N

Nash Equilibrium

David M. Kreps

The concept of a *Nash equilibrium* plays a central role in noncooperative game theory. Due in its current formalization to John Nash (1950, 1951), it goes back at least to Cournot (1838). This entry begins with the formal definition of a Nash equilibrium and with some of the mathematical properties of equilibria. Then we ask: To what question is ‘Nash equilibrium’ the answer? The answer that we suggest motivates further questions of *equilibrium selection*, which we consider in two veins: the informal notions, such as Schelling’s (1960) *focal points;* and the formal theories for *refining* or *perfecting* Nash equilibria, due largely to Selten (1965, 1975). We conclude with a brief discussion of two related issues: Harsanyi’s (1967–8) notion of a *game of incomplete information* and Aumann’s (1973) *correlated equilibria*.

Definition and Simple Mathematical Properties

We give the definition in the simple setting of a finite player and action game in normal form. There are I players, indexed by $i = 1,\ldots,I$. Player i chooses from N_i (pure) strategies; we write S_i for this set of strategies, and s_i for a typical member of S_i. A *strategy profile*, written $s = (s_1,\ldots,s_I)$, is a vector of strategies for the individual players – we write S for $\Pi_{i=1}^{I} S_i$, the set of all strategy profiles. For a strategy profile $s = (s_1,\ldots,s_I) \in S$ and a strategy $s_i' \in S_i$ for player i, we write $s|s_i'$ for the strategy profile $\left(s_1,\ldots,s_{i-1},s_i',s_{i+1},\ldots,s_I\right)$, or s with the part of i changed from s_i to s_i'. For each player i and strategy profile s, $u_i(s)$ denotes I’s expected utility or payoff if players employ strategy profile s.

Definition. A *Nash equilibrium* (in pure strategies) is a strategy profile s such that for each i and $s_i' \in S_i, u_i(s) = u_i\left(s|s_j'\right)$. In words, no single player, by changing his own part of s, can obtain higher utility if the others stick to their parts.

The basic definition is often extended to independently mixed strategy profiles, as follows. Given S_i, write $\sum_i$ for the set of mixed strategies for player i; that is, all probability distributions over S_i. Write $\sum$ for $\Pi^{I}_{i=1} \sum_i \sigma = (\sigma_1,\ldots,\sigma_I), \sigma|\sigma_i'$, and so on, as before. Extend the utility functions u_i from domain S to domain by letting $u_i\ (\sigma)$ be player I’s expected utility:

$$u_i(\sigma) = \sum_{s_1} \cdots \sum_{s_1} u_i(s_1,\ldots,s_I)\sigma_1(s_1)\ldots\sigma_I(s_I).$$

Then define a Nash equilibrium in mixed strategies just as above, with σ in place of s and σ_i in place of s_i. Equivalently, player i puts positive weight on pure strategy s_i only if s_i is among the pure strategies that give him the greatest expected utility.

© Macmillan Publishers Ltd 2018

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

This formal concept is due to John Nash (1950, 1951). Luce and Raiffa (1957) provided an important and influential early commentary. Nash also proved that in a finite player and finite action game, there always exists at least one Nash equilibrium, albeit existence can only be guaranteed if we look at mixed strategies – standard examples (such as matching pennies) shows that there are games with no pure strategy equilibria. The proof that a Nash equilibrium always exists is an application of Brouwer's fixed point theorem. The concept of a Nash equilibrium is extended in natural fashion to games with infinitely many players and/or pure strategies, although in such cases existence can be problematic; we do not discuss these matters further here.

The Philosophy of Nash Equilibrium

To what question is 'Nash equilibrium' the answer? This has been and continues to be the subject of much discussion and debate. Most authors take a position that is a variation on the following.

Suppose that, in a particular game, players *by some means unspecified at the moment* arrive at an 'agreement' as to how each will play the game. This 'agreement' specifies a particular strategy choice by each player, and each player is aware of the strategies chosen by each of his fellow players, although players may not resort to enforcement mechanisms except for those given as part of the formal specification of the game. One would not consider this agreement *self-enforcing* (or strategically stable) if some one of the players, hypothesizing that others will keep to their parts of the agreement, would prefer to deviate and choose some strategy other than that specified in the agreement. Thus, to be self-enforcing in this sense, it is *necessary* that the agreement form a Nash equilibrium. (If players could perform a public randomization as part of the agreement, we would get convex combinations of Nash equilibria as candidate self-enforcing agreements. See section VI for what can be done with partially private randomizations.)

This does not say that every Nash equilibrium is a self-enforcing agreement. For example, in the context being modelled, it might be appropriate to consider multi-player defections (and the concept of a *strong equilibrium*, a strategy assignment in which no coalition can profitably deviate, then comes into play). It does not say how this agreement comes about, nor what will transpire if there is no agreement. Indeed, in the latter case the concept of a Nash equilibrium has no particular claim upon us.

We are moved to ask, then: What other necessary conditions might be added to the condition that the agreement forms a Nash equilibrium? Some (but certainly not all) answers are given in section IV. What does transpire if no agreement arises? We will not touch on this question here, except to send the reader to recent work by Bernheim (1984) and Pearce (1984). And how might an agreement arise? This we take up next.

Reaching an 'Agreement'

One means to an agreement on how to play the game might be explicit negotiation among the players, conducted prior to play of the game. (If this happens, it may be important that negotiations take place before any player possesses private information, as such information might become revealed during the course of the negotiations.) We cannot guarantee that the players will come to an agreement, nor can we say what agreement will be reached. But, if the agreement is to be self-enforcing as above, it must be an equilibrium. That is, the range of possible self-enforcing agreements, arrived at via preplay negotiation, is contained within the set of Nash equilibria.

Any story about preplay negotiation contains within it an opportunity to choose among Nash equilibria, depending on the mechanism one imagines for the preplay negotiation. For example, if we imagine that exactly one player is allowed to make a speech, after which play occurs, then it is natural to suppose that the player, if he proposes an equilibrium at all, would propose one that is advantageous to him (see Farrell 1985). How the type of preplay negotiation affects the nature of any agreement that is reached is a relatively unexplored topic. (We return to preplay

Nash Equilibrium, Table 1

	Left	Right
Top	1, 0	5, 5
Bottom	2, 2	0, 1

Nash Equilibrium, Table 2

	Column 1	Column 2	Column 3	Column 4
Row 1	20, 5	0, 4	1, 3	2, -10^4
Row 2	0, -10^4	1, -10^3	3, 3	5, 10

negotiation later, in our discussion of correlated equilibria.)

But what if there is no explicit, preplay negotiation? Even then, *in some contexts, for some particular games*, player may *know* what each will do (at least, with high probability). A very simple example is the two player bimatrix game in Table 1: The two players are called Row and Col, and each is asked, simultaneously and without consultation, to make a choice: Row must choose either the top row or the bottom, and Col must choose either the left column or the right. Given these choices, payoffs are as in the chosen cell with Row's payoff listed first; so, for example, in Table 1, if the choices are Top and Left, then Row gets 1 and Col gets 0. For the game in Table 1, players usually have very little problem deciding what to do: Row chooses Top, and Col chooses Right. note that this is a Nash equilibrium. But Bottom and Left is another. Being Nash is only necessary, and not sufficient.

Another bimatrix game illustrates the point that such implicit agreements do not always arise. Consider the game in Table 2, where Row picks between rows 1 and 2, and Col selects one of four columns. This game possesses three Nash equilibria, two in pure strategies and one in mixed strategies, and in none of the three equilibria is column 3 played with positive probability. None the less, in the majority of cases (in informal experiments with students, with payoffs in units such as nickels), Col selects column 3, and Row selects row 2. A nontrivial fraction of Row players pick row 1, enough so that column 3 is an optimal choice for Col. Because there is no clear 'agreement', Col may well optimize by choosing a column that appears in *no* equilibrium.

The game in Table 1 seems too simple to be of consequence, but a similar phenomenon can be found in much more complex games. Consider the following game. There are two players, both American college students. A list of eleven cities in the United States is given: Atlanta, Boston, Chicago, Dallas, Denver, Kansas City, Los Angeles, New York, Philadelphia, Phoenix, San Francisco. Each city has been assigned an 'index' reflecting its importance to commerce, the arts, etc. All that the students know about this index is that New York is highest, with index 100, and Kansas City is lowest, with index 1. Each student is asked to choose, independently and without consultation, a subset of the cities, with one told that he must list Boston, and the other told that he must list San Francisco. (All these rules are common knowledge among the players.) After the two lists have been prepared, they are compared. If a city appears on one list and not the other, the student listing that city wins as many dollars as the city's index. If a city appears on both lists, each loses twice as many dollars as the city's index. And if the students manage to partition the eleven cities between them, their total winnings are tripled.

In pure strategies, this game has 512 Nash equilibria. Yet when played, students achieve a quite striking level of coordination. The Boston list nearly always contains New York, and Philadelphia, with Chicago less likely (but still very likely), and Atlanta a bit less still; the San Francisco list almost invariably includes Los Angeles, Phoenix and Denver, with Dallas a bit less likely, and Kansas City less likely still. (When there is contention, it nearly always involves Atlanta and/or Kansas City.) The reader will, or course, recognize what is going on here: Students focus very quickly on a division based on geographical principles. They do this without consultation – something in the game seems to focus attention in this manner.

This is an example of a *focal point* Nash equilibrium, as proposed and discussed by Schelling (1960). Schelling discusses a number of properties that focal points tend to possess (or, rather, that in some cases become the focus of the focal point): symmetry, qualitative uniqueness, equity. Beyond these vague generalities, it is clear that the context and presentation of the game matter. If instead of eleven cities we had eleven letters: A, B, C, D, E, K, L, N, P, Q and S, then the B list would contain A, C, D, E and (perhaps) K, while the S list would contain L, N, P, Q and (perhaps) K. (In simulation, K tends to go to the B list, presumably on grounds that players know that N has the highest index, and some sort of equity consideration intrudes.) The identities of the participants matter: if the cities game is played by two foreign students (each of whom knows that the other is foreign), there is increased use of the alphabetical rule. And experience matters: Roth and Schoumaker (1983) examine a bargaining game that admits two natural focal points; they show experimentally that players are conditioned by experience to key on one or the other.

The theory of focal points, while clearly quite important (both with regard to the use of Nash equilibrium and by itself), remains undeveloped. Until formal development occurs, the application of Nash equilibrium in many contexts relies for justification on a very vague idea.

The experimental work of Roth and Schoumaker suggests another explanation that is sometimes given for how agreements arise; namely through a dynamic process of adaptive expectations. Imagine a population of players engaged in a particular game over and over, learning after each round of play how opponents have played, and adapting subsequent choices to what has been learned. We might imagine that, in this process, there is convergence to some stationary equilibrium, which then would be a Nash equilibrium. But an imagination this vivid should be tempered: If the players are engaged with the same (or a small and recognizable set of) opponents over and over, then in the large (super-)game that they play, there are many more equilibria than in the single-shot game. Even if opponents change, players may carry with them reputations from past play, which will enlarge the set of equilibria. To nullify these effects, the players must face changing opponents, with no record of anyone's past play brought to bear. This is far from realistic; and still one must be careful concerning the amount of information that is passed after each round, if a 'dynamic stationary equilibrium' of such a process is to be a Nash equilibrium. With all these caveats, some study has been made of such dynamic processes, providing a further way in which 'agreements' might arise.

Finally, and again in the spirit of focal points, 'agreements' would arise if there were a single, unanimously adopted theory as to how games (or the game in question) are played. An example of such a theory is the tracing procedure of Harsanyi (1975).

Further Necessary Conditions: Perfection and Other Refinements

Consider the bimatrix game depicted in Table 3. There are two Nash equilibria in pure strategies here: Top-Left and Bottom-Right. Suppose that, somehow, Bottom-Right is agreed upon. (For example, imagine a process of pre-play negotiation in which only Col is allowed to speak, so that Col proposes the equilibrium that is most advantageous to him.) Would we consider this a selfenforcing agreement?

Note that Col, by picking Right, is picking a weakly dominated strategy. That is, no matter what Row does, Col. does as well with Left, and Col does strictly better if Row picks Top. Bottom-Right is a Nash equilibrium because Col does just as well with Right as with Left if Row can be trusted to play Bottom, but we might think that Col, entertaining the slightest doubts about whether Row will indeed stick to the agreement, would move to Left. If we think this, then Bottom-Right would not seem to be a self-enforcing agreement.

Consider next the following *extensive game* (hereafter called game A). Here one player, named Row, begins the game by choosing one of two actions, called *T* and *B*. If Row chooses

Nash Equilibrium, Table 3

	Left	Right
Top	2, 1	0, 0
Bottom	1, 2	1, 2

B, then the game is over, with Row receiving 1 and a second player, Col, receiving 2. If Row chooses *T*, then Col must select between two actions, called *L* and *R*. The choices of *T* and *L* net 2 for Row and 1 for Col, while the choices of *T* and *R* net 0 for each. Now if Row does choose *T*, then Col, it seems, would pick *L;* it is better to get 1 than 0. And if Col is going to pick *L*, then Row prefers *T* to *B*. Indeed, *T, L* is a Nash equilibrium for game A. But *B,R* is another Nash equilibrium. (Note that, although the choice of *B* by Row moots any choice by Col, we specify a choice, in this case *R*, so that Row can evaluate what will happen if he should choose *T* instead.) If Row thinks that Col will choose *R*, then Row responds with *B*. And if Row is to choose *B*, then Col's choice of *R* costs Col nothing. This second equilibrium, however, does not seem to be a self-enforcing agreement: If Row does choose *T*, then Col is put on the spot; will he really choose *R*, faced with the *fait accompli* of *T?*

The connection between these two games should be clear: Table 3 gives the normal form representation of game A. In each case, for (perhaps) slightly different reasons, we see that there can be Nash equilibria that do not seem viable candidates for self-enforcing agreements. These examples raise the general question: What further formal necessary criteria can be stated for selfenforcing agreements?

Game A is a finite game of complete and perfect information: there are finitely many moves and countermoves, and a player who is moving always knows what has transpired previously. It seems obvious how to solve (and play) games of this sort. Beginning at the end of the game tree, one finds how the last player to move will move. Then one can move back one step, and find the move of the penultimate player, and so on, using backwards induction to derive the solution. Going back to Kuhn (1953) (and perhaps earlier), it has been known that this procedure generates a Nash equilibrium. (And if there are never any ties at any stage of the backwards induction, it will generate a unique solution.) Correspondingly, in the normal form one sometimes comes across games that are *dominance solvable* – where the iterated elimination of dominated (weak or strict) strategies leads one to a single strategy combination. When such criteria apply, it seems sensible to use them. (Although in some applications the application of these criteria does lead to counter-intuitive results: see Selten (1978) and the literature that follows on the chain-store paradox.)

The intuition applied in game A can be generalized beyond the class of finite games with complete and perfect information. Beginning with the seminal work of Selten (1965, 1975), several authors have refined or 'perfected' the concept of a Nash equilibrium, to capture further necessary conditions for self-enforcing agreements. The first of these refinements is Selten's (1965) notion of subgame perfection: If at any point in an extensive game, all players agree as to what has transpired, then 'what remains' is, by itself, an extensive game. We might require that, in such circumstances, players expect that the agreement for this subgame constitutes a Nash equilibrium for the subgame. This applies to game A and, generally, to all finite games with complete and perfect information. But it applies fruitfully as well to games that are not finite (e.g. Rubinstein 1982) or that do not have complete and perfect information. Selten (1975) proposes further conditions called *perfection* (or, sometimes in the literature, trembling hand perfection). This is somewhat harder to describe, but the basic idea is that each player's strategy should be a best response to the others' strategies, where the first player does not rule out the possibility that his opponents might (with very small probability) fail to keep to the agreement. So, for example, in Table 3, Col, fearing that Row might play Top 'by mistake' as it were, will select Left.

Following these ideas, a number of alternative refinements (both stronger and weaker) have been proposed. Three are mentioned here (with apologies to those omitted): Myerson (1978) strengthens perfection to what is called *properness*, where (roughly) it is assumed that the chances of a 'mistake' made by some player are related to how severe that mistake is. Kreps and Wilson (1982) propose a weaker (than perfection) criterion for extensive games called sequential equilibrium: The basic idea is that behaviour in all parts of a game tree should be rationalized by some beliefs as to the play of the game that are not contradicted by what the player knows for sure. This bites wherever subgame perfection does; in game A, Col, asked to move, can no longer believe that Row will choose *B;* the fact that he is asked to move contradicts this. So his choice must be made optimally given the beliefs that, in this case, he must hold, once he is asked to move. But the notion is stronger than subgame perfection; indeed, it is 'almost equivalent' to perfection. Finally, Kohlberg and Mertens (1982), noting that the other criteria fail in certain applications and fail to possess natural properties such as invariance to alternative extensive form representations of the same normal form game, propose *stability*, a set-valued concept, which captures a number of very intuitive restrictions.

At the time of writing this entry, work on refinements is an active and ongoing subject. This brief description is probably outdated as it is written, and it will surely be outdated by the time it is read. Still, the programme of this work should be clear: Nash equilibrium gives a necessary condition for 'self-enforcing agreements' that is far from sufficient; there is much room for further formal criteria against which candidate agreements can be measured.

Games with Incomplete Information

In a Nash equilibrium, it is (essentially) presumed that players are all aware of the strategies their opponents are selecting. This presumption would seem especially incredible in cases where some players initially possess knowledge that other players lack, concerning their own tastes, abilities, and even the rules of the game. Imagine, for example, that Row and Col are playing the game A, but that Row is not certain what Col's payoffs are. In particular, Row entertains the possibility that Col might well prefer *R* to *L* if faced with the choice by Row of *T.* This is not so fanciful as it may seem; it might, perhaps, represent situations where Row is uncertain to what extent Col derives 'psychological utility' from seeing Row hurt. In economic applications, the uncertainty (if Row and Col are firms) might reflect one firm's initial uncertainty about the financial or human capital resources of its rivals, and so on. To apply Nash equilibrium analysis (and game theory generally) to such situations, therefore, seems a witless exercise.

There is, however, a standard technique to deal with such situations. This involves what is called a *game with incomplete information*, as developed by John Harsanyi (1967–8). The concept is subtle, but a brief description can be given. We imagine that the differences in players' initial information can be traced to a two-step preplay procedure. At the start, every player is on an equal informational footing. There is initial uncertainty as to what rules of the game, etc., will prevail when the game is played, and players have their prior assessments as to how that uncertainty will resolve. (It is almost always assumed that these prior assessments are identical; indeed, this assumption is held by many to be the only philosophically sensible assumption to make, and it is called the *Harsanyi doctrine* in many places.) Nature resolves this uncertainty and *selectively* reveals to the players part of that resolution. That is, one player may learn (in this initial round of revelation) things not revealed to another. *Then* the game begins; the 'initial' differences in what players know about the rules of the game trace to differences in what the players were told by nature before the game 'begins'. So, for example, to model Row's uncertainty about Col's payoffs in game A, we imagine: There are several possible games that the players might play, distinguished by Col's payoff structure. There is an initial probability distribution over what Col's payoffs will be. Nature picks a payoff structure for Col, and nature reveals to Col *but not to Row* what that structure is. Hence the game begins with Row uncertain about Col's payoffs.

