedly, Owen's tendency to stress the socially harmonious future and the ultimate reconcilability of class antagonism, rather than the social hostilities of the present, left its quietistic mark upon Owenite socialism. Yet for socialist writers such as Thompson, Gray and

Bray, Owen's real legacy was methodological rather than ideological. What they imbibed from Owen was a particular, social scientific way of approaching the condition of labour rather than any unwillingness to unearth the roots of social antagonism.

Selected Works

1813. *A new view of society, essays on the formation of human character*. London: Printed for Cadell and Davies.

1815. *Observations on the effect of the manufacturing system,* 2nd ed. London.

1817. Report to the committee for the relief of the manufacturing poor. In *The life of Robert Owen written by himself,* 2 vols. London, 1857–8.

1818. Two memorials on behalf of the working classes. In *The life of Robert Owen written by himself,* 2 vols. London, 1857–8.

1819. *An address to the master manufacturers of Great Britain.* Bolton.

1821. *Report to the county of Lanark of a plan for relieving public distress.* Glasgow: Glasgow University Press.

1823. *An explanation of the cause of distress which pervades the civilized parts of the world.* London: Printed for the British and Foreign Philanthropic Society.

1832. *An address to all classes in the state. London.*

1849. *The revolution in the mind and practice of the human race.* London.

Bibliography

Beales, H.L. 1933. *The early English socialists*. London: Hamish Hamilton.

Beer, M. 1953. *A history of British socialism,* 2 vols. London: Allen & Unwin.

Butt, J. 1971. Robert Owen in his own time 1771–1858. In *Robert Owen and his relevance to our time,*. Cooperative college papers no. 14. Loughborough: Cooperative Union.

Cole, G.D.H. 1930. *Robert Owen*. London: Macmillan.

Cole, M. 1971. Owen's mind and methods. In *Robert Owen, prophet of the poor*, ed. S. Pollard and J. Salt. London: Macmillan.

Cole, G.D.H. 1977. *A history of socialist thought,* 5 vols, vol. 1, Socialist thought: The forerunners, 1789–1850. London: Macmillan.

Engels, F.W. 1844. *The condition of the working class in England.* London: Panther, 1974.

Garnett, R.G. 1971. *Co-operative and Owenite socialist communities in Britain and America, 1825–45*. Manchester: Manchester University Press.

Gray, A. 1967. *The socialist tradition, Moses to Lenin*. London: Longman.

Harrison, J.F.C. 1969. *Robert Owen and the Owenites in Britain and America: The quest for the new moral world*. London: Routledge & Kegan Paul.

Oliver, W.H. 1958. The labour exchange phase of the co-operative movement. *Oxford Economic Papers* 10: 355–367.

Podmore, F. 1906. *Robert Owen: A biography.* London/Honolulu: Hutchinson/University Press of the Pacific, 2004.

Pollard, S. 1971. Robert Owen as an economist. In *Robert Owen and his relevance to our times,* Co-operative college papers no. 14. Loughborough: Co-Operative Union.

Royle, E. 1998. *Robert Owen and the commencement of the Millennium: The harmony community at queenwood farm, Hampshire, 1839–1845*. Manchester: Manchester University Press.

Thompson, N.W. 1984. *The people's science: The popular political economy of exploitation and crisis, 1816–34*. Cambridge: Cambridge University Press.

Woodward, L. 1962.*The age of reform: 1915–1870.* Oxford: Clarendon Press (for broad historical context.)

Own Rates of Interest

John Eatwell

The concept of the own-rate of interest on a commodity was introduced (though not named) by Piero Sraffa in his review (1932) of Friedrich von Hayek's book *Prices and Production* (1931), and was later taken up, and labelled, by Maynard Keynes in his analysis of the role of money in the theory of employment (1936, ch. 17). Sraffa introduced the concept by means of the example of a cotton spinner who borrows money to purchase a quantity of raw cotton today (at the spot price) which he simultaneously sells forward (Sraffa 1932, p. 50). The spinner is actually borrowing cotton for the period of the

transaction, say, one year. The own-rate of interest on cotton is then the spot price of a bale of cotton for divided by the future price of a bale discounted at the going money rate of interest; less one. So if the price of 100 bales of cotton for delivery today is \$20, and the price to be paid for delivery of 100 bales in one year's time is \$21.40, whilst the money rate of interest is 5%, then the own-rate of interest on cotton is

$$\frac{20}{21.40/1.05} - 1 = c. - 2\% (\text{See Keynes 1930, p. 223}).$$

Sraffa's interpretation of the role of the money rate of interest in the calculation was *not* that it was simply the rate of interest on a numeraire. 'Money' in his discussion, is the actual financial medium. So the money rate represents the normal rate of interest (which is assumed equal to rate of profit) in the economy as a whole. The difference between the money rate and own-rate of interest on a commodity therefore indicates that the *spot market* for that commodity is not in normal long-run equilibrium.