In this model, Col is aware of the nature of Row's uncertainty. And, in doing Nash equilibrium analysis, Col will (if he can) take advantage of that uncertainty. In a Nash equilibrium, we specify the players' choices of actions, as before, for the particular 'rules' that nature has indeed selected. But we *also* specify how players would have acted had nature chosen (and informed them) differently. This is necessary because when one player is uncertain about part of nature's choice, it is important what his fellow players would have done had nature chosen differently.

The example we have given is too simple to see the full power of this construction, but the reader need not go far into the literature to find examples. This technique has been applied in many instances, to extend the reach of Nash (and game theoretic) analysis. Applied skillfully, it can be used to model all sorts of situations, and while (in order to retain tractability) one must be content with highly stylized models, qualitative insights that have considerable intuitive appeal have been derived.

Correlated Equilibrium

One of the stories told to justify Nash equilibria holds that players meet prior to play, and they (perhaps) negotiate a self-enforcing agreement. It turns out that, in some cases, by being clever, players can do better than they can with any Nash equilibrium.

Consider the bimatrix game in Table 4, taken from Aumann (1985). There are three Nash equilibria here, Bottom-Left, Top-Right, and a mixed strategy equilibrium in which each player has an expected payoff of 14/3.

Nash Equilibrium, Table 4

	Left	Right
Top	6, 6	2, 7
Bottom	7, 2	0, 0

Now imagine that, in preplay negotiation, one player suggests to the other that they hire a referee to perform the following steps. The referee will roll a six-sided die. If he die comes up with one or two dots on top, the referee will privately instruct Row to pick Top and Col to pick Left. For three or four dots, the instructions will be Top to Row and Right to Col. For five or six the instructions will be Bottom to Row and Left to Col. And, what is crucial, the instructions to each will not include what is being told to the other side; each player is told by the referee *only* what that player should do.

Are these instructions self-enforcing, in the sense that each player, assuming the other will carry out his instructions, would do so as well? Consider Row. If told to play Bottom, Row knows that the die came up with five or six up, and so Col must have been told to play Left. Thus Bottom is indeed Row's best choice. If told to play Top, Row only knows that the die came up with between one and four spots. Hence Col may have been told to play Right, and may have been told Left, each with probability 1/2. But if Row assesses that Col is choosing between Left and Right, each with probability 1/2, then Top is indeed better than Bottom. Symmetric reasoning shows that this arrangement is self-enforcing on Col.

With these instructions, the vector expected payoff to the players is (5, 5), which lies outside the set of Nash equilibria; indeed, it lies outside the convex hull of the set of Nash payoffs. Apparently (the convex hull of) the set of Nash equilibrium *is not* the entire set of potential self-enforcing agreements to the game, at least, if the players can hire and instruct referees that act to *correlate* the actions of the players.

The last sentence is the key. In a Nash equilibrium, the players are presumed to select their strategies independently of one another. Through the intervention of a referee, they can achieve correlation in their choices. This is the basic insight of Aumann (1973). It has been extended by Forges (1986) and Myerson (1984), who note that the possibilities for correlation may expand still further if the referee can send messages during the course of an extensive game, and further still if players can, during the course of play, communicate privately to the referee information that they possess or will come to possess.

The set of correlated equilibria, unlike the set of Nash equilibria, has a very simple mathematical structure; it is a convex polyhedron, which is easy to compute, using simple mathematical programming techniques. (Computing Nash equilibria is much more difficult.) Perhaps most importantly, Aumann (1987) establishes a beautiful linkage between correlated equilibria, a particular class of games with incomplete information, and 'the common knowledge of Bayesian rationality by all players'.

See Also

- ▶ Bargaining
- ▶ Bidding
- ▶ Exchange
- ▶ Game Theory
- ▶ Repeated Games

Bibliography

Aumann, R. 1973. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics* 1: 67–96.

Aumann, R. 1987. Correlated equilibrium as an expression of Bayesian rationality. *Econometrica* 55: 1–18.

Bernheim, D. 1984. Rationalizable strategic behavior. *Econometrica* 52: 1007–1028.

Cournot, A. 1838. *Recherches sur les principes mathématiques de la théorie des richesses*. Paris. Trans. as *Researches into the mathematical principles of the theory of wealth*. New York: Macmillan and Company, 1897.

Farrell, J. 1985. *Communication equilibria in games*. Waltham: GTE Laboratories.

Forges, F. 1986. An approach to communication equilibrium. *Econometrica* 54(6): 1375–1385.

Harsanyi, J. 1967–8. Games with incomplete information played by Bayesian players. Parts I, II, and III. *Management Science* 14: 159–82, 320–334. 486–502.

Harsanyi, J. 1975. The tracing procedure. *International Journal of Game Theory* 4: 61–94.

Kohlberg, E., and J.-F. Mertens. 1982. On the strategic stability of equilibrium. Working paper, CORE, Catholic University of Louvain, forthcoming in *Econometrica*.

Kreps, D., and R. Wilson. 1982. Sequential equilibrium. *Econometrica* 50: 863–894.

Kuhn, H. 1953. Extensive games and the problem of information. In *Contributions to the theory of games*, vol. 2, ed. H. Kuhn and A. Tucker. Princeton: Princeton University Press.

Luce, D.R., and H. Raiffa. 1957. *Games and decisions*. New York: Wiley.

Myerson, R. 1978. Refinements of the Nash equilibrium concept. *International Journal of Game Theory* 7: 73–80.

Myerson, R. 1984. Sequential equilibria of multistage games. DMSEMS discussion paper no. 590, Northwestern University.

Nash, J.F. 1950. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences USA* 36: 48–49.

Nash, J.F. 1951. Non-cooperative games. *Annals of Mathematics* 54: 286–295.

Pearce, D. 1984. Rationalizable strategic behavior and the problem of perfection. *Econometrica* 52: 1029–1050.

Roth, A., and F. Schoumaker. 1983. Expectations and reputations in bargaining: An experimental study. *American Economic Review* 73: 362–372.

Rubinstein, A. 1982. Perfect equilibrium in a bargaining model. *Econometrica* 50: 97–109.

Schelling, T. 1960. *The strategy of conflict*. Cambridge, MA: Harvard University Press.

Selten, R. 1965. Spieltheoretische Behandlung eines Oligopolmodells mit Nachfragetragheit. *Zeitschrift für die gesamte Staatswissenschaft* 121: 301–324.

Selten, R. 1975. Re-examination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory* 4: 25–55.

Selten, R. 1978. The chain-store paradox. *Theory and Decision* 9: 127–159.

Nash Equilibrium, Refinements of

Srihari Govindan and Robert B. Wilson

Abstract

This article describes ways that the definition of an equilibrium among players' strategies in a game can be sharpened by invoking additional criteria derived from decision theory. Refinements of John Nash's 1950 definition aim primarily to distinguish equilibria in which implicit commitments are credible due to incentives. One group of refinements requires sequential rationality as the game progresses. Another ensures credibility by considering perturbed games in which every contingency occurs with positive probability, which has the further advantage of excluding weakly dominated strategies.

Keywords

Backward induction; Bayes' rule; Dynamic games; Game theory; Harsanyi, J.; Imperfect information; Nash equilibrium; Nash, J.; Noncooperative games; Perfect information; Refinements of Nash equilibrium; Sequential rationality; Small-worlds criterion; Subgame perfection; Symmetric games

JEL Classifications

C7

Game theory studies decisions by several persons in situations with significant interactions. Two features distinguish it from other theories of multi-person decisions. One is explicit consideration of each person's available strategies and the outcomes resulting from combinations of their choices; that is, a complete and detailed specification of the 'game'. Here a person's strategy is a complete plan specifying his action in each contingency that might arise. In non-cooperative contexts, the other is a focus on optimal choices by each person separately. John Nash (1950, 1951) proposed that a combination of mutually optimal strategies can be characterized mathematically as an *equilibrium.* According to Nash's definition, a combination is an equilibrium if each person's choice is an optimal response to others' choices. His definition assumes that a choice is optimal if it maximizes the person's expected utility of outcomes, conditional on knowing or correctly anticipating the choices of others. In some applications, knowledge of others' choices might stem from prior agreement or communication, or accurate prediction of others' choices might derive from 'common knowledge' of strategies and outcomes and of optimizing behaviour. Because many games have multiple equilibria, the predictions obtained are incomplete. However, equilibrium is a weak criterion in some respects, and therefore one can refine the criterion to obtain sharper predictions (Harsanyi and Selten 1988; Hillas and Kohlberg 2002; Kohlberg 1990; Kreps 1990).

Here we describe the main refinements of Nash equilibrium used in the social sciences. Refinements were developed incrementally, often relying on ad hoc criteria, which makes it difficult for a non-specialist to appreciate what has been accomplished. Many refinements have been proposed but we describe only the most prominent ones. First we describe briefly those refinements that select equilibria with simple features, and then we focus mainly on those that invoke basic principles adapted from single-person decision theory.

Equilibria with Simple Features

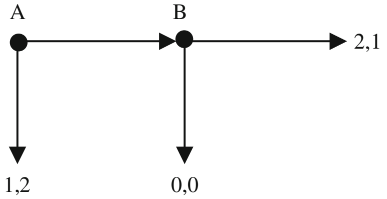
Nash's construction allows each person to choose randomly among his strategies. But randomization is not always plausible, so in practice there is a natural focus on equilibria in 'pure' strategies, those that do not use randomization. There is a similar focus on strict equilibria, those for which each person has a unique optimal strategy in response to others' strategies. In games with some symmetries among the players, the symmetric equilibria are those that reflect these symmetries. In applications to dynamic interactions the most useful equilibria are those that, at each stage, depend only on that portion of prior history that is relevant for outcomes in the future. In particular, when the dynamics of the game are stationary one selects equilibria that are stationary or that are Markovian in that they depend only on state variables that summarize the history relevant for the future. Applications to computer science select equilibria or, more often, approximate equilibria, using strategies that can be implemented by simple algorithms. Particularly useful are equilibria that rely only on limited recall of past events and actions and thus economize on memory or computation.

Refinements That Require Strategies to Be Admissible

One strategy is strictly dominated by another if it yields strictly inferior outcomes for that person regardless of others' choices. Because an equilibrium never uses a strictly dominated strategy, the same equilibria persist when strictly dominated strategies are deleted, but after deletion it can be that some remaining strategies become strictly dominated. A refinement that exploits this feature deletes strictly dominated strategies until none remain, and then selects those equilibria that remain in the reduced game. If a single equilibrium survives then the game is called 'dominance solvable' . An equilibrium can, however, use a strategy that is weakly dominated in that it would be strictly dominated were it not for ties – in decision theory such a strategy is said to be inadmissible. A prominent criterion selects equilibria that use only admissible strategies, and sometimes this is strengthened by iterative deletion of strictly dominated strategies after deleting the inadmissible strategies. A stronger refinement uses *iterative deletion of* (both strictly and weakly) *dominated strategies* until none remain; however, this procedure is ambiguous because the end result can depend on the order in which weakly dominated strategies are deleted.

A particular order is used for dynamic games that decompose into a succession of subgames as time progresses. In this case, those strategies that are weakly dominated because they are strictly dominated in final subgames are deleted first, then those in penultimate subgames, and so on. In games with 'perfect information' as defined below this procedure implements the criterion called 'backward induction' and the equilibria that survive are among those that are 'subgame-perfect' (Selten 1965). In general a subgame-perfect equilibrium is one that induces an equilibrium in each subgame. Fig. 1 depicts an example in which there are two Nash equilibria, one in which A moves down because she anticipates that B will move down, and a second that is subgame-perfect because in the subgame after A moves across, B also moves across, which yields him a higher payoff than down.

The informal criterion of 'forward induction' has several formulations. Kohlberg and Mertens (1986) require that a refined set of equilibria contains a subset that survives deletion of strategies that are not optimal responses at any equilibrium in the set. Van Damme (1989, 1991) requires that if player A rejects a choice X in favour of Y or Z then another player who knows only that Y or Z was chosen should consider Z unlikely if it is chosen only in equilibria that yield player A outcomes worse than choosing X, whereas Y is chosen in an equilibrium whose outcome is better. A typical application mimics backward induction but in reverse – if a person previously rejected a choice with an outcome that would have been superior to the outcomes from all but one equilibrium of the ensuing subgame, then presumably the person is anticipating that favourable equilibrium and intends to use his strategy in that equilibrium of the subgame. In Fig. 2, if A rejects the payoff 5 from Down then B can infer that A intends to play Top in the ensuing subgame, yielding payoff 6 for both players.

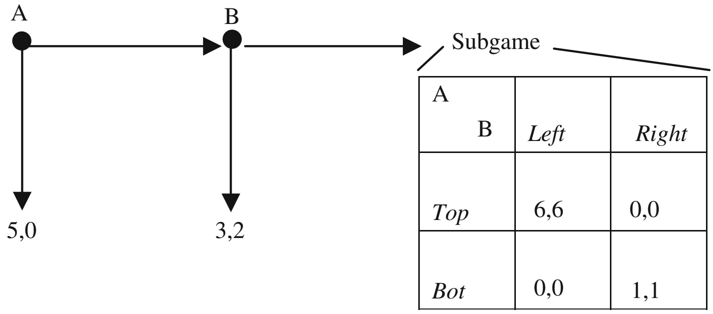


Nash Equilibrium, Refinements of, Fig. 1 Player A moves down or across, in which case player B moves down or across. Payoffs for A and B are shown at the end of each sequence of moves

Dynamic Games

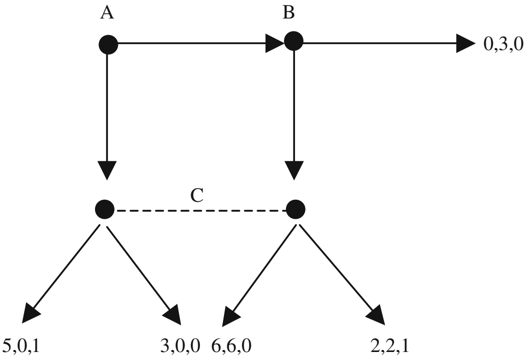
Before proceeding further we describe briefly some relevant features of dynamic games, that is, games in which a player acts repeatedly, and can draw inferences about others' strategies, preferences, or private information as the game progresses. A dynamic game is said to have 'perfect information' if each person knows initially all the data of the game, and the prior history of his and others' actions whenever he acts, and they do not act simultaneously. In such a game each action initiates a subgame; hence backward induction yields a unique subgame-perfect equilibrium if there are no ties. But in many dynamic games there are no subgames. This is so whenever some person acts without knowing all data of the game relevant for the future. In Fig. 3 player C acts without knowing whether player A or B chose down.

The source of this deficiency is typically that some participant has private information – for example, about his own preferences or about outcomes – or because his actions are observed imperfectly by some others. Among parlour games, chess is a game with perfect information (if players remember whether each king has been



A B	*Left*	*Right*
Top	6,6	0,0
Bot	0,0	1,1

Nash Equilibrium, Refinements of, Fig. 2 First A and then B can avoid playing the subgame in which simultaneously each chooses between two options



Nash Equilibrium, Refinements of, Fig. 3 Player A moves down or across, in which case player B moves down or across. Player C does not observe whether it was A or B who moved down when she chooses to move left or right

castled). Bridge and poker are games with imperfect information because the cards in one player's hand are not known to others when they bet. In practical settings, auctions and negotiations resemble poker because each party acts (bids, offers, and so on) without knowing others' valuations of the transaction. Analyses of practical economic games usually assume (as we do here) 'perfect recall' in the sense that each player always remembers what he knew and did previously. If bridge is treated as a two-player game between teams, then it has imperfect recall because each team alternately remembers and forgets the cards in one member's hand as the bidding goes round the table, but bridge has perfect recall if it is treated as a four-player game. In card games like bridge and poker each player can derive the probability distribution of others' cards from the assumption that the deck of cards was thoroughly shuffled. Models of economic games impose analogous assumptions; for example, a model of an auction assumes that each bidder initially assesses a probability distribution of others' valuations of the item for sale, and then updates this assessment as he observes their bids. More realism is obtained from more complicated scenarios; for example, it could be that player A is uncertain about player B's assessment of player A's valuation. In principle the model could allow a hierarchy of beliefs – A's probability assessment of B's assessment of A's assessment of . . .. To adopt a proposal by John Harsanyi (1967–1968) developed by Mertens and Zamir (1985), such situations are modelled by assuming that each player is one of several types. The initial joint distribution of types is commonly known among the players, but each player knows his own type, which includes a specification of his available strategies, his preferences over outcomes, and, most importantly, his assessment of the conditional probabilities of others' types given his own type. In poker, for instance, a player's type includes the hand of cards he is dealt, and his hand affects his beliefs about others' hands.

Refinements of Nash equilibrium are especially useful in dynamic games. Nash equilibria do not distinguish between the case in which each player commits initially and irrevocably to his strategy throughout the game, and the case in which a player continually re-optimizes as the game progresses. The distinction is lost because the definition of Nash equilibrium presumes that players will surely adhere to their strategies chosen initially. Most refinements of Nash equilibrium are intended to resurrect this important distinction. Ideally one would like each Nash equilibrium to bear a label telling whether it assumes implicit commitment or relies on incredible threats or promises. Such features are usually evident in the equilibria of trivially simple games, but in more complicated games they must be identified by augmenting the definition of Nash equilibrium with additional criteria.

In the sequel we describe two classes of refinements in detail, but first we summarize their main features, identify the main selection criteria they use, and mention the names of some specific refinements. Both classes are generalizations of backward induction and subgame perfection, and they obtain similar results, but their motivation and implementation differ.

The Criterion of Sequential Rationality

The presumption that commitment is irrevocable is flawed if other participants in the game do not view commitment to a strategy as credible. Commitment can be advantageous, of course, but if commitment is possible (for example, via enforceable contractual arrangements) then it should properly be treated as a distinct strategy. Absent commitment, some Nash equilibria are suspect because they rely implicitly on promises or threats that are not credible. For example, one Nash equilibrium might enable an incumbent firm to deter another firm from entering its market by threatening a price war. If such a threat succeeds in deterring entry then it is costless to the incumbent because it is never challenged; indeed, it can be that this equilibrium is sustained only by the presumption that the incumbent will never need to carry out the threat. But this threat is not credible if, after entry occurs, the incumbent would recognize that accommodation is more profitable than a price war. In such contexts, the purpose of a refinement is to select an alternative Nash equilibrium that anticipates correctly that entry will be followed by accommodation. For instance, the subgame-perfect equilibrium in Fig. 1 satisfies this criterion.

Refinements in the first class exclude strategies that are not credible by requiring explicitly that a strategy is optimal in each contingency, even if it comes as a surprise. (We use the term ‘contingency’ rather than the technical term ‘information set’ used in game theory – it refers to any situation in which the player chooses an action.) These generally require that a player’s strategy is optimal initially (as in the case of commitment), *and* that in each subsequent contingency in which the player might act his strategy remains optimal for the remainder of the game, even if the equilibrium predicts that the contingency should not occur. This criterion is called ‘ sequential rationality’. As described later, three such refinements are *perfect Bayes, sequential,* and *lexicographic* equilibria, each of which can be strengthened further by imposing additional criteria such as *invariance,* the *intuitive criterion* and *divinity.*

The Criterion of Perfection or Stability

The presumption that commitment is irrevocable is also flawed if there is some chance of deviations. If a player might ‘tremble’ or err in carrying out his intended strategy, or his valuation of outcomes might be slightly different from others anticipated, then other players can be surprised to find themselves in unexpected situations. Refinements that exploit this feature are implemented in two stages. In the first stage one identifies the Nash equilibria of a perturbation of the original game, usually obtained by restricting each player to randomized strategies that assign positive probabilities to all his original pure strategies. In the second stage one identifies those equilibria of the original game that are limits of equilibria of the perturbed game as this restriction is relaxed to allow inferior strategies to have zero probabilities.

Refinements in the second class also exclude strategies that are not credible, but refinements in this class implement sequential rationality indirectly. The general criterion that is invoked is called ‘perfection’ or ‘stability’, depending on the context. In each case a refinement is obtained from analyses of perturbed games. This second class of refinements is typically more restrictive than the first class due to the stronger effects of perturbations. As described later, two such refinements are *perfect* and *proper* equilibria. These are equilibria that are perturbed slightly by *some* perturbation of the players’ strategies. A more stringent refinement selects a subset of equilibria that is *truly perfect* or *stable* in the sense that it is perturbed only slightly by *every* perturbation of players' strategies. This refinement selects a subset of equilibria rather than a single equilibrium because there need not exist a single equilibrium that is *essential* in that it is perturbed slightly by every perturbation of strategies. A stringent refinement selects a subset that is *hyperstable* in that it is stable against perturbations of both players' strategies and their valuations of outcomes, or against perturbations of their optimal responses; and further, it is *invariant* in that it is unaffected by addition or deletion of redundant strategies.

The crucial role of perturbations in the second class of refinements makes them more difficult for non-specialists to understand and appreciate, but they have a prominent role in game theory because of their desirable properties. For example, in a two-player game a perfect equilibrium is equivalent to an equilibrium that uses only admissible strategies. In general, refinements in the second class have the advantage that they satisfy several selection criteria simultaneously.

After this overview, we now turn to detailed descriptions of the various refinements.

Refinements That Require Sequential Rationality

In dynamic games with perfect information, the implementation of backward induction is unambiguous because in each contingency the player taking an action there knows exactly the subgame that follows. In chess, for example, the current positions of the pieces determine how the game can evolve subsequently. Moreover, if he anticipates his opponent's strategy then he can predict how the opponent will respond to each possible continuation of his own strategy. Using this prediction he can choose an optimal strategy for the remainder of the game by applying the *principle of optimality* – his optimal strategy in the current subgame consists of his initial action that, when followed by his optimal strategies in subsequent subgames, yields his best outcome. Thus, in principle (although not in practice, since chess is too complicated) his optimal strategy can be found by working backward from final positions through all possible positions in the game.

In contrast, in a game with imperfect information a player's current information may be insufficient to identify the prior history that led to this situation, and therefore insufficient to identify how others will respond in the future, even if he anticipates their strategies. In poker, for example, knowledge of his own cards and anticipation of others' strategies are insufficient to predict how they will respond to his bets. Their strategies specify how they will respond conditional on their cards but, since he does not know their cards, he remains uncertain what bets they will make in response to his bets. In this case, it is his assessment of the probability distribution of their cards that enables construction of his optimal strategy. That is, this probability distribution can be combined with their strategies to provide him with a probabilistic prediction of how they will bet in response to each bet he might make. Using this prediction he can again apply the principle of optimality to construct an optimal strategy by working backward from the various possible conclusions of the game.

Those refinements that select equilibria satisfying sequential rationality use an analogous procedure. The analogue of the probability distribution of others' cards is a system of 'beliefs', one for each contingency in which the player might find himself. Each belief is a conditional probability distribution on the prior history of the game given the contingency at which he has arrived. Thus, to whatever extent he is currently uncertain about others' preferences over final outcomes or their prior actions, his current belief provides him with a probability distribution over the various possibilities. As in poker, this probability distribution can be combined with his anticipation of their strategies to provide him with a probabilistic prediction of how they will act in response to each action he might take – and again, using this prediction he can apply the principle of optimality to construct an optimal strategy by working backward from the various possible conclusions of the game.

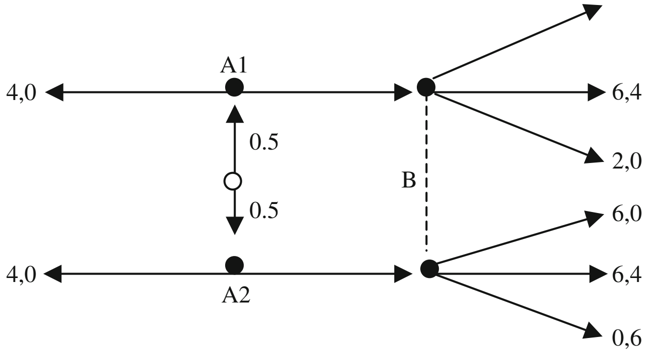
There is an important proviso, however. These refinements require that, whenever one contingency follows another with positive probability, the belief at the later one must be obtained from the belief at the earlier one by Bayes' rule. This ensures consistency with the rules of conditional probability. But, importantly, it does not restrict a player's belief at a contingency that was unexpected, that is, had zero probability according to his previous belief and the other players' strategies.

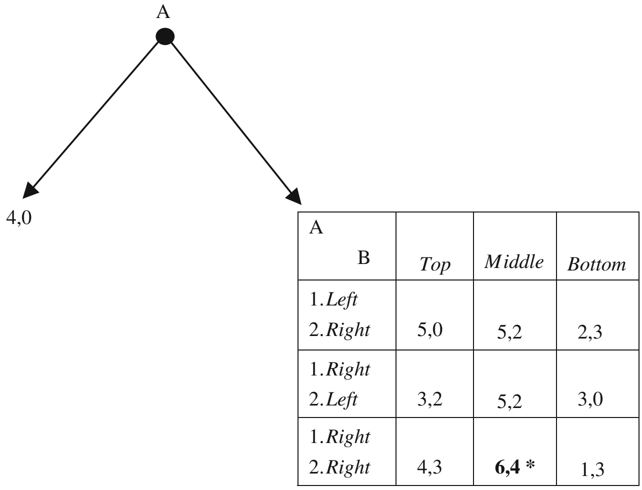
In Fig. 3, in one Nash equilibrium A chooses down, B chooses across, and C chooses left. This is evidently not sequential because if A were to deviate then B could gain by choosing down. In a sequential equilibrium B chooses down and each of A and C randomizes equally between his two strategies. The strategies of A and B imply that C places equal probabilities on which of A and B chose down.

The weakest refinement selects a *perfect-Bayes* equilibrium (Fudenberg and Tirole 1991). This requires that each player's strategy is consistent with some system of beliefs such that (*a*) his strategy is optimal given his beliefs and others' strategies, and (*b*) his beliefs satisfy Bayes' rule (wherever it applies) given others' strategies. A stronger refinement selects *sequential* equilibria (Kreps and Wilson 1982). A sequential equilibrium requires that each player's system of beliefs is consistent with the structure of the game. Consistency is defined formally as the requirement that each player's system of beliefs is the limit of the conditional probabilities induced by players' strategies in some perturbed game, as described previously. A further refinement selects *quasi-perfect* equilibria (van Damme 1984), which requires admissibility of a player's strategy in continuation from each contingency, excluding any chance that he himself might deviate from his intended strategy. And even stronger are *proper* equilibria (Myerson 1978), described later. This sequence of progressively stronger refinements is typical. Because proper implies quasi-perfect implies sequential implies perfect-Bayes, one might think that it is sufficient to always use properness as the refinement. However, the prevailing practice in the social sciences is to invoke the weakest refinement that suffices for the game being studied. This reflects a conservative attitude about using unnecessarily restrictive refinements. If, say, there is a unique sequential equilibrium that uses only admissible strategies, then one refrains from imposing stronger criteria.

Additional criteria can be invoked to select among sequential equilibria. In Fig. 4 there is a sequential equilibrium in which both types of A move left and B randomizes equally between middle and bottom, and another in which both types of A move right and B chooses middle. An alternative justification for the second, due to Hillas (1998), is shown in Fig. 5, where the game is restructured so that A either commits initially to left or they play the subgame with simultaneous choices of strategies. The criterion of subgame perfection selects the second equilibrium in Fig. 4 because in Fig. 5 the subgame has a unique equilibrium with payoff 6 for A that is superior to his payoff 4 from committing to left.

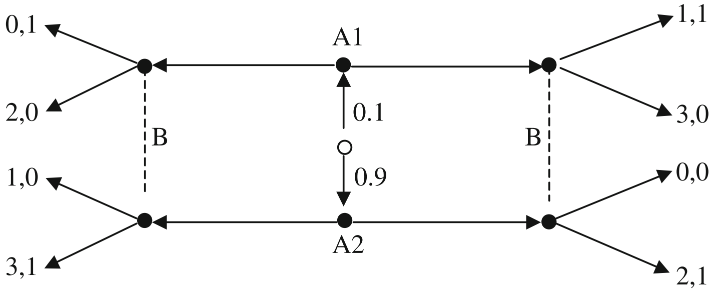
Nash Equilibrium, Refinements of, Fig. 4 Nature chooses whether player A's type is A1 or A2 with equal probabilities. Then A chooses Left or Right, in which case player B, without knowing A's type, chooses one of three options





A B	*Top*	*Middle*	*Bottom*
1.*Left* 2.*Right*	5,0	5,2	2,3
1.*Right* 2.*Left*	3,2	5,2	3,0
1.*Right* 2.*Right*	4,3	**6,4 ***	1,3

Nash Equilibrium, Refinements of, Fig. 5 The game in Fig. 4 restructured so that either A commits to Left regardless of his type, or plays a subgame with simultaneous moves in which he chooses one of his other three type-contingent strategies. The payoffs 6,4 to A and B from the unique Nash equilibrium of the subgame are shown with an asterisk



Nash Equilibrium, Refinements of, Fig. 6 A signalling game in which Nature chooses A's type A1 or A2, then A chooses left or right, and then B, without knowing A's type, chooses up or down

These refinements can be supplemented with additional criteria that restrict a player' s beliefs in unexpected contingencies. The most widely used criteria apply to contexts in which one player B could interpret the action of another player A as revealing private information; that is, A's action might signal something about A's type. These criteria restrict B's belief (after B observes A deviating from the equilibrium) to one that assigns positive probability only to A's types that might possibly gain from the deviation, provided it were interpreted by B as a credible signal about A's type. The purpose of these criteria is to exclude beliefs that are blind to A's attempts to signal what his type is when it would be to A's advantage for B to recognize the signal. In effect, these criteria reject equilibria that commit a player to unrealistic beliefs. Another interpretation is that these criteria reject equilibria in which A is 'threatened by B's beliefs' because B stubbornly retains these beliefs in spite of plausible evidence to the contrary.

The simplest version requires that B's belief assigns zero probability to those types of A that cannot possibly gain by deviating, regardless of how B responds. The *intuitive* criterion (Cho and Kreps 1987) requires that there cannot be some type of A that surely gains from deviating in every continuation for which B responds with a strategy that is optimal based on a belief that assigns zero probability to those types of A that cannot gain from the deviation. That is, an equilibrium fails the intuitive criterion if B's belief fails to recognize that A's deviation is a credible signal about his type. They apply this criterion to the game in Fig. 6, which has two sequential equilibria. In one both types of A choose left and B chooses down or up contingent on left or right. In another both types choose right and B chooses up or down contingent on left or right. In both equilibria B's belief in the unexpected event (right or left respectively) assigns probability greater than 0.5 to A's type A1. The intuitive criterion rejects the second equilibrium because if A2 were to deviate by choosing left, and then B recognizes that this deviation credibly signals A's type A2 (because type A1 cannot gain by deviating regardless of B's response) and therefore B chooses down, then type A2 obtains payoff 3 rather than his equilibrium payoff 2.

Cho and Kreps also define an alternative version, called the 'equilibrium domination' criterion. This criterion requires that, for each continuation in which B responds with a strategy that is optimal based on a belief that assigns zero probability to those types of A that cannot gain from deviating, there cannot be some type of A that gains from deviating. More restrictive is the criterion **D1** (Banks and Sobel 1987), also called 'divinity' when it is applied iteratively, which requires that, if the set of B's responses for which one type of A gains from deviating is larger than the set for which a second type gains, then B's beliefs must assign zero probability to the second type. The criterion **D2** is similar except that some (rather than just one) types of A gain. All these criteria are weaker than the *never weak best reply* criterion that requires an equilibrium to survive deletion of a player's strategy that is not an optimal reply to any equilibrium with the same outcome. In Fig. 6 this criterion is applied by observing that the second equilibrium does not survive deletion of those strategies of A in which type A2 chooses left.

The above criteria are all weak versions of forward induction. Govindan and Wilson (2009a, b) propose the following formal definition of forward induction for a game in extensive form with perfect recall. Say that an equilibrium is weakly sequential if it is sequential except that a player's strategy need not be optimal at an information set that the strategy excludes from being reached. A player's strategy is called relevant for an outcome of the game if there exists a weakly sequential equilibrium with that outcome for which the strategy is an optimal reply at every information set it does not exclude. The outcome satisfies forward induction if it results from a weakly sequential equilibrium in which players' beliefs assign positive probability only to relevant strategies at each information set reached by a profile of relevant strategies. They prove that if there are two players and payoffs are generic, then an outcome satisfies forward induction if every game with the same reduced normal form (obtained by eliminating redundant pure strategies) has a sequential equilibrium with an equivalent outcome. Thus in this case forward induction is implied by decision-theoretic criteria.

A *lexicographic* equilibrium (Blume et al. 1991a, b) uses a different construction. Each player is supposed to rely on a sequence of ' theories' about others' strategies. He starts the game by assuming that his first theory of others' strategies is true, and uses his optimal strategy according to that theory. He continues doing so until he finds himself in a situation that cannot be explained by his first theory. In this case, he abandons the first theory and assumes instead that the second theory is true – or if it too cannot explain what has happened then he proceeds to the next theory in the sequence. This provides a refinement of Nash equilibrium because each player anticipates that deviation from his optimal strategy for any theory will provoke others to abandon their current theories and strategies and thus respond with their optimal strategies for their next theories consistent with his deviant action. Lexicographic equilibria can be used to represent nearly any refinement. The hierarchy of a player's theories serves basically the same role as his system of beliefs, but the focus is on predictions of other players' strategies in the future rather than probabilities of what they know or have done in the past. The lexicographic specification has the same effect as considering small perturbations of strategies; for example, the sequence of strategies approximating a perfect or proper equilibrium can be used to construct the hierarchy of theories.

Refinements Derived from Perturbed Games

The other major class of refinements relies on perturbations to select among the Nash equilibria. The motive for this approach stems from a basic principle of decision theory – the *equivalence* of alternative methods of deriving optimal strategies.

This principle posits that constructing a player's optimal strategy in a dynamic game by invoking auxiliary systems of beliefs and the iterative application of the principle of optimality (as in perfect-Bayes and sequential equilibria) is a useful computational procedure, but the same result should be obtainable from an initial choice of a strategy, that is, an optimal plan for the entire game of actions taken in each contingency. Indeed, the definition of Nash equilibrium embodies this principle. Proponents therefore argue that whatever improvements come from dynamic analysis can and should be replicated by static analysis of initial choices among strategies, supplemented by additional criteria. (We use the terms 'static' and 'dynamic' analysis rather than the technical terms 'normal-form' and 'extensive-form' analysis used in game theory.) The validity of this argument is evident in the case of subgame-perfect equilibria of games with perfect information, which can be derived either from the principle of optimality using backward induction, or by iterative elimination of weakly dominated strategies in a prescribed order. The argument is reinforced by major deficiencies of dynamic analysis; for example, we mentioned above that a sequential equilibrium can use inadmissible strategies. Another deficiency is failure to satisfy the criterion of *invariance,* namely, the set of sequential equilibria can depend on which of many equivalent descriptions of the dynamics of the game is used (in particular, on the addition or deletion of redundant strategies).

On this view one should address directly the basic motive for refinement, which is to exclude equilibria that assume implicitly that each player commits initially to his strategy – since Nash equilibria do not distinguish between cases with and without commitment. Thus one considers explicitly that during the game any player might deviate from his equilibrium strategy for some exogenous reason that was not represented in the initial description of the game. Recognition of the possibility of deviations, however improbable they might be, then ensures that a player's strategy includes a specification of his optimal response to others' deviations from the equilibrium. The objective is therefore to characterize those equilibria that are affected only slightly by small probabilities of deviant behaviours or variations in preferences. This programme is implemented by considering perturbations of the game. These can be perturbations of strategies or payoffs, but actually the net effect of a perturbation of others' strategies is to perturb a player's payoffs.

In the following we focus on the perturbations of the static (that is, the normal form) of the game but similar perturbations can also be applied to the dynamic version (that is, the extensive form) by applying them to each contingency separately. This is done by invoking the principle that a dynamic game can also be analysed in a static framework by treating the player acting in each contingency as a new player (interpreted as the player's agent who acts solely in that contingency) in the ' agent-normal-form' of the game, where the new player's payoffs agree with those of the original player.

The construction of a *perfect* equilibrium (Selten 1975) illustrates the basic method, which uses two steps.

1. For each small positive number ε one finds an *ε-perfect* equilibrium, defined by the requirement that each player's strategy has the following property: every one of his pure strategies is used with positive probability, but any pure strategy that is an inferior response to the others' strategies has probability no more than ε. Thus an ε-perfect equilibrium supposes that every strategy, and therefore every action during the game, might occur, even if it is suboptimal.
2. One then obtains a perfect equilibrium as the limit of a convergent subsequence of ε-perfect equilibria.