In equilibrium the spot and forward price coincide, for cotton as for any other commodity; and all the 'natural' or commodity rates are equal to one another, and to the money rate. But if, for any reason, the supply and the demand for a commodity are not in equilibrium (*i.e.,* its market price exceeds or falls short of its cost of production), its spot and forward prices diverge, and the 'natural' rate of interest on that commodity diverges from the 'natural' rates on other commodities. Suppose there is a change in the distribution of demand between various commodities; immediately some will rise in price, and other will fall; the market will expect that after a certain time, the supply of the former will increase, and the supply of the latter fall, and accordingly the forward price, for the date on which equilibrium is expected to be restored, will be below the spot price in the case of the former and above it in the case of the latter; to the effecting of the [restoration of equilibrium] as is the divergence of prices from the costs of production; it is, in fact, another aspect of the same thing (1932, p. 50).

In terms of the example, the equilibrium price is \$21.40 and the equilibrium rate of interest is 5%. However, the current price of 100 bales of cotton is \$20, which invested at the going rate of interest, would be worth only \$21, at the end of a year, and this would buy c.98 bales of cotton at the equilibrium price then ruling. Thus the own-rate of interest on cotton is c. –2%. The concept of the own-rate of interest on a commodity interpreted in this way, can only be defined with respect to normal prices and to the normal interest rate, represented by the money rate of interest. For example, if the money rate of interest in the above instance were 10% the own-rate of interest on cotton would be c. 3%; if 0%, then c. –7%.

Keynes used Sraffa's idea in his analysis of the determination of the level of investment. His theory of the rate of interest was derived from an analysis of the demand for the stock of monetary assets – that demand being the sum of transactions, precautionary, and speculative demands – with only the latter being regarded as a function of the rate of interest.

The elasticity of the liquidity preference schedule with respect to the rate of interest was based on two rates of interest, the rate which actually holds, and the rate which is expected to hold in the future (the long-run rate). The ambiguity introduced into Keynes's analysis by the construction of the liquidity preference schedule on the basis of the short-run and the long-run rate of interest was not totally clear in the *General Theory*, other than in Keynes's ambivalence over whether the rate of interest was a 'psychological' or a 'conventional' variable (see Keynes 1936, pp. 200–202). The reference to 'convention' established the idea that 'institutional' or 'historical' factors might be the underlying determinants of the long-run rate of interest; he is content to point to forces other than supply and demand and leave the issue there.

The ambiguity in Keynes's theory is exposed in his theory of investment. There, he associates the equalization of rates of return on different categories of assets with the determination of the volume of investment. The idea that rates of return are equalised is characteristic of long-run analyses. Yet Keynes is suggesting that this equality is attained with respect to a rate of interest which

is determined as a short-run phenomenon. This ambiguity is unresolved in the *General Theory*. Subsequent discussion by Kaldor (1960) although the issue was clearly identified, left the problem unresolved.

The definition and interpretation of the own-rate of interest in modern general equilibrium theory (see Debreu 1959) are quite different from those advanced by Sraffa – although there is a formal similarity in the method of calculation. The set of equilibrium prices refer to commodities located at different points in time and yet include no interest charge. The price are discounted prices *which would be paid today* for commodities to be traded at a future date. The rate of interest at which the prices are discounted is not specified.

The own-rate of interest on a commodity in one production period, say time t to time $t+1$, is defined as the ratio of the appropriate discounted prices, less one: i.e., the own-rate of interest on commodity q over the time period t is

$$p_{qt} = \frac{p_{qt}}{p_{qt+1}} - 1$$

where p_{qt}, p_{qt+1} are the discounted prices in period 1 of the commodity q available at the beginning (resp. the end) of period t, the prices being determined in the manner shown. The calculation jsut shown contains no reference to a normal rate of interest. The own-rate of return is defined independently of any normal or money rate. By analogy with the case of a-capitalistic production, the ratio of the discounted prices p_{qt}, p_{qt+1} is equal to the marginal rate of substitution in consumption between q_t and q_{t+1}, and to the marginal rate of transformation in production.

So although commodities q_t and q_{t+1}, are defined as *different* commodities for the purpose of price determination, they are regarded as the *same* commodity for the purpose of the definition of the own-rate of interest, the difference in the prices being due to their temporal location.

Although there is some technical similarity in the calculation of the own-rate of interest on a commodity by both Sraffa and Debreu, the definition advanced by Sraffa is quite different from that adopted by Debreu. This difference stems from their different conceptions of prices and their formation. In Sraffa's formulation the own-rate of interest is a reflection of the divergence of the market price from normal equilibrium price (and the normal rate of interest). In Debreu's definition this latter distinction has no meaning. The discounted prices used in his calculation are equilibrium prices, but there is no normal rate of interest of normal long-run prices in the Marshallian sense of those terms. Thus differences in the own-rate of interest as between commodities arise not out of market price 'deviations', but out of his definition of a 'commodity'.

It should also be noted that markets of the type referred to by Debreu, on which payment is made today for commodities to be traded in the future, do not exist. On such futures markets as there are the prices set are those which will be paid at the time the trade is actually made (Debreu 1959, p. 33). Such prices could not be the basis for the calculation of the own-rate of interest on a commodity in the manner of Debreu.

Bibliography

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

Kaldor, N. 1960. Keynes' theory of the own-rates of interest. In *Essays on economic stability and growth*, ed. N. Kaldor. London: Duckworth.

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

Sraffa, P. 1932. Dr Hayek on money and capital. *Economic Journal* 42(March): 42–53.

von Hayek, F.A. 1931. *Prices and production*. London: Routledge & Kegan Paul.