One method of constructing an ε-perfect equilibrium starts by specifying for each player i a small probability $\delta_i < \varepsilon$ and a randomized strategy σ_i that uses every pure strategy with positive probability – that is, the strategy combination σ is 'completely mixed'. One then finds an ordinary Nash equilibrium of the perturbed game in which each player's payoffs are as follows: his payoff from each combination of all players' pure strategies is replaced by his expected payoff when each player i's pure strategy is implemented only with probability 1 – δ_i and with probability δ_i that player uses his randomized strategy σ_i instead. In this context one says that the game is perturbed by less than e toward σ – we use this phrase again later when we describe stable sets of equilibria. An equilibrium of this perturbed game induces an ε-perfect equilibrium of the original game.

An alternative definition of perfect equilibrium requires that each player's strategy is an optimal response to a convergent sequence of others' strategies for which all their pure strategies have positive probability – this reveals explicitly that optimality against small probabilities of deviations is achieved, and that a perfect equilibrium uses only admissible strategies. In fact, a perfect equilibrium of the agent-normal-form induces a sequential equilibrium of the dynamic version of the game. Moreover, if the payoffs of the dynamic game are generic (that is, not related to each other by polynomial equations) then every sequential equilibrium is also perfect.

A stronger refinement selects *proper* equilibria (Myerson 1978). This refinement supposes that the more inferior the expected payoff from a strategy is, the less likely it is to be used. The construction differs only in step 1: if one pure strategy S is inferior to another T in response to the others' strategies then S has probability no more than *e* times the probability of T. A proper equilibrium induces a sequential equilibrium in every one of the equivalent descriptions of the dynamic game.

A perfect or proper equilibrium depends on the particular perturbation used to construct an ε-perfect or ε-proper equilibrium. Sometimes a game has an equilibrium that is *essential* or *truly perfect* in that any σ can be used when perturbing the game by less than ε toward σ, as above. This is usual for a static game with generic payoffs because in this case its equilibria are isolated and vary continuously with perturbations. However, such equilibria rarely exist in the important case that the static game represents a dynamic game, since in this case some strategies have the same equilibrium payoffs. This occurs because there is usually considerable freedom about how a player acts in contingencies off the predicted path of the equilibrium; in effect, the same outcome results whether the player 'punishes' others only barely enough to deter deviations, or more than enough. Indeed, for a dynamic game with generic payoffs, all the equilibria in a connected set yield the same equilibrium outcome because they differ only off the predicted path of equilibrium play. One must therefore consider sets of equilibria when invoking stringent refinements like truly perfect. One applies a somewhat different test to sets of equilibria. When considering a set of equilibria one requires that every sufficiently small perturbation (within a specified class) of the game has an equilibrium near some equilibrium in the set. Some refinements insist on a minimal closed set of equilibria with this property, but here we ignore minimality.

The chief refinement of this kind uses strategy perturbations to generate perturbed games. Kohlberg and Mertens (1986) say that a set of equilibria is *stable* if for each neighbourhood of the set there exists a positive probability ε such that, for every completely mixed strategy combination σ*,* each perturbation of the game by less than ε toward σ has an equilibrium within the neighbourhood. Stability can be interpreted as truly perfect applied to sets of equilibria and using the class of payoff perturbations generated by strategy perturbations. Besides the fact that a stable set always exists, it satisfies several criteria: it uses only *admissible* strategies, it contains a stable set of the reduced game after deleting a strategy that is weakly dominated or an inferior response to all equilibria in the set (these assure *iterative elimination of weakly dominated strategies* and a version of *forward induction),* and it is *invariant* to addition or deletion of redundant strategies. However, examples are known in which a stable set of a static game does not include a sequential equilibrium of the dynamic game it represents. This failure to satisfy the backward induction criterion can be remedied in various ways that we describe next.

One approach considers the larger class of all payoff perturbations. In this case, invariance to redundant strategies is not assured so it is imposed explicitly. For this, say that two games are equivalent if deletion of all redundant strategies results in the same reduced game. Similarly, randomized strategies in these two games are equivalent if they yield the same randomization over pure strategies of the reduced game. Informally, a set of equilibria is hyperstable if, for every payoff perturbation of every equivalent game, there is an equilibrium equivalent to one near the set. Two formal versions are the following. Kohlberg and Mertens (1986) say that a set S of equilibria is *hyperstable* if, for each neighbourhood N of those strategies in an equivalent game that are equivalent to ones in S, there is a sufficiently small neighbourhood P of payoff perturbations for the equivalent game such that every game in P has an equilibrium in N. A somewhat stronger version is the following. A set S of equilibria of a game G is *uniformly hyperstable* if, for each neighbourhood N of S, there is a $\delta > 0$ such that every game in the δ-neighbourhood of any game equivalent to G has an equilibrium equivalent to one in N. This version emphasizes that uniform hyperstability is closely akin to a kind of continuity with respect to payoff perturbations of equivalent games. Unfortunately, both of these definitions are complex, but the second actually allows a succinct statement in the case that the set S is a ‘component’ of equilibria, namely, a maximal connected set of the Nash equilibria. In this case the component is uniformly hyperstable if and only if its topological index is non-zero (Govindan and Wilson 2005), and thus *essential* in the sense used in algebraic topology to characterize a set of fixed points of a function that is slightly affected by every perturbation of the function. This provides a simply computed test of whether a component is uniformly hyperstable.

Hyperstable sets tend to be larger than stable sets of equilibria because they must be robust against a larger class of perturbations, but for this same reason the criterion is actually stronger. Within a hyperstable component there is always a stable set satisfying the criteria listed previously. There is also a proper equilibrium that induces a sequential equilibrium in every dynamic game with the same static representation – thus, the criterion of *backward induction* is also satisfied. Selecting a stable subset or a proper equilibrium inside a hyperstable component may be necessary because there can be other equilibria within a hyperstable component that use inadmissible strategies. Nevertheless, for a dynamic game with generic payoffs, all the equilibria within a single component yield the same outcome, since they differ only off the path of equilibrium play, so for the purpose of predicting the outcome rather than players’ strategies it is immaterial which equilibrium is considered. However, examples are known in which an inessential hyperstable component contains two stable sets with opposite indices with respect to perturbations of strategies.

The most restrictive refinement is the revised definition of stability proposed by Mertens (1989). Although this definition is highly technical, it can be summarized briefly as follows for the mathematically expert reader. Roughly, a closed set of equilibria is (Mertens-) *stable* if the projection map (from its neighbourhood in the graph of the Nash equilibria into the space of games with perturbed strategies) is essential. Such a set satisfies all the criteria listed previously, and several more. For instance, it satisfies the *small-worlds* criterion (Mertens 1992), which requires that adding other players whose strategies have no effect on the payoffs for the original players has no effect on the selected strategies of the original players. The persistent mystery in the study of refinements is why such sophisticated constructions seem to be necessary if a single definition is to satisfy all the criteria simultaneously. The clue seems to be that, because Nash equilibria are the solutions of a fixed-point problem, a fully adequate refinement must ensure that fixed points exist for every perturbation of this problem.

Govindan and Wilson (2009a, b) characterize Mertens-stability by three axioms adapted from decision theory. They consider refinements of the Nash equilibria of games with perfect recall that select connected closed subsets called solutions. (1) Undominated Strategies: no player uses a weakly dominated strategy in any equilibrium in a solution. (2) Backward Induction: each solution contains a quasi-perfect equilibrium and thus a sequential equilibrium in strategies that provide conditionally admissible optimal continuations from information sets. (3) Generalized Small

Worlds: A refinement is immune to embedding a game in a larger game with additional players provided the original players' strategies and payoffs are preserved, i.e. solutions of a game are the same as those induced by the solutions of any larger game in which it is embedded. This third axiom implies small worlds and invariance. For games with two players and generic payoffs, they prove that these axioms are equivalent to requiring that each solution is an essential component of equilibria in undominated strategies, and thus a stable set as defined by Mertens (1989).

The State of the Art of Refinements

The development of increasingly stronger refinements by imposing ad hoc criteria incrementally was a preliminary to more systematic development. Eventually, one wants to identify decision-theoretic criteria that suffice as axioms to characterize refinements. The two groups of refinements described above approach this problem differently. Those that consider perturbations seek to verify whether there exist refinements that satisfy many or (in the case of Mertens-stability) most criteria. From its beginning in the work of Selten (1975), Myerson (1978), and Kohlberg and Mertens (1986), this has been a productive exercise, showing that refinements can enforce more stringent criteria than Nash (1950, 1951) requires. However, the results obtained depend ultimately on the class of perturbations considered, since Fudenberg et.al. (1988) show that each Nash equilibrium of a game is the limit of strict equilibria of perturbed games in a very general class. Perturbations are mathematical artefacts used to identify refinements with desirable properties, but they are not intrinsic to a fundamental theory of rational decision making in multi-person situations. Those in the other group directly impose decision-theoretic criteria – admissibility, iterative elimination of dominated or inferior strategies, backward induction, invariance, small worlds, and so on. Their ultimate aim is to characterize refinements axiomatically. But so far none has obtained an ideal refinement of the Nash equilibria.

See Also

- ▶ Behavioural Game Theory
- ▶ Epistemic Game Theory: Incomplete Information
- ▶ Game Theory
- ▶ Harsanyi, John C. (1920–2000)
- ▶ Markov Equilibria in Macroeconomics
- ▶ Nash, John Forbes (Born 1928)
- ▶ Nash Program
- ▶ Selten, Reinhard (Born 1930)
- ▶ Signalling and Screening

Bibliography

Banks, J., and J. Sobel. 1987. Equilibrium selection in signaling games. *Econometrica* 55: 647–661.

Blume, L., A. Brandenburger, and E. Dekel. 1991a. Lexicographic probabilities and choice under uncertainty. *Econometrica* 59: 61–79.

Blume, L., A. Brandenburger, and E. Dekel. 1991b. Lexicographic probabilities and equilibrium refinements. *Econometrica* 59: 81–98.

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

Fudenberg, D., D. Kreps, and D. Levine. 1988. On the robustness of equilibrium refinements. *Journal of Economic Theory* 44: 351–380.

Fudenberg, D., and J. Tirole. 1991. Perfect Bayesian equilibrium and sequential equilibrium. *Journal of Economic Theory* 53: 236–260.

Govindan, S., and R. Wilson. 2005. Essential equilibria. *Proceedings of the National Academy of Sciences, USA* 102: 15706–15711.

Govindan, S., and R. Wilson. 2009a. On forward induction. *Econometrica* 77: 1–28.

Govindan, S. and Wilson, R. 2009b. Axiomatic theory of equilibrium selection for generic two-player games, Stanford Business School Research Paper 2021. https://gsbapps.stanford.edu/researchpapers/library/RP2021.pdf.

Harsanyi, J. 1967–1968. Games with incomplete information played by 'Bayesian' players, I–III. *Management Science* 14: 159–82, 320–34, 486–502.

Harsanyi, J., and R. Selten. 1988. *A general theory of equilibrium selection in games*. Cambridge, MA: MIT Press.

Hillas, J. 1998. How much of 'forward induction' is implied by 'backward induction' and 'ordinality'? Mimeo: Department of Economics, University of Auckland.

Hillas, J., and E. Kohlberg. 2002. The foundations of strategic equilibrium. In *Handbook of game theory*, ed. R. Aumann and S. Hart, Vol. 3. Amsterdam: North-Holland/Elsevier Science Publishers.

Kohlberg, E. 1990. Refinement of Nash equilibrium: the main ideas. In *Game theory and applications*, ed. T. Ichiishi, A. Neyman, and Y. Tauman. San Diego: Academic Press.

Kohlberg, E., and J.-F. Mertens. 1986. On the strategic stability of equilibria. *Econometrica* 54: 1003–1038.

Kreps, D. 1990. *Game theory and economic modeling*. New York: Oxford University Press.

Kreps, D., and R. Wilson. 1982. Sequential equilibria. *Econometrica* 50: 863–894.

Mertens, J.-F. 1989. Stable equilibria – a reformulation, Part I: definition and basic properties. *Mathematics of Operations Research* 14: 575–624.

Mertens, J.-F. 1992. The small worlds axiom for stable equilibria. *Games and Economic Behavior* 4: 553–564.

Mertens, J.-F., and S. Zamir. 1985. Formulation of Bayesian analysis for games with incomplete information. *International Journal of Game Theory* 14: 1–29.

Myerson, R. 1978. Refinement of the Nash equilibrium concept. *International Journal of Game Theory* 7: 73–80.

Nash, J. 1950. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences USA* 36: 48–49.

Nash, J. 1951. Non-cooperative games. *Annals of Mathematics* 54: 286–295.

Selten, R. 1965. Spieltheoretische Behandlung eines Oligopolmodells mit Nachfragetragheit. *Zeitschrift fur die gesamte Staatswissenschaft* 121(301–24): 667–689.

Selten, R. 1975. Reexamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory* 4: 25–55.

van Damme, E. 1984. A relation between perfect equilibria in extensive form games and proper equilibria in normal form games. *International Journal of Game Theory* 13: 1–13.

van Damme, E. 1989. Stable equilibria and forward induction. *Journal of Economic Theory* 48: 476–496.

van Damme, E. 1991. *Stability and perfection of Nash equilibria*. Berlin: Springer-Verlag.

Nash Program

Roberto Serrano

Abstract

This article is a brief survey on the Nash program for coalitional games. Results of non-cooperative implementation of the Nash solution, the Shapley value and the core are discussed.

Keywords

Cooperative games; Core; Edgeworth, F.; Nash program; Nash solution; Non-cooperative games; Shapley value; Subgame perfect equilibrium; Walrasian outcome

JEL Classifications

C7

In game theory, ‘Nash program’ is the name given to a research agenda, initiated in Nash (1953), intended to bridge the gap between the cooperative and non-cooperative approaches to the discipline.

Many authors have contributed to the program since its beginnings (see Serrano, 2005, for a comprehensive survey). The current article concentrates on a few salient contributions. One should begin by introducing some preliminaries and providing definitions of some basic concepts.

Preliminaries

The non-cooperative approach to game theory provides a rich language and develops useful tools to analyse strategic situations. One clear advantage of the approach is that it is able to model how specific details of the interaction may affect the final outcome. One limitation, however, is that its predictions may be highly sensitive to those details. For this reason it is worth also analysing more abstract approaches that attempt to obtain conclusions that are independent of such details. The cooperative approach is one such attempt.

Here are the primitives of the basic model in cooperative game theory. Let $N = \{1, \ldots, n\}$ be a finite set of players. For each S, a non-empty subset of N, we shall specify a set $V(S)$ containing jSj-dimensional payoff vectors that are feasible for coalition S. Thus, a reduced form approach is taken because one does not explain what strategic choices are behind each of the payoff vectors in $V(S)$. In addition, in this formulation, referred to as the characteristic function, it is implicitly assumed that the actions taken by the complement coalition (those players not in S) cannot prevent S from achieving each of the payoff vectors in $V(S)$. There are more general models in which these sorts of externalities are considered, but for the most part the contributions to the Nash program have been confined to the characteristic function model. Given a collection of sets $V(S)$, one for each S, the theory formulates its predictions on the basis of solution concepts.

A solution is a mapping that assigns a set of payoff vectors in $V(N)$ to each characteristic function $(V(S))_{S \subseteq N}$. Thus, a solution in general prescribes a set, although it can be single-valued (when it assigns a unique payoff vector as a function of the fundamentals of the problem). The leading set-valued cooperative solution concept is the core, while the most used single-valued ones are the Nash bargaining solution and the Shapley value.

There are several criteria to evaluate the reasonableness or appeal of a cooperative solution. One could start by defending it on the basis of its definition alone. In the case of the core, this will be especially relevant: in a context in which players can freely get together in groups, the prediction should be payoff vectors that cannot be improved upon by any coalition. Alternatively, one can propose axioms, abstract principles, that one would like the solution to have, and the next step is to pursue their logical consequences. Historically, this was the first argument to justify the Nash solution and the Shapley value. However, some may think that the definition may be somewhat arbitrary, or one may object that the axiomatic approach is 'too abstract'. By proposing non-cooperative games that specify the details of negotiation, the Nash program may help to counter these criticisms. First, the procedure will tell a story about how coalitions form and what sort of interaction among players is happening. In that process, because the tools of non-cooperative game theory are used for the analysis, the cooperative solution will be understood as the outcome of a series of strategic problems facing individual players. Second, novel connections and differences among solutions may now be uncovered from the distinct negotiation procedures that lead to each of them. Therefore, a result in the Nash program, referred to as a 'non-cooperative foundation' or 'non-cooperative implementation' of a cooperative solution, enhances its significance, being looked at now from a new perspective. Focusing on the features of the rules of negotiation that lead to different cooperative solutions takes one a long way in opening the 'black box' of how a coalition came about, and contributes to a deeper understanding of the circumstances under which one solution versus another may be more appropriate to use.

The Nash Bargaining Solution

A particular case of a characteristic function is a two-player bargaining problem. In it, $N = \{1, 2\}$ is the set of players. The set $V(\{1, 2\})$, a compact and convex subset of $\mathbb{R}$, is the set of feasible payoffs if the two players reach an agreement. Compactness may follow from the existence of a bounded physical pie that the parties are dividing, and convexity is a consequence of expected utility and the potential use of lotteries. The sets $(V(\{i\}))_{i \in N}$ are subsets of $\mathbb{R}$, and let $d_i = \max V(\{i\})$ be the disagreement payoff for player i, that is, the payoff that i will receive if the parties fail to reach an agreement. It is assumed that $V(\{1, 2\})$ contains payoff vectors that Pareto dominate the disagreement payoffs. A solution assigns a feasible payoff pair to each bargaining problem.

This is the framework introduced in Nash (1950), where he proposes four axioms that a solution to bargaining problems should have. First, expected utility implies that, if payoff functions are rescaled via positive affine transformations, so must be the solution (scale invariance). Second, the solution must prescribe a Pareto efficient payoff pair (efficiency). Third, if the set $V(\{1, 2\})$ is symmetric with respect to the 45 degree line and $d_1 = d_2$, the solution must lie on that line (symmetry). Fourth, the solution must be independent of 'irrelevant' alternatives, that is, it must pick the same point if it is still feasible after one eliminates other points from the feasible set (IIA). Because of scale invariance, there is no loss of generality in normalizing the disagreement payoff to 0. We call the resulting problem a normalized problem.

Nash (1950) shows that there exists a unique solution satisfying scale invariance, efficiency, symmetry and IIA, and it is the one that assigns to each normalized bargaining problem the point (u_1, u_2) that maximizes the product $v_1 v_2$ over all (v_1, v_2) $\in V(\{1, 2\})$. Today we refer to this as the 'Nash solution'. The use of the Nash solution is pervasive in applications and, following the axioms in Nash (1950), it is usually viewed as a normatively appealing resolution to bargaining problems.

In the first paper of the Nash program, Nash (1953) provides a non-cooperative approach to his axiomatically derived solution. This is done by means of a simple demand game. The two players are asked to demand simultaneously a payoff: player 1 demands v_1 and player 2 demands v_2. If the pair (v_1, v_2) is feasible, so that (v_1, v_2) $\in V(\{1, 2\})$, the corresponding agreement and split of the pie takes place to implement these payoffs. Otherwise, there is disagreement and payoffs are 0. To fix ideas, let us think of the existence of a physical pie of size 1 that is created if agreement is reached, while no pie is produced otherwise. Thus, player i's demand v_i corresponds to demanding a share x_i of the pie, $0 \leq x_i \leq 1$, such that player i's utility or payoff from receiving x_i is v_i.

The Nash demand game admits a continuum of Nash equilibria. Indeed, every point on the Pareto frontier of $V(\{1, 2\})$ is a Nash equilibrium outcome, as is the disagreement payoff point if each player demands the payoff corresponding to having the entire pie. However, Nash (1953) introduces uncertainty concerning the exact size of the pie. Now players, when formulating their demands, must have to take into account the fact that with some probability the pair of demands may lead to disagreement, even if they add up to less than 1. Then, it can be shown that the optimal choice of demands at a Nash equilibrium of the demand game with uncertain pie converges to the Nash solution payoffs as uncertainty becomes negligible. Hence, the Nash solution arises as the rule that equates marginal gain (through the increase in one's demanded share) and marginal loss (via the increase in the probability of disagreement) for each player when the problem is subject to a small degree of noise and demands/commitments are made simultaneously.

Rubinstein (1982) proposes a different non-cooperative procedure. In it, time preferences – impatience – and credibility of threats are the main forces that drive the equilibrium. The game is a potentially infinite sequence of alternating offers. In period 0, player 1 begins by making the first proposal. If player 2 accepts it, the game ends; otherwise, one period elapses and the rejector will make a counter-proposal in period 1, and so on. Let $\delta \in [0, 1)$ be the common per period discount factor, and let $v_i(\cdot)$ be player i's utility function over shares of the pie, assumed to be concave and strictly monotone. Thus, if player i receives a share x_i in an agreement reached in period t, his payoff is $\delta^{t\text{-}1} v_i(x_i)$. Perpetual disagreement has a payoff of 0.

Using subgame perfect equilibrium as the solution concept (the standard tool to rule out non-credible threats in dynamic games of complete information), Rubinstein (1982) shows that there exists a unique prediction in his game. Specifically, the unique subgame perfect equilibrium prescribes an immediate agreement on the splits (x,1 – x) – offered by player 1 – and (y, 1 – y) – by player 2 – which are described by the following equations:

$$\begin{aligned} v_1(y) &= \delta v_1(x) \\ v_2(1-x) &= \delta v_2(1-y). \end{aligned}$$

That is, at the unique equilibrium, the player acting as a responder in a period is offered a share that makes him exactly indifferent between accepting and rejecting it to play the continuation: the bulk of the proof is to show that any other behaviour relies on non-credible threats.

As demonstrated in Binmore, Rubinstein and Wolinsky (1986), the unique equilibrium payoffs of the Rubinstein game, regardless of who is the first proposer, converge to the Nash solution payoffs as $\delta \to 1$. First, note that the above equations imply that, for any value of δ, the product of payoffs $v_1(x)v_2(1 - x)$ is the same as the product $v_1(y)v_2(1 - y)$. Thus, both points, ($v_1(x)$,$v_2(1 - x)$) and ($v_1(y)$,$v_2(1 - y)$), lie on the same hyperbola of equation $v_1 v_2 = K$ and, in addition, since they correspond to efficient agreements, both points also lie on the Pareto frontier of $V(\{1,2\})$. Finally, as $\delta \to 1$, one has that $x \to y$ so that the two proposals (the one made by player 1 and the other by player 2) converge to one and the same, the one that yields the Nash solution payoffs. Thus, credible threats in dynamic negotiations in which both players are equally and almost completely patient also lead to the Nash solution.

The Shapley Value

Now consider an n-player coalitional game where payoffs are transferable in a one-to-one rate among different players (for instance, because utility is money for all of them). This means that $V(S)$, the feasible set for coalition S, is the set of payoffs $(x_i)_{i \in S}$ satisfying the inequality $\sum_{i \in S}$ $xi \leq v(S)$ for some real number $v(S)$. This is called a transferable utility or TU game in characteristic function form. The number v (S) is referred to as the 'worth of S', and it expresses S's initial position (for example, the maximum total utility that the group S of agents can achieve in an exchange economy by redistributing their endowments when utility is quasi-linear).

Therefore, without loss of generality, we can describe a TU game as a collection of real numbers $(v(S))_{S \subseteq N}$. A solution is then a mapping that assigns to each TU game a set of payoffs in the set $V(N)$, that is, vectors $(x_1, \ldots, x_n)$ such that $\sum i_{\in N}$ $x_i \leq v(N)$. In this section, as in the previous one, we shall require that the solution be single-valued. Shapley (1953) is interested in solving in a fair way the problem of distribution of surplus among the players, when taking into account the worth of each coalition. To do this, he resorts to the axiomatic method. First, the payoffs must add up to $v(N)$,which means that the entire surplus is allocated (efficiency). Second, if two players are substitutes because they contribute the same to each coalition, the solution should treat them equally (symmetry). Third, the solution to the sum of two TU games must be the sum of what it awards to each of the two games (additivity). Fourth, if a player contributes nothing to every coalition, the solution should pay him nothing (dummy).

The result in Shapley (1953) is that there is a unique single-valued solution to TU games satisfying efficiency, symmetry, additivity and dummy. It is what today we call the Shapley value, the function that assigns to each player i the payoff

$$\mathrm{Sh}_i(N, v) = \sum_{S, i \in S} \frac{(|S| - 1)!(|N| - |S|)!}{|N|!} \times [v(S) - v(S/\{i\})].$$

That is, the Shapley value awards to each player the average of his marginal contributions to each coalition. In taking this average, all orders of the players are considered to be equally likely. Let us assume, also without loss of generality, that $v(\{i\}) = 0$ for each player i.

Hart and Mas-Colell (1996) propose the following non-cooperative procedure. With equal probability, each player $i \in N$ is chosen to publicly make a feasible proposal to the others: $(x_1, \ldots, x_n)$ is such that the sum of its components cannot exceed $v(N)$. The other players get to respond to it in sequence, following a pre-specified order. If all accept, the proposal is implemented; otherwise, a random device is triggered. With probability $0 \leq \delta < 1$, the same game continues being played among the same n players (thus, a new proposer will be chosen again at random among them), but with probability $1 - \delta$ the proposer leaves the game. He is paid 0 and his resources are removed so that, in the next period, proposals to the remaining $n - 1$ players cannot add up to more than $v(N/\{i\})$. A new proposer is chosen at random among the set $N/\{i\}$, and so on.

As shown in Hart and Mas-Colell (1996), there exists a unique stationary subgame perfect equilibrium payoff profile of this procedure, and it actually coincides with the Shapley value payoffs for any value of δ . (Stationarity means that strategies cannot be history dependent.) As $\delta \to 1$, the Shapley value payoffs are also obtained not only in expectation but independently of who the proposer is. One way to understand this result, as done in Hart and Mas-Colell (1996), is to check that the rules of the procedure and stationary behaviour in it are in agreement with Shapley's axioms. That is, the equilibrium relies on immediate acceptances of proposals, stationary strategies treat substitute players similarly, the equations describing the equilibrium have an additive structure, and dummy players will have to receive 0 because no resources are destroyed if they are asked to leave. It is also worth stressing the important role in the procedure of players' marginal contributions to coalitions: following a rejection, a proposer incurs the risk of being thrown out and the others of losing his resources, which seem to suggest a 'price' for them.

The Core

The idea of agreements that are immune to coalitional deviations was first introduced to economic theory in Edgeworth (1881), who defined the set of coalitionally stable allocations of an economy under the name 'final settlements'. Edgeworth envisioned this concept as an alternative to Walrasian equilibrium (Walras, 1874), and was also the first to investigate the connections between the two concepts. Edgeworth's notion, which today we refer to as 'the core', was rediscovered and introduced to game theory in Gillies (1959). Therefore, the origins of the core were not axiomatic. Rather, its simple definition appropriately describes stable outcomes in a context of unfettered coalitional interaction. (The axiomatizations of the core came much later: see, for example, Peleg, 1985, 1986; Serrano and Volij, 1998).

For simplicity, let us continue to assume that we are studying a TU game. In this context, the core is the set of payoff vectors $x = (x_1, \ldots, x_n)$ that are feasible, that is, $\sum_{i \in N} x_i \leq v(N)$, and such that there does not exist any coalition $S \subseteq N$ for which $\sum_{i \in S} x_i \leq v(S)$. If such a coalition S exists, we shall say that S can improve upon or block x, and x is deemed unstable. The core usually prescribes a set of payoffs instead of a single one, and it can also prescribe the empty set in some games.

To obtain a non-cooperative implementation of the core, the procedure must embody some feature of anonymity, since the core is usually a large set and it contains payoffs where different players are treated very differently. Perry and Reny (1994) build in this anonymity by assuming that negotiations take place in continuous time, so that anyone can speak at the beginning of the game instead of having a fixed order. The player that gets to speak first makes a proposal consisting of naming a coalition that contains him and a feasible payoff for that coalition. Next, the players in that coalition get to respond. If they all accept the proposal, the coalition leaves and the game continues among the other players. Otherwise, a new proposal may come from any player in N. It is shown that, if the TU game has a non-empty core (as well as any of its subgames), the stationary subgame perfect equilibrium outcomes of this procedure coincide with the core. If a core payoff is proposed to the grand coalition, there are no incentives for individual players to reject it. Conversely, a non-core payoff cannot be sustained because any player in a blocking coalition has an incentive to make a proposal to that coalition, who will accept it (knowing that the alternative, given stationarity, would be to go back to the non-core status quo). Moldovanu and Winter (1995) offer a discrete-time version of the mechanism: in their work, the anonymity required is imposed on the solution concept by looking at order-independent equilibria.

Serrano (1995) sets up a market to implement the core. The anonymity of the procedure stems from the random choice of broker. The broker announces a vector $(x_1, \ldots, x_n)$, where the components add up to $v(N)$. One can interpret x_i as the price for the productive asset held by player i. Following an arbitrary order, the remaining players either accept or reject these prices. If player i accepts, he sells his asset to the broker for the price x_i and leaves the game. Those who reject get to buy from the broker, at the called out prices, the portfolio of assets of their choice if the broker still has them. If a player rejects but does not get to buy the portfolio of assets he would like because someone else took them before, he can always leave the market with his own asset. The broker's payoff is the worth of the final portfolio of assets that he holds, plus the net monetary transfers that he has received. Serrano (1995) shows that the prices announced by the broker will always be his top-ranked vectors in the core. If the TU game is such that gains from cooperation increase with the size of coalitions, the set of all subgame perfect equilibrium payoffs of this procedure will coincide with the core. Core payoffs are here understood as those price vectors where all arbitrage opportunities in the market have been wiped out. Finally, yet another way to build anonymity in the procedure is by allowing the proposal to be made by brokers outside of the set N, as done in Pérez-Castrillo (1994).

See Also

- Bargaining
- Non-Cooperative Games (Equilibrium Existence)
- Shapley Value

Bibliography

Binmore, K., A. Rubinstein, and A. Wolinsky. 1986. The Nash bargaining solution in economic modelling. *RAND Journal of Economics* 17: 176–188.

Edgeworth, F. 1881. Mathematical psychics. In *F. Y. Edgeworth's mathematical psychics and further papers on political economy*, ed. P. Newman. Oxford: Oxford University Press.

Gillies, D. 1959. Solutions to general non-zero-sum games. In *Contributions to the theory of games IV*, ed. A. Tucker and R. Luce. Princeton, NJ: Princeton University Press.

Hart, S., and A. Mas-Colell. 1996. Bargaining and value. *Econometrica* 64: 357–380.

Moldovanu, B., and E. Winter. 1995. Order independent equilibria. *Games and Economic Behavior* 9: 21–34.

Nash, J. 1950. The bargaining problem. *Econometrica* 18: 155–162.

Nash, J. 1953. Two person cooperative games. *Econometrica* 21: 128–140.

Peleg, B. 1985. An axiomatizationof the core of cooperative games without side payments. *Journal of Mathematical Economics* 14: 203–214.

Peleg, B. 1986. On the reduced game property and its converse. *International Journal of Game Theory* 15: 187–200.

Pérez-Castrillo, D. 1994. Cooperative outcomes through non-cooperative games. *Games and Economic Behavior* 7: 428–440.

Perry, M., and P. Reny. 1994. A non-cooperative view of coalition formation and the core. *Econometrica* 62: 795–817.

Rubinstein, A. 1982. Perfect equilibrium in a bargaining model. *Econometrica* 50: 97–109.

Serrano, R. 1995. A market to implement the core. *Journal of Economic Theory* 67: 285–294.

Serrano, R. 2005. Fifty years of the Nash program, 1953–2003. *Investigaciones Económicas* 29: 219–258.

Serrano, R., and O. Volij. 1998. Axiomatizations of neoclassical concepts for economies. *Journal of Mathematical Economics* 30: 87–108.

Shapley, L. 1953. A value for *n*-person games. In *Contributions to the Theory of Games II*, ed. A. Tucker and R. Luce. Princeton, NJ: Princeton University Press.

Walras, L. 1874. *Elements of pure economics, or the theory of social wealth*, Trans. W. Jaffé. Philadelphia: Orion Editions, 1984.

Nash, John Forbes (Born 1928)

Joel Watson

Abstract

Nash originated general non-cooperative game theory in seminal articles in the early 1950s by formally distinguishing between non-cooperative and cooperative models and by developing the concept of equilibrium for non-cooperative games. Nash developed the first bargaining solution characterized by axioms, pioneered methods and criteria for relating cooperative-theory solution concepts and non-cooperative games, and also made fundamental contributions in mathematics. Nash was the 1994 recipient of the Bank of Sweden Prize in Economic Sciences in Memory of Alfred Nobel, jointly with John C. Harsanyi and Reinhard Selten.

Keywords

Coalitions; Commitment; Contract curve; Cooperative games; Cournot, A. A; Dominance; Equilibrium; Equilibrium refinements; Evolutionary stability; Expected utility; Fixed-point methods; Game theory; Maximin strategy; Morgenstern, O; Multiple equilibria; Nash bargaining solution; Nash demand game; Nash equilibrium; Nash program; Nash. J. R., Jr; Non-cooperative games; Prisoner's dilemma; Rational behaviour; Strategic and extensive-form games; Strategic independence; von Neumann, J

JEL Classifications

B31

The Context of for Nash's Work: Von Neumann and Morgenstern

Nash's contributions to the theory of games were fundamental to the development of the discipline and its interface with applied fields of study. This section provides is a short account of the state of affairs before Nash's work. For a more detailed account, see the suggestions for further reading at the end of this article.

The first significant step in mathematical modelling of strategic situations was Augustin Cournot's (1838) book on oligopoly, where Cournot presented models of firm interaction that were analysed using what we now call Nash equilibrium. But Cournot did not attempt, or perhaps even recognize, how the analysis might generalize. Further, in the ensuing years confusion persisted regarding whether it would be appropriate for a firm to incorporate a response by its rivals when considering whether to change its own action. The concept of *strategic independence* – that the players' strategies can be considered to be chosen simultaneously and independently – began to be clarified by Emile Borel's (1921) description of a *method of play.*

Game theory became a discipline with the work of John von Neumann (1928), which was incorporated into the path-breaking book by von Neumann and Oscar Morgenstern (1944, 1947). In the book, von Neumann and Morgenstern formally defined both the extensive form (tree-based) and normal form (strategy-based) representations of games, related by the notion of a strategy; they studied for the first time a general class of games, defining solutions and proving existence using fixed- point methods; they introduced the idea of analysing how coalitions of players can take advantage of binding agreements; and they provided a theory of utility and decision-making under risk (the expected utility criterion). With one book, game theory was created and put on solid footing.

Von Neumann and Morgenstern were interested in developing a positive theory of behaviour in games – for any given game, a 'solution'. In a nutshell, their analysis progresses as follows:

1. Formulate a solution concept for two-player *zero-sum games,* which have the defining property that, for each *strategy profile* (one strategy for each player), the players' payoffs sum to zero. Such a game is special because the only economic concern is distributional; in other words, the game models a situation of pure conflict between the players, where one player's winnings come at the other's expense.
2. Analyse n-player zero-sum games by assuming that coalitions of players could bind together and play as a team against the other players. This requires assuming that coalitions can communicate before the game and make binding agreements on how to play. The value of forming a coalition is calculated in reference to the implied zero-sum game that the coalitions play against one another, which ultimately is a two-player game to which the solution from Part 1 above is applied.
3. To evaluate a non-zero-sum, n-player game, imagine the existence of a fictitious player $n + 1$ whose payoff is defined as negative of the sum of the other players' payoffs. This creates a zero-sum game to which the preceding applies.

For an illustration of von Neumann and Morgenstern's analysis of two-player zero- sum games (Part 1 above), consider a simple example. Suppose that players 1 and 2 interact in the normal form game depicted in the following table.

1\2	X	Y	Z
A	4, −4	0, 0	−2, 2
B	3, −3	1, −1	1, −1
C	2, −2	1, −1	1, −1

Player 1 selects between strategies A, B, and C. Simultaneously, player 2 chooses between X, Y, and Z. The players' payoffs, which might as well be in monetary terms, are shown in the cells of the table, with player 1's payoff written first. Note that this is a zero-sum game in that, in each cell of the table, the players' payoffs sum to zero.

Von Neumann and Morgenstern motivated their solution concept by considering sequential variations of games in which one player would move first and then the other player, having seen what the first selected, would respond. Their key concept is what is generally known as a 'maximin strategy', also called a 'security strategy'. A security strategy for a given player is a strategy that gives the highest guaranteed payoff level; that is, it maximizes the minimum that the player could get, where the minimum is calculated over all of the strategies of the other player.

In the example, B and C are both security strategies for player 1 because, regardless of what player 2 does, player 1 gets a payoff of at least 1 when using either of these strategies, whereas it is feasible for player 1 to obtain a lower payoff (0 or -2, in particular) by selecting strategy A. For player 2, Y and Z are security strategies and they guarantee a payoff of at least -1.

Von Neumann and Morgenstern's general analysis focuses on mixed strategies (probability distributions over pure strategies) in finite two-player games, to which the maximin definition extends. They prove that the players' security levels (the amounts that the security strategies guarantee) sum to zero. Thus, when each player selects his security strategy, each player obtains exactly his security level payoff. Further, when one player selects his security strategy, the other player can do no better than select her own security strategy; that is, the two players' security strategies are optimal responses to each other. Security strategies also describe optimal play in zero-sum games that are played sequentially. For example, if player 1 had the privilege of selecting among A, B, and C *after* observing player 2's choice, both players would still select security strategies. Finally, security strategies are interchangeable in that the preceding conclusions hold equally well for any combination of security strategies, for instance (B, Y) as well as (B, Z).

Although von Neumann and Morgenstern had developed a theory that applied to all finite games, their theory is essentially empty for non-zero-sum games. For example, in converting a two-player game into a three-player game by adding the fictitious player 3, von Neumann and Morgenstern basically change the rules of the game for the original two players, who now can make binding agreements. The resulting prediction is that the two players will bind themselves to a strategy profile that maximizes the sum of their payoffs, with each player getting at least his security level. Von Neumann and Morgenstern's theory is therefore incomplete and unsatisfying on two fronts. First, for non-zero-sum games, it offers no treatment of rationality in the absence of binding commitments. Second, it offers no way of predicting the outcome of a two-player bargaining problem beyond Francis Ysidro Edgeworth's (1881) contract curve and it relies on transferable utility. Nearly all interesting economic examples involve efficiency concerns and hence are not zero- sum in nature, so economics had little to benefit from game theory until another significant step could be made in the modelling of rational behaviour.

Nash's Contributions

Nash's contributions to the emerging discipline of game theory were equally as bold as were von Neumann and Morgenstern's and, in terms of applicability, even more significant. Nash's main contributions were made in a series of four papers published between 1950 and 1953 and summarized in this section.

In his articles in the *Proceedings of the National Academy of Sciences* in 1950 and the *Annals of Mathematics* in 1951, which reported his dissertation research, Nash (*a*) introduced and made clear the distinction between cooperative and non-cooperative games – the latter being games in which players act independently (that is, without the assumption about coalitions that von Neumann and Morgenstern adopted) – and (*b*) defined a solution concept for non-cooperative games. The first four paragraphs from Nash's *Annals of Mathematics* article describe the context and the contribution succinctly:

> Von Neumann and Morgenstern have developed a very fruitful theory of two- person zero-sum games in their book *Theory of Games and Economic Behavior*. This book also contains a theory of n-person games of a type which we would call cooperative. This theory is based on an analysis of the interrelationships of the various coalitions which can be formed by the players of the game.
>
> Our Theory, in contradistinction, is based on the *absence* of coalitions in that it is assumed that each participant acts independently, without collaboration or communication with any of the others.
>
> The notion of an *equilibrium point* is the basic ingredient in our theory. This notion yields a generalization of the concept of the solution of a two-person zero-sum game. It turns out that the set of equilibrium points of a two- person zero-sum game is the set of all pairs of opposing 'good strategies.'
>
> In the immediately following sections we shall define equilibrium points and prove that a finite non-cooperative game always has at least one equilibrium point. We shall also introduce the notions of solvability and strong solvability of a non-cooperative game and prove a theorem on the geometrical structure of the set of equilibrium points of a solvable game. (1951, p. 286)

Nash's equilibrium concept became known as 'Nash equilibrium'. It and the cooperative/non-cooperative distinction were cited by the Royal Swedish Academy of Sciences in awarding Nash the Nobel Prize.

In more mathematical and modern language, here are the definitions of *best response* (in Nash's words, a 'good strategy') and Nash equilibrium. Consider any game defined by a number n of players; a strategy set S_i for each player $i = 1,2,\ldots,n$; and, for each player i, a payoff function $u_i : S \rightarrow \mathbf{R}$, where S is the set of strategy profiles. The strategy sets may be defined as mixed strategies for some underlying set of pure strategies, in which case the payoff functions, as expectations, are linear in the mixed strategies. For a player i, we write '$-i$' to refer to the other players. Given a strategy vector s_{-i} for the other players, player i's strategy s_i is called a best response if player i can do no better than to select s_i; that is, we have $u_i(s_i, s_{-i}) \geq u_i\left(s_i', s_{-i}\right)$ for every strategy s_i' of player i. Then strategy profile $s* = \left(s_1^*, s_2^*, \ldots, s_n^*\right)$ is called a Nash equilibrium if every player is best responding to the others – that is, if for each player i, it is the case that $s*$ is a best response to s_{-i}^*.

For an illustration of Nash equilibrium and its relation to security strategies, consider the game depicted in the following table.

1\2	X	Y	Z
A	2, 3	1, 2	6, 5
B	1, 0	0, 2	4, 0
C	3, 4	2, 2	2, 0

Observe that, in this game, C and Y are the players' security strategies, so a naive application of von Neumann and Morgenstern's maximin theory (absent binding agreements) would predict that strategy profile (C, Y) be played. However, this strategy profile is plainly inconsistent with the idea that players are rational in responding to each other. In particular, if player 1 is expected to select C then player 2 behaves quite irrationally by choosing Y. In fact, strategy Y is *not even rationalizable* for player 2; it does not survive iterated removal of dominated strategies (see below). Thus, the notion of a security strategy is not a good theory of behaviour for non-zero-sum games, demonstrating the limits of von Neumann and Morgenstern's analysis.

Next, observe that the game has two Nash equilibria in pure strategies, (C, X) and (A, Z). Both of these are reasonable predictions in the sense that, in both cases, the players are best responding to one another. For example, if player 1 is sure that player 2 will select X, then it is best for player 1 to select C; likewise, if player 2 is convinced that player 1 will select C, then it is optimal for player 2 to choose X. There is also a mixed-strategy Nash equilibrium in which player 1 randomizes between A and C, and player 2 randomizes between X and Z. That the game has multiple Nash equilibria demonstrates the general economic problem of coordination, in particular the possibility that the players will coordinate on the less efficient Nash equilibrium. Other games, such as the *Prisoner's Dilemma,* have only inefficient equilibria and thus reveal a fundamental tension between individual and joint incentives.

Nash's intuitive concept of equilibrium facilitated the analysis of *all* noncooperative games, opening the door to widespread application of game theory. Indeed, Nash equilibrium has become the dominant solution concept for the analysis of games. Through an ingenious fixed-point argument, Nash also proved the existence of an equilibrium point in every finite game. Further, in his dissertation (1950b) Nash offered two interpretations of the concept, one based on rational reasoning by individual players and the other describing stability of the distribution of strategies chosen by a population of individuals who interact over time. The latter is a precursor to the methodology of the literature on learning in games and to the modern theories of *evolutionary stability* in biology (Maynard Smith 1984). Nash's 1951 *Annals of Mathematics* article also contains a section that defines 'dominance' (meaning one strategy yields a strictly higher payoff than another, regardless of what the other players do) and explains how an iterated dominance procedure can be used to rule out strategies that are not equilibria. Thus, Nash also made observations that would resurface in the concept of 'rationalizable strategic behaviour' (Bernheim 1984; Pearce 1984), the main nonequilibrium notion of rationality. Nash even was among the first to perform game experiments, as his co-authored article in the volume *Decision Processes* (Kalisch et al. 1954) attests.

In his 1950 *Econometrica* article, Nash tackled the two-person bargaining problem with the objective of determining a unique solution (a precise 'value' that eluded von Neumann and Morgenstern) from the underlying set of alternatives and the players' preferences. Nash took a cooperate-theory approach by positing a system of four axioms that reasonably characterize properties one might expect the outcome of a bargaining process to exhibit: (*a*) a notion of equal bargaining power, (*b*) invariance to inessential utility transformations, (*c*) efficiency, and (*d*) independence of the solution to the removal of so-called irrelevant alternatives. Nash proved that a particular function of parameters (which maximizes the product of surpluses) is exactly characterized by the axioms. The analysis showed that it is possible to reasonably identify a precise outcome of a bargaining problem. It also initiated the axiomatic method for the analysis of bargaining (where theorists explore how different axioms characterize various functional solutions), starting a literature that thrived for several decades. The *Nash bargaining solution* is still the dominant solution in applied economic models.

Nash's second paper on bargaining (the 1953 *Econometrica* article) took another major step by connecting the non-cooperative and cooperative approaches to strategic analysis. At the heart of this theoretical exercise is an underlying noncooperative game, which gives a set of feasible payoffs, and a technology for the players to make binding commitments about the mixed strategies that they will play in the underlying game. In the model, players first simultaneously make threats, which are mixed strategies they are bound to play if they do not reach an agreement. Then the players interact in a non-cooperative bargaining game in which they simultaneously make payoff demands – this stage is now called the 'Nash demand game'. If their payoff demands are feasible in the underlying game, then the players obtain their demanded payoffs; otherwise, the players get what their threats imply.

Nash observed that the demand game has generally an infinite number of equilibria, revealing a coordination aspect to the bargaining problem. But Nash went further in developing a brilliant method to 'escape from this troublesome non--uniqueness' by looking at the limit of 'smooth' approximations of the demand game. Amazingly, Nash showed that the limit is unique and coincides with the prediction of his axiomatic model; that is, the limit is the Nash bargaining solution. Nash's limit argument was the forerunner to the enormous literature on *equilibrium refinements*, an area of research that thrived decades later and was the primary subject of Nash's Nobel co-recipients. More significantly, Nash argued that the relation between the cooperative solution concept and the equilibrium in the non-cooperative model justifies wide use of the cooperative solution as a reasonable shorthand for the actual non-cooperative setting. Nash's argument, and fascinating theoretical result, established the profession's understanding of the connection between cooperative and non-cooperative models and initiated the literature on what is now called the 'Nash program'.

After completing the work in game theory just described, Nash made fundamental contributions in pure mathematics – contributions that, in terms of mathematical depth and originality, were of an even higher order of sophistication and importance. According to leading mathematician John Milnor, Nash's

> subsequent mathematical work is far more rich and important [in this mathematical sense]. During the following years he proved that every smooth compact manifold can be realized as a sheet of a real algebraic variety, proved the highly anti-intuitive C1-isometric embedding theorem, introduced powerful and radically new tools to prove the far more difficult C1-isometric embedding theorem in high dimensions, and made a strong start on fundamental existence, uniqueness, and continuity theorems for partial differential equations. (Milnor 1998, p. 1330.

It is not appropriate to provide here details on Nash's pure mathematics work (nor is it possible, due to the limitations of the author's fields of expertise).

Nash's Personal Life

Nash's character became legendary with the publication of a biography by Sylvia Nasar (1998) and a 2001 feature film produced by Brian Grazer and Ron Howard. Nash's remarkable personal journey began in Bluefield, West Virginia, where he was born and raised. He explored mathematics and conducted science experiments as a child, and attended Carnegie Institute of Technology, where the mathematics department discovered in him a budding genius. Nash's ideas on bargaining that were published as 'The Bargaining Problem' (1950c) were developed while he was an undergraduate student at Carnegie, during the only economics course he took, on international trade.

Nash studied mathematics in the graduate program at Princeton University, where, as his biography describes, he was boorish, cocky, and a renowned adversary in strategic contests. At Princeton, Nash added to his prodigious achievements, finishing his dissertation – the work on non-cooperative games and equilibrium that would bring him the Nobel Prize – in his second year. (Nash also invented the board game *Hex,* a game independently created by Danish mathematician Piet Hein.) Nash taught at Princeton for one year and then took a position at Massachusetts Institute of Technology, where he was on the faculty until 1959. There he conducted the research that won him great acclaim in the mathematics community.

Nash's genius in advancing game theory and mathematics was paired with deep personal challenges. In 1959 Nash began experiencing the severe mental disturbances of paranoid schizophrenia. He resigned from MIT and began a phase of life marked by delusional thinking, an escape to Europe, repeated hospitalizations, unsuccessful medical treatments, and then a long, disengaged presence at Princeton. In the mid-1980s Nash miraculously began to emerge from the delusional haze in what he describes as a gradual rejection of psychotic thinking on intellectual grounds (Nash 1995). After a quarter century of detachment, Nash's life regained a measure of normality.

Nash's Legacy in Game Theory and Economics

There is no simple way of quantifying the enormous reach of Nash's ideas. The notions of Nash equilibrium, the Nash bargaining solution, the Nash demand game, and the Nash program have found such widespread acceptance and application that it has become customary, and perhaps even appropriate, for researchers to forgo formally citing Nash's articles when utilizing these concepts. Nash ideas helped to propel game theory from a mathematical sub-field into a full discipline, with major use and application in not only economics, where it is the main and worthy alternative to the competitive-market framework, but also in theoretical biology, political science, international relations and law.

Beyond its theoretical content, Nash's work also made a stylistic departure from that of von Neumann and Morgenstern, whose book methodically records definitions, examples, and analysis for numerous special cases in the process of developing general theory. Nash, in contrast, used the terse style of the mathematician, presenting his ideas with minimal obscuring features. His 1950 *Proceedings of the National Academy of Sciences* entry, for instance, is generously allotted two pages and could have been typeset on one. The benefit of focusing on the basic mathematical concepts is that it allows for a broad range of interpretations and extensions. For example, there are several motivations for Nash equilibrium, including as a condition for self-enforcement of a contract (which is an important topic in the current literature). A hallmark of excellent theoretical modelling is precise and straightforward expression of assumptions and conclusions, with their relation shown in the most simple and elegant way possible.

Mathematician Milnor, after offering the assessment of Nash's work in pure mathematics that is quoted above, continues with by saying: 'However, when mathematics is applied to other branches of human knowledge, we must really ask a quite different question: To what extent does the new work increase our understanding of the real world? On this basis, Nash's thesis was nothing short of revolutionary' (1998, p. 1330). Two leading game theorists of today say 'Nash's theory of non-cooperative games should now be recognized as one of the outstanding intellectual advances of the twentieth century' (Myerson 1999, p. 1067) and 'His work lay the foundation of non-cooperative game theory, now the predominant mode of analysis of strategic interactions in economics, political science, and biology' (Crawford 2002, p. 380).

When viewed from the perspective of five short decades, game theory has caused a revolution in economics and other fields of study. It was with the work of John Nash that the flame so exquisitely ignited by von Neumann and Morgenstern became the torch that would eventually set the social sciences ablaze.

See Also

- ▶ Bargaining
- ▶ Game Theory
- ▶ Morgenstern, Oskar (1902–1977)
- ▶ Nash Program
- ▶ Non-cooperative Games (Equilibrium Existence)
- ▶ von Neumann, John (1903–1957)

Acknowledgments The author thanks Vincent Crawford, Joel Sobel and Martin Dufwenberg for comments on a preliminary draft.

Bibliography

Items indicated with an asterisk provide good further background reading on John F. Nash, Jr. Also, the *Scandinavian Journal of Economics,* vol. 97, issue 1 (1995), contains articles on John Nash and his co-Nobel Prize recipients, John C. Harsanyi and Reinhard Selten. For a complete list of Nash's publications, including his papers in pure mathematics, see Milnor (1998).

Bernheim, B.D. 1984. Rationalizable strategic behavior. *Econometrica* 52: 1007–1028.

Borel, E. 1921. La théorie du jeu et les équations intégrales à noyau symétrique gauche. *Comptes Rendus de l'Académie des Sciences* 173: 1304–1308. English translation by L.J. Savage, *Econometrica* 21: 97–100 (1953).

Cournot, A. 1838. *Recherches sur les Principes Mathématiques de la Théorie des Richesses.* Paris: Hatchette. English translation by N.T. Bacon, *Researches into the mathematical principles of the theory of wealth*. New York: Macmillan, 1927.

Crawford, V.P. 2002. John Nash and the analysis of strategic behavior. *Economics Letters* 75: 377–382.

Edgeworth, F.Y. 1881. *Mathematical psychics*. London: Kegan Paul.

Hammerstein, P., et al. 1996. The work of John Nash in game theory: Nobel seminar, December 8, 1994. *Journal of Economic Theory* 69: 153–185.

Kalisch, C., J. Milnor, J. Nash, and E. Nering. 1954. Some experimental n-person games. In *Decision processes*, ed. R.M. Thrall, C.H. Coombs, and R.L. Davis. New York: Wiley.

Mayberry, J.P., J.F. Nash, and M. Shubik. 1953. A comparison of treatments of a duopoly situation. *Econometrica* 21: 141–154.

Maynard Smith, J. 1984. *Evolution and the theory of games*. New York: Cambridge University Press.

*Milnor, J. 1995. A Nobel Prize for John Nash. *The Mathematical Intelligencer* 17: 11–17.

*Milnor, J. 1998. John Nash and 'a beautiful mind'. *Notices of the American Mathematical Society* 45: 1329–1332.

Myerson, R.B. 1999. Nash equilibrium and the history of economic theory. *Journal of Economic Literature* 37: 1067–1082.

*Nasar, S. 1998. *A beautiful mind*. New York: Simon and Schuster.

Nash Jr., J.F. 1950a. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences, USA* 36: 48–49.

Nash Jr., J.F. 1950b. Non-cooperative games. Doctoral dissertation, Princeton University.
Nash Jr., J.F. 1950c. The bargaining problem. *Econometrica* 18: 155–162.
Nash Jr., J.F. 1951. Non-cooperative games. *Annals of Mathematics* 54: 286–295.
Nash Jr., J.F. 1953. Two-person cooperative games. *Econometrica* 21: 128–140.
*Nash Jr., J.F. 1995. Autobiography. In *Les Prix Nobel. The Nobel Prizes 1994,* ed. Frängsmyr, T. Stockholm: Nobel Foundation. Online. Available at http://nobelprize.org/nobel_prizes/economics/laureates/1994/nash-autobio.html. Accessed 29 Nov 2006.
Pearce, D. 1984. Rationalizable strategic behavior and the problem of perfection. *Econometrica* 52: 1029–1050.
von Neumann, J. 1928. Zur theories der gesellschaftsspiele. *Mathematische Annalen* 100: 295–320. English translation by S. Bergmann in *Contributions to the theory of games IV,* ed. R. D. Luce and A. W. Tucker. Princeton: Princeton University Press, 1959.
von Neumann, J., and O. Morgenstern. 1944. *Theory of games and economic behavior.* Princeton: Princeton University Press (2nd ed. 1947).

Nathan, Robert Roy (Born 1908)

J. K. Galbraith

Nathan was born in Dayton, Ohio, and had his undergraduate and graduate training at the University of Pennsylvania and his legal training at Georgetown University. Strongly influenced by Simon Kuznets, he was one of the handful of innovating statisticians who brought National Income and Gross National Product accounting into active use in the United States government, where, from 1934 to 1940, he was Chief of the National Income Division of the Bureau of Foreign and Domestic Commerce of the Department of Commerce. With the increased threat of war he moved in 1940 from the Department of Commerce to the Office of Production Management, later the War Production Board, where he brought national production accounting to bear on the problems of war production. Showing therefrom that unused capacity and possible weapons production were far greater than commonly believed, he was largely responsible for the huge Victory Program approved by President Roosevelt a few weeks before the attack on Pearl Harbor. Then, with Simon Kuznets, who had joined him in Washington, he worked out feasible schedules for weapons production in the early months of the war. The importance of this work for the success of the American war effort cannot be exaggerated. The Germans, having no analysis of comparable value, had no way of knowing their production possibilities and, in consequence, greatly underestimated them.

Nathan's work also brought him into sharp conflict with the business executives who had been drawn to Washington from private industry and who, relying confidently on their experience and presumed knowledge, regarded his figures as extravagantly optimistic, an exercise in grave academic impracticality. For some months in these years the war with Hitler and the Japanese sank into the background in competition with the conflict with Nathan and Kuznets. In 1943, to the wholly undisguised relief of the businessmen, Nathan went into the Army.

In the four decades following World War II, Nathan headed a highly successful, socially oriented consulting firm in Washington, Robert R. Nathan Associates, Inc., which extended advice on economic development to a score or more of governments, including those of France, Korea, Burma, Colombia, Afghanistan, El Salvador, Nigeria, Indonesia, Venezuela and Thailand. Additionally, he has had an active role in a wide range of academic and public-service organizations, has been a figure of importance in liberal Washington politics and an active member of the American Statistical Association and has served on various corporate boards of directors.

National Accounting, History Of

André Vanoli

Abstract

With antecedents as far back as the late 17th century, national accounting is a product of the Great Depression, the Second World War and the subsequent period of recovery and economic growth. Soon after the war, country experiences and international harmonization processes interacted, eventually leading to a complete accounting framework with the 1993 SNA/ESA 1995. Until the mid-1970s, national accounting experienced a kind of golden age, after which greater difficulties arose, in terms of the increased complexity of economic life, widened social concerns and theoretical challenges. In that context, impressive achievements and a sense of frustration have coexisted.

Keywords

Aggregates; Aukrust, O; Balance sheets; Basic prices; Boisguilbert, P; Brown, A; Clark, C. G; Computable general equilibrium models; Copeland, M.A; Denison, E; Domestic product; Economic and Monetary Union; European System of Accounts; Factor cost valuation; Fair value valuation; Fisher, I; Frisch, R. A. K; Gilbert, M; Goldsmith R; Gordon, R; Great Depression; Grüning, F; Gruson, C; Hicks, J. R; Historical cost valuation; Income; Indicative planning; Input–output analysis; International Monetary Fund; Kampmann, V; Keynes, J. M; King, G; Kravis, I; Kuznets, S; Leontief, W; Lindahl, E. R; Maastricht criteria; Market price valuation; Marx, K; Material Product System; Mesoeconomic level; National accounting; National accounts statistics; National income; Ohlsson, I; Petty, W; Popov, P. I; Purchasing power parities; Quesnay, F; Ruggles, N; Ruggles, R; Satellite accounts; Smith, A; Social accounting matrices; Social accounting,; Stone, J. R. N; Structural change; Sustainability; System of Balances of the National Economy; System of National Accounts; *Tableau économique*; Tinbergen, J; Value added; Van Cleeff, E; Vauban, S; Vincent, A; Walras, L; Warburton, C; Wealth; Welfare indicators

JEL Classifications

B2

National accounting is a product of the 20th century, more precisely of the Great Depression, the Second World War and the subsequent period of recovery and economic growth. However, two and a half centuries earlier, estimates of national income had started with William Petty and Gregory King in England, and Vauban and Boisguilbert in France. This innovation in England, by the end of the 17th century, has been attributed to 'the spirit of the age' (Phyllis Deane 1955), 'an age of great intellectual vigour, scientific curiosity and inventiveness' (Richard Stone 1986). This early work had two main purposes: on the one hand, taxation and fiscal reforms, and on the other the assessment of the nations' comparative economic strength in an age when England, France and the Netherlands were frequently at war. Exceptionally, King, an outstanding pioneer, made consistent estimates of various economic magnitudes (income, expenses, increase or decrease in wealth, and so on) for a series of years. However, as a rule, national income was estimated as an isolated magnitude using various methods. Estimates were intermittent and extended slowly (according to Studenski 1958, national income had been estimated at least once for only eight countries by the end of the 19th century, and for some 20 by 1929. From 1850, earlier in England, evaluations of fortune or wealth, more numerous, were disconnected from national income estimates.

From National Income Estimate to National Accounting

The influence of the First World War was limited, with some exceptions (for example, an NBER 1909–19 series in current and constant dollars published by Wesley Mitchell et al. in 1921–22). The 1929 crisis was a turning point. Official demand appeared (US Senate 1932; Carson 1975, p. 156) leading to a 1934 report prepared by Simon Kuznets and his assistants *(National Income 1929–1932,* in current prices, by type of economic activity and distributed income). Estimates were then extended to expenditures (final consumption and capital formation) by Clark

Warburton. In a number of countries – the Netherlands (Jan Tinbergen), Sweden, Denmark (Viggo Kampmann) – large programs were developed, such as the one resulting in *National Income in Sweden 1861–1930* published in 1937 by Erik Lindahl, Einar Dahlgren and Karin Koch. Working on his own, Colin Clark in the United Kingdom extended his previous 1932 estimates to a quite comprehensive coverage (*National Income and Outlay* 1937).

The 1930s were a period of maturation in economics, apart from the conceptual and methodological deepening directly involved in this stream of quantitative estimates. The stimulus to quantitative macroeconomics given by Keynes's *General Theory* (1936) provided the theoretical basis for the estimation of interdependent economic aggregates, for the relationships between income and expenditure and between saving and investment were central to his argument. Such interrelationships had not previously been absent from economic theories (think of Quesnay's *Tableau économique,* Marx's reproduction schemes or Walras's general equilibrium analysis). However, after the Great Depression, such concepts and their statistical representations became central to macroeconomic concerns and policies. Keynes's works were focused on macroeconomic relations, but others sought representations of the economic system as a whole in different ways. Ferdinand Grüning in Germany (1933) analysed the economic circuit at a level later called 'mesoeconomic', halfway between the macro and micro levels. Wassily Leontief's research (1941) introduced input–output analysis at the level of homogeneous industrial groups, with a much broader view, in terms of general equilibrium, than the descriptive detailed balances of relations between branches (industries) prepared by P.I. Popov (1926) in the Soviet Union. The idea of an accounting approach for the economy as a whole, similar to the business accounting approach, was introduced either as a tool for improving national income estimates (as by Morris A. Copeland, following an intuition of Irving Fisher) or as part of a new proposed economic organization (André Vincent in France, Ed Van Cleeff in the Netherlands). The idea of micro/macro relationships was present in much of this work. Coming from a very different perspective, Ragnar Frisch developed an axiomatic, bottom-up representation of economic circulation.

The Second World War was the second, decisive, turning point. National accounting, often called at the beginning social accounting, crystallized in a direct response to the problem of war finance in the UK, as explicitly stated in the April 1941 White Paper (UK Treasury, *An Analysis of the Sources of War Finance and Estimate of the National Income and Expenditure in 1938 and 1940).* This was backed up by a technical paper by James Meade and Richard Stone in 1941. A more elaborated 'social accounting' system was soon proposed by Stone in an appendix to *Measurement of National Income and The Construction of Social Accounts* (published by the United Nations in 1947). Inspired by business accounting, it included sector accounts grouping accounting entities and their transactions organized according to a sequence of sub-accounts, with a set of detailed definitions and the discussion of many unsettled issues. Although it covered neither balance sheets nor a detailed analysis of the productive system, this accounting system was well ahead of its time. Actually, before and during the war, the United States was in advance in both national income and related aggregates estimates and their use, as for instance in the 1942 feasibility study of the Victory Program led by Kuznets or the analysis of the inflationary gap (Carson 1975, p. 174–7). However, the National Income Division of the Commerce Department, with Milton Gilbert, evolved towards a simple accounting framework rather than a developed accounting system.

Though they encountered many difficulties and though it was a very uneven development, mostly due to deficiencies in statistical information and staffing, national accounting experienced a kind of golden age in the three decades following the war. Economic reconstruction and growth policies, the large increase in the economic role of government and the welfare state, the extension of international cooperation (for example, the Marshall Plan and, later, the Common Market in Europe), with the consequent emphasis on measuring of the rate of growth, led to a great demand for national accounts. This comprised the requirements of Keynesian macroeconomic demand management for short-term economic budget forecasts and longer-term projections needed for various types of indicative planning (the latter being particularly important in France). The development of econometric techniques and national accounts estimates reinforced each other. This trend towards greater use of national accounting data was general, even though the economies involved ranged from basically liberal economies such as the United States to more controlled economies such as France, the Netherlands and Norway.

International Harmonization and Extensions

Country experiences interacted with the process of international harmonization very early. Discussion between Canada, the UK and the USA took place in September 1944. There was a meeting of a League of Nations Committee, for which Stone prepares a memorandum, in December 1945. Stone played a prominent role in the first generation of standardized systems (OEEC 1950, 1952; United Nations 1952). This first attempt at standardization across the Western world as a whole, however, was too limited in scope, and was very far from the ambitions of the 1945 accounting scheme. Conceived as a simplified model for countries that were only beginning to develop their national accounts, it could not meet the needs of countries that were already more advanced, such as Scandinavian countries (Odd Aukrust in Norway, Ingvar Ohlsson in Sweden) or even a country like France. Under the impulse of Claude Gruson, France was, in the 1950s, in order to implement far- reaching economic policies, beginning the process of building a comprehensive and ambitious system of its own, integrating accounts for economic agents, input–output tables and financial transactions in a way that was more integrated than the Copeland’s money-flows accounts in the United States.

Until the end of the 1960s the Western stage was characterized by the existence of a variety of national systems that were difficult to reconcile, even among those countries that adopted, in principle, the same comprehensive concept of production, including non-market government services. The new French system adopted a narrower concept of production, limited to market goods and services. The Soviet Union and its satellites used the even more restricted concept of material production, limited to goods and the so-called material services (mostly the transport of goods), following the old tradition of Smith and Marx. However, during the 1960s intense international discussions took place, on the basis of the wide range of national experiences in Europe and North America and the demands of international organizations. The result was the adoption of a second generation of standardized systems, the 1968 System of National Accounts (SNA) and the new European System of Accounts (ESA 1970), prepared on the basis of a report by Stone for the UN (the OECD deleting its system) and a French expert for the European Community. The European Community, thinking the 1952 system was too narrow and unsuited to harmonizing the accounts of its original six members and to meeting the needs of Community policies, had decided in 1964 to establish its own system.

The new system (they can be described as a single system, for SNA and ESA were very close) was closer to Stone’s 1945 inspiration and to the French, Scandinavian and British systems than to the 1952 standardized system, in terms of coverage (in particular of input–output tables and financial accounts), integration and institutional orientation. The main weakness remained the absence of balance sheets, despite the pioneering work of Raymond Goldsmith in the United States at the beginning of the 1960s. Fixed capital formation was limited to tangible assets and the relation between income and changes in wealth was not fully shown.

The System of Balances of the National Economy, built around the material product concept, was also standardized, though little innovation was involved, through the framework of the Council of Mutual Economic Assistance, and then published by the United Nations (1971). Careful comparisons between the SNA and the Material Product System (MPS) were carried out in the UN European Economic Commission in Geneva.

France decided to leave its own peculiar system and join, via ESA 1970, the international system, this being achieved by 1976. The USA was not actively involved in the elaboration of the 1968 SNA, keeping its National Income and Product Accounts, whose accounting and conceptual framework had evolved little since 1947.

A quarter of a century later, a third generation of normalized systems has taken the trend towards a universal system a step further. The 1993 SNA/ESA 1995 closed the accounting framework by including balance sheets and completing the accumulation accounts with the introduction of a revaluation account (holding gains and losses) and an account for other types of capital gains and losses. Intangible capital formation was partly accounted for. In the current accounts, the analysis of income distribution was deepened (primary income distribution, secondary distribution, and redistribution in kind), actual final consumption was differentiated from final consumption expenditures, via the re-routing of social transfers in kind from government to households. This clarification of the accounting relation between income and changes in wealth (net worth) has deep implications (see below).

Nearly full integration was achieved between the SNA and the International Monetary Fund manuals (Balance of Payments, Government Finance Statistics, Monetary and Financial Statistics). The MPS disappeared at the beginning of the 1990s with the collapse of the Soviet Union and the fast transition of China towards a market economy. Paradoxically, the USA followed a slower path towards adopting the SNA framework.

During this long process of extension and harmonization of the accounting framework, the substance of the accounts changed dramatically in comparison with what was involved when the focus was on estimating national income. The product aggregate soon became the most important one, on a par with the expenditure aggregate. The income aggregate not only lost its position of being the single aggregate, but was often given a secondary position. From that, a series of consequences resulted.

The factor cost method of valuation, when still in use, was reduced to a lower rank than the market price valuation (in spite of the recurrent objection of ‘double counting’). The latter was much more convenient for the valuation of expenditure and the analysis of consumer behaviour. In an integrated framework, the market price valuation was then applied also to the product aggregate (domestic product takes progressively the first place) and much later on to the income aggregate. In the 1993 SNA, full recognition was given to the concept of national income at market prices, which is in fact the new name given to the earlier concept of national product (which was not actually a product but an income concept).

Partly for similar reasons, gross concepts have generally come to be preferred in practice, even though net concepts, that is, after deduction of consumption of fixed capital (depreciation in the usual business terminology), were considered closer to what was generally understood by the idea of national income. Both gross and net concepts of product, income and expenditure are finally considered part of the SNA/ ESA.

The analysis and measurement of production and flows of products (goods and services), both in current value and in volume, have been given an increasing importance in relation to the integration of supply and use or input–output tables (a characteristic feature of the 1968 SNA/ESA 1970). This is increasingly done using the framework of annual tables. The integration with income estimates is less clear in practice, though the concept of value added, a significant improvement, and not only in words, on the old expression ‘net output’ or ‘net product’, provides the necessary link.

In this context, thanks to Stone’s contribution, significant improvements in valuation concepts were made in the 1968 SNA. This widens and differentiates the usual notion of market prices. Basic prices, excluding net taxes on products, were introduced on the output side, resulting in the measurement of value added at basic prices.

All taxes, minus subsidies, on products are then introduced. On the use side, acquisition prices are defined as purchasers' prices including only non-deductible taxes.

Measures in constant prices (described as volume measures), combining quantity and quality changes, also changed significantly. The trend was from globally deflating national income using a single price index in the 1930s, to deflating each of the main items in the balance of products (output, final consumption, and so on) using specific indices, and finally to an integrated system of price and volume measures, at a detailed level, using an input–output framework when annual tables were available (with Denmark, France, the Netherlands and Norway leading here). Double deflation, of output and inputs respectively, was used for value added in this context. International manuals by Stone (1956, 1968 SNA, ch. 4) and Peter Hill (1972; United Nations 1979) recommended such an approach. Later on the 1993 SNA/ESA 1995 recommended replacing the traditional fixed base indices with chain indices, preferably Fisher volume and price indices or acceptable alternatives.

Much more complex, both conceptually and practically, international comparisons of volume levels of aggregates were the object of an International Comparison Project (ICP), launched in 1968, after the pioneering research of Colin Clark (1940) and Gilbert and Irving Kravis (1954) at the OEEC. Purchasing power parities, more significant than exchange rates, were calculated. The results of the ICP, however, were not as widely implemented or as widely accepted as national volume measures, something that is unfortunate in a globalized world.

Beyond the progressive completion of its integrated framework, attempts were made to broaden the scope of national accounting by developing semi-integrated additional constructs, such as the satellite accounts whose idea was introduced (by Vanoli) by the end of the 1960 (for example, accounts for social protection, health, education, and environmental protection). In such an approach, the fully integrated system itself becomes the central framework (the expression often used, 'core accounts', is ambiguous).

Social accounting matrices (SAMs) were designed by Stone and Alan Brown in 1962, in order to achieve more flexibility than was possible using the usual account presentation. Though the word 'social' here means only 'for the whole economy', it gave rise to a certain ambiguity. SAMs are sometimes presented as a kind of alternative framework.

In the late 1980s, the Dutch proposed an ambitious 'system of economy-related statistics' as a way of organizing a vast array of statistics. A 'core system', narrower than the SNA central framework, was linked with 'system modules', such as social and environmental modules. This proposal had some similarity with the unsuccessful attempt by Stone, in the first half of the 1970s, to design for the United Nations a system of social and demographic statistics. It echoes the growing importance given to the micro–macro linkages (for example, Richard and Nancy Ruggles 1986), in parallel with the increased availability of micro-databases.

Concern for statistical coordination had, of course, been present in national accounting from the very beginning.

New Challenges Since the Mid-1970s

The achievements of national accounting, in the face of an enormous development of statistics, have been impressive. However, many countries are still far from fully implementing the international system (for example, few countries prepare integrated balance sheets), and economic and social conditions have changed drastically, especially since the mid-1970s. As a result national accounting, often questioned, sometimes radically, has had to face new challenges.

Since around 1980, after the supply shocks of the 1970s and the decreasing role played by macroeconometric models, national accounting has no longer been supported by the Keynesian paradigm. Some people even think it is obsolete. However, the demand for national accounts continues to grow, even if it also changes. Predominantly short-term concerns have led to a pressing demand for quarterly accounts, and even sometimes for a monthly GDP, resulting in conflicts between timeliness (early estimates are required) and accuracy. Though more accurate, through successive revisions, annual accounts seem less used and their results are less commented upon.

In the opposite direction, computable general equilibrium models have multiplied since the mid-1970s as a means of studying policies aimed at structural change. Without any concern for the setting up of time series, they are based on the accounts of a single year supplemented, as required, by other data dictated by the models' specificities and purposes. Although they use the somewhat misleading SAM terminology, they actually need national accounts bases.

It remains true, however, that for the study of structural and social policies economists and social researchers, since the last two decades of the 20th century, have generally preferred to make use of micro-simulation models. The role of national accounts data is relatively reduced in this context.

In contrast, a considerable extension of the institutional and political role of national accounting took place during the 1990s, mostly in Europe. Certain aggregates (GDP or GNP) had been used fairly early for administrative purposes such as country contributions to international organizations, eligibility thresholds to preferential World Bank loans, regional allocation of European structural funds, and the 'Fourth own budgetary resource' of the Community budget. However, the debate over accession criteria to the European Economic and Monetary Union (the creation of the euro) marked a qualitative jump in the consideration of national accounting by policymakers and public opinion. Most Maastricht criteria were defined in reference to the ESA (ratios of public deficit and public debt to GDP). The ESA became compulsory for member states of the European Union. This marked the culmination of the European statistical strategy adopted in the 1960s. Closely related to the international statistical systems, like the SNA, European statistical tools are in effect very often legally based.

The policy uses of the ESA necessitate effective harmonization of the content of the accounts. A procedure of verification and evaluation of the comparability and representativeness of GDP is established. Full harmonization is, however, difficult. Because conceptual and statistical issues and political considerations intervene, especially in the procedure for identifying excessive deficits, specific cases have to be studied, sometimes through a rather difficult process. Here, and in issues such as the ratio between compulsory levies and GDP, national accounts appear at the forefront of sensitive political concerns. While it clearly shows their importance, this situation may also have less positive aspects for the national accounts. There is the possibility of political pressures, though this is rare; there may be lack of flexibility; official obligations and procedures can be very time-consuming and, as a result of limited human resources, European national accountants may become insufficiently involved in research work.

No similar policy-led process is taking place at the world level. However the need for regulation on a global scale is increasingly felt. Monitoring and intervention aimed at remedying local and regional crises and at preventing systemic crises falls to the International Monetary Fund, in agreement with the principal economic powers. Hence the growing role of the IMF in the supply, by member states, of timely and well-documented harmonized information. In the last decade of the 20th century, the Fund set up a system of standards to guide countries in data dissemination, including meta-information concerning various characteristics of the data. The structuring role of national accounts has been particularly highlighted. The Fund has conducted assessment missions in order to evaluate the quality of countries' national accounts and data systems.

The impressive increase in the demand for and use of national accounts statistics has taken place against the background of economies which have become much more complex, and hence more difficult to describe and measure, than was the case in the three decades following the Second World War. The number and sophistication of available products have grown; changes in product quality have become more rapid; the share of services, generally more difficult to measure, especially in volume, has increased. The effects of technical change, opening the global economy, the transformation of enterprises and groups, refinements of price policies and consumer behaviour, continuing financial innovations, frequent extension of informal activities, and so on have caused a tendency for economic information systems to maladjust. Hence many controversies arise, notably on price and volume measurements of capital goods – quality change based on performances (Robert Gordon) or on resource cost (the traditional solution championed by Edward Denison) – or measurement of consumption goods and services, where the Boskin Report in the United States (Boskin et al. 1996) argued that the price increase was overestimated.

Significant methodological progress has been in areas such as the measurement of quality change of durable goods based on the change in their performance, the US having taken the lead. However the field is huge, and research is mostly concentrated on information and communication technology products. The measurement of financial and insurance services is in progress. Intangible assets are increasingly investigated. For non-market services, the necessary focusing on direct output–volume measurement instead of the traditional input–volume approach opens, at the start of the 21st century, another wide field of research. It soon appears that the relationship between the concepts of output and outcome must be clarified. On the other hand, some very important issues, like interest and inflation, the treatment of R&D expenditures and the extraction of subsoil resources, have remained outstanding for a long time, defying consensus, though relevant solutions do exist.

After a long emphasis on the relationship between production, income and expenditure, national accounting concerns have in recent decades been extended to the full set of relations between production, income, accumulation and wealth. This raises complex issues concerning the analysis and measurement of capital, particularly intangible assets, and consequently income. By the end of the 20th century business accountants faced similar difficulties with the emerging international accounting standards, moving from historical cost, which national accounting always rejected, to fair value valuation of assets.

Thus, national accounting is fighting for a better coverage of its traditional object at the same time that, at least since the early 1970s, new social concerns have given rise to requests for aggregate monetary indicators synthesizing broader sets of phenomena. There remain things that national accountants cannot do. One is the provision of a welfare indicator, a function that Kuznets assigned to national income, and which gave rise, in the 1940s, to an intense debate involving John Hicks and Paul Samuelson that reached negative conclusions (William Nordhaus and James Tobin 1973, later tried to provide such a measure with their 'measure of economic welfare'). Another is the measurement of an environmentally adjusted domestic product. The suggestions in this direction included in the 1993 United Nations Handbook, *Integrated Environmental and Economic Accounting,* do not reach a consensus and are not implemented. There was then a move towards wanting a sustainable product or income measure, but this does not make any answer easier, though Hicks's concept of income (the maximum amount that can be consumed in a period while expecting total wealth to be unchanged at the end of it) has increasingly been advocated in recent decades.

Most difficulties relate to the observation and measurement of non-market nonmonetary flows and stocks. Economists propose at least partial measurement solutions, within the framework of standard economic theory, using, for instance, contingent valuation methods (which raises problems of combination with actual exchange values, transfer of results and aggregation), or theoretical constructs with idealized conditions, seeking to justify a possible interpretation of net domestic product in terms of both welfare and sustainability. Other approaches, however, lean towards synthetic indicators combining both monetary and non-monetary variables.

Tensions between social concerns, theoretical issues and observation constraints of actual economies are increasingly at stake.

See Also

- ▶ Green National Accounting
- ▶ Kuznets, Simon (1901–1985)
- ▶ National Accounting, History Of
- ▶ National Income
- ▶ Stone, John Richard Nicholas (1913–1991)
- ▶ *Tableau économique*

Bibliography

Aukrust, O. 1994. The Scandinavian contribution to national accounting. In *The accounts of nations*, ed. Z. Kenessey. Amsterdam: IOS Press.

Boskin, M.J., E.R. Dulberger, and Z. Griliches. 1996. *Toward a more accurate measure of the cost of living.* Final report to the senate finance committee from the advisory commission to study the consumer price index. Washington, DC: Government Printing Office.

Carson, C.S. 1975. The history of the United States national income and product accounts: the development of an analytical tool. *Review of Income and Wealth* 21: 153–181.

Clark, C. 1937. *National income and outlay.* London: Macmillan.

Clark, C. 1940. *The conditions of economic progress*. London: Macmillan.

Commission of the European Communities, International Monetary Fund, Organisation for Economic Co-operation and Development, United Nations, World Bank. 1993. *System of national accounts 1993*. Brussels/Luxembourg/New York/Paris/Washington, DC: Commission of the European Communities/ United Nations/Organisation for Economic Co-operation and Development/World Bank.

Deane, P. 1955. The implications of early national income estimates for the measurement of long term economic growth in the United Kingdom. *Economic Development and Cultural Change* 4: 3–38.

Eurostat. 1996. *European system of accounts ESA 1995*. Luxembourg: Eurostat.

Gilbert, M., and I.B. Kravis. 1954. *An international comparison of national products and the purchasing power of currencies*. Paris: OEEC.

Grüning, F. 1933. *Der Wirtschaftskreislauf.* München: Beck.

Hill, T.P. 1972. *A system of integrated price and volume measures (Indices)*. Luxembourg: Statistical Office of the European Communities.

Kenessey, Z. (ed.). 1994. *The accounts of nations*. Amsterdam: IOS Press.

Kuznets, S. 1934. *National income 1929–1932.* US Senate Document No. 124, 73rd Congress, 2nd session. Washington, DC: Government Printing Office.

Kuznets, S. 1942. U.S. War Production Board, Planning Committee Document No. 151. A memorandum to the Planning Committee from Simon Kuznets on ‘Analysis of the Production program’, dated 12 August.

Leontief, W. 1941. *The structure of the American economy 1919–1929: An empirical application of equilibrium analysis*. Cambridge, MA: Harvard University Press.

Lindahl, E., E. Dahlgren, and K. Koch. 1937. *National income in Sweden 1861–1930*. London: P.S. King.

Meade, J., and R. Stone. 1941. The construction of tables of national income, expenditure, savings and investment. *Economic Journal* 51: 216–233.

Mitchell, W.C., W.I King, F.R. Macaulay, and O.W. Knauth. 1921–1922. *Income in the United States: Its amount and distribution, 1909–1919,* Parts I and II. New York: NBER.

Nordhaus, W., and J. Tobin. 1973. Is growth obsolete? In *The measurement of economic and social performance*, ed. M. Moss. New-York: Columbia University Press for NBER.

OEEC (Organisation for European Economic Co-operation). 1950. *A simplified system of national accounts*. Paris: OEEC.

OEEC (Organisation for European Economic Co-operation). 1952. *A standardised system of national accounts*. Paris: OEEC.

Popov, P.I., ed. 1926. *Balans narodnogo khoziaistva Soyuza SSSR 1923–1924 goda.* Moskva: Trudi Tsentralnogo Statisticheskogo Upravlenia, Tom XXIX.

Ruggles, R., and N.D. Ruggles. 1986. The integration of macro and micro data for the household sector. *Review of Income and Wealth* 32: 245–276.

Statistical Office of the European Communities. 1970. *European system of integrated economic accounts (ESA)*. Luxembourg: OSCE.

Stone, R. 1947. Definition and measurement of the national income and related totals. Appendix to *Measurement of national income and the construction of social accounts.* Geneva: United Nations.

Stone, R. 1956. *Quantity and price indexes in national accounts*. Paris: OEEC.

Stone, R. 1986. Nobel memorial lecture 1984: the accounts of society. *Journal of Applied Econometrics* 1: 5–28.

Studenski, P. 1958. *The income of nations*. New York: New York University Press.

UK Treasury. 1941. *An analysis of the sources of war finance and an estimate of the national income and expenditure in 1938 and 1940.* Cmd. 6261. London: HMSO.

United Nations. 1952. *A system of national accounts and supporting tables*. New York: United Nations.

United Nations. 1968. *A system of national accounts.* Studies in methods series F n° 2 Rev. 3. New-York: United Nations.

United Nations. 1971. *Basic principles of the system of balances of the national economy.* New York: United Nations.

United Nations. 1979. *Manual on national accounts at constant prices*. New York: United Nations.

United Nations. 1993. *Integrated environmental and economic accounting. Interim version*. New York: United Nations.

US Senate. 1932. S. Res. 220, 72nd Cong., 1st sess., *Congressional record* 75, 12285.
Vanoli, A. 2005. *A history of national accounting*. Amsterdam: IOS Press.

National Bureau of Economic Research

Malcolm Rutherford

Abstract

The National Bureau of Economic Research was founded in 1920 and has been regarded as one of the leading research organizations in economics ever since. This entry deals briefly with the founding of the NBER, its early research on national income and business cycles, its later research directions and contributions, and some of the more important changes in organization and direction that have occurred up to 2007.

Keywords

Burns, A. F.; Business cycle measurement; Commons, J. R.; Cowles Commission; Fabricant, S.; Feldstein, M.; Friedman, M.; Gay, E.; Kuznets, S.; Meyer, J.; Mitchell, W. C.; Moore, G.; National Bureau of Economic Research; National income; National income accounting; Rorty, M.; Schwartz, A.; Stigler, G.; Stone, N. I.

JEL Classifications

B2

The National Bureau of Economic Research (NBER) was founded in January 1920, and from the moment of its founding was seen as one of the leading independent research organizations in economics in the world (Fabricant 1984).

The NBER was established as an independent, non-partisan, research organization focused on empirical investigation. The original research orientation was towards 'basic' knowledge of the economy, but was, nevertheless, clearly intended to inform and improve the policymaking process. More recently the research focus has shifted to become more explicitly applied and policy orientated, but empirical work is still central to the bureau's mission. From the first, its Board included a large number of directors from various universities, scientific associations and other organizations. This and the system of manuscript review were designed to ensure the scientific impartiality of its work. These aspects of bureau organization still exist today.

The idea for an independent research bureau in economics sprang from discussions between Malcolm Rorty and N.I. Stone in 1916. Rorty was a statistician with AT&T, Stone an economist working as an arbitrator and economic advisor. Their policy views clashed but they could agree on the need for more reliable information. They involved Wesley Mitchell (Columbia), Edwin Gay (Harvard), and John R. Commons (Wisconsin, and then President of the American Economic Association). The First World War interrupted progress, but the experience of the war made the lack of quantitative information concerning the economy even more apparent, and by the AEA meeting of December 1919 all the necessary elements were in place.

The NBER began with a research agenda directed at the measurement of the size and distribution of national income, and the problem of business cycles. Wesley Mitchell was the first director of research, Edwin Gay the first president, while Rorty and Stone were members of the Board of Directors. Funding was obtained for a small research staff, originally consisting of Mitchell, Willford King, Frederick Macaulay and Oswald Knauth. The major financial contributors were the Commonwealth Fund, followed by the Carnegie Corporation, and, after 1923, the Laura Spelman Rockefeller Memorial Foundation (and its successor organization, the Social Science Division of the Rockefeller Foundation). The NBER also sold subscriptions and engaged in research commissioned by the President's Conference on Unemployment. In 1921 and 1922 the NBER published its first national income estimates: *Income in the United States: Its Amount and Distribution*. This was followed in 1923 by *Business Cycles and Unemployment,* produced by a special staff of the NBER for the President's Conference on Unemployment.

The NBER grew and prospered during the 1920s and early 1930s. The senior research staff were paid a modest stipend by the bureau, but generally held university appointments in the New York area. The bureau also employed research assistants and received funding for research fellowships and for statistical laboratory and library facilities. The research staff came to include Leo Wolman, F.C. Mills, Simon Kuznets, Arthur Burns and Solomon Fabricant. The bureau's research expanded to include Wolman's work on trade union membership, a substantial project on the topic of labour migration (undertaken by Harry Jerome, who was 'borrowed' from Wisconsin), F.C. Mills's extensive series of price studies, as well as further work on national income and business cycles. Mitchell produced the first of his projected volumes on business cycles, *Business Cycles: The Problem and its Setting,* in 1927. The bureau also continued its association with the President's Conference on Unemployment by contributing the research for *Recent Economic Changes in the United States* (1929). Kuznets took over the work on national income from King in 1931, and from 1933 he was 'loaned' to the Department of Commerce to work on the construction of official national income estimates. The first result of Kuznets's efforts was his report *National Income, 1929–32,* published in 1934.

A financial crisis in 1932 resulted in significant retrenchment at the bureau, which had suffered loss of income due to the Depression and faced uncertainty over the future of Carnegie support. The crisis was overcome thanks to the flexibility shown by Edmund Day of the Social Science Division of the Rockefeller Foundation, but Day expressed concerns with the bureau – its dependence on Rockefeller funding, its domination by a staff drawn heavily from Columbia University, and its lack of interaction with the broader academic community (Rutherford 2005).

Rockefeller continued to fund the NBER core programmes on national income, business cycles, price and price relationships, the labour market, and savings and capital formation. The bureau also took on a programme of financial research funded by the Association of Reserve City Bankers and headed by Ralph Young. Mitchell and Burns developed what became known as the 'NBER method' of specific and reference cycles to deal with the variations they found between cycles, but the project became ever larger. By the late 1930s the bureau's financial position had recovered and staff numbers again grew substantially, with Milton Friedman joining as an assistant to Kuznets in 1937 (he took over Kuznets's work on *Incomes from Independent Professional Practice*), Moses Abramovitz and Julius Shiskin arriving in 1938, and Geoffery Moore, among numerous others, in 1939.

Day's concerns were not without results. A Universities National Bureau Committee was established in 1935 to examine the potential of NBER–university cooperation. Out of this came the Conference on Income and Wealth (headed by Kuznets) and the Conference on Prices (headed by Mills). The first of these was particularly successful, producing the series Studies in Income and Wealth from 1938 onwards. In addition, Joseph Willits joined the bureau in 1936 as executive director, to deal with administration and fund raising. In 1939, Willits was appointed as Director of the Division of Social Science of the Rockefeller Foundation, and the NBER enjoyed strong support from Rockefeller until Willits left that position in 1954 Rutherford 2005.

Mitchell retired as Director of Research and was succeeded by Arthur Burns in 1945. Kuznets and Burns disagreed over the future direction of the bureau. Kuznets wished to shift the research emphasis to long-run growth, while Burns wished to maintain the focus on business cycles. Kuznets was to pursue his interests through the Conference on Income and Wealth with the financial support of the Social Science Research Council. Burns stayed as Director of Research until appointed to the Council of Economic Advisers in 1953. He was succeeded by Solomon Fabricant. Burns, however, returned to the bureau as President in 1957 and regained much of his previous authority within the organization.

In 1946, Burns and Mitchell published *Measuring Business Cycles,* the result of almost 20 years of effort on the business-cycle project, and the much delayed second volume of the three that were planned. The final, theoretical, volume was never completed. *Measuring Business Cycles* drew sharp criticism from Tjalling Koopmans of the Cowles Commission for its failure to utilize a formal model. Although Koopman's 1947 characterization of the work as 'measurement without theory' is a misrepresentation of the Mitchell–Burns programme, there can be no doubt that Burns and others at the bureau were sceptical of what might be achieved by the econometric methods being pioneered at Cowles. Also at this time Burns was engaged in a criticism of Keynesian economics as represented by Alvin Hansen. For Burns, Keynesian theorizing was too speculative and not sufficiently well grounded empirically (Burns 1946).

The period from the late 1940s through to the mid-1960s was a mixed time for the bureau. Some excellent projects were undertaken. Milton Friedman and Anna Schwartz began their work on US monetary history in 1948, a project that took until 1963 to publish. Friedman did other important work, particularly on consumption theory. Abramovitz worked on inventories and business cycles. George Stigler, who had joined the bureau staff in1943, worked on output and employment trends. Geoffrey Moore refined the system of leading indicators for business cycles, and Morris Copeland developed the analysis of money flows, later to become flow of funds accounts. All the same, the focus of the bureau's efforts had become less sharp; it was conducting much work of lesser value, and running into considerable financial difficulty. Once Willits left Rockefeller, those at Rockefeller were not so sympathetic to the bureau's plight. With the exception of a programme on international economic relations, Rockefeller declined to continue funding the NBER, and in 1958 the bureau turned to the Ford Foundation. Ford established a review committee of Gardiner Ackley, Richard Ruggles, and George Stocking. They criticized the bureau, but recommended that Ford provide funding, which they did. This allowed the bureau to continue, with relatively few changes until 1965. The research conducted over this period covered a wide range of projects that were loosely grouped into the categories of economic fluctuations, economic growth, wages and other incomes, the economic impact of government and international economic relations.

In 1965, Solomon Fabricant retired as Director of Research and was replaced by Geoffrey Moore, which was seen by many as a decision by the bureau to stay pretty much on its existing track. At the same time, Ford embarked on a major review of the bureau, again with a committee, but this time consisting of Emile Despres, R.A. Gordon, Lawrence Klein, Lloyd Reynolds, Theodore Schultz, George Shultz and James Tobin. This committee was sharply critical of the bureau, its leadership, project selection and research methods. Burns resigned as President and was replaced by John Meyer of Harvard. Meyer took over many of the functions previously held by the Director of Research, created two Vice Presidents of Research, and reorganized the bureau's efforts into specific programmes under their own Directors. Meyer also shifted the focus of the bureau's research into a number of new areas of social policy importance such as urban economics, health, human resources, education, environmental standards, the economics of the family, and crime and punishment. A number of important NBER studies were published during Meyer's term on subjects such as these by Theodore Schultz, Gary Becker, William Landes, Jacob Mincer and Victor Fuchs. Work on cycles was carried on, but no longer using the older NBER methods (Rutherford 2005).

Meyer left the bureau in 1977 and was replaced as President by Martin Feldstein, also of Harvard. Feldstein has remained as President except for a few years when he was with the Council of Economic Advisors (1982–4), and Eli Shapiro took over. Feldstein brought about further changes at the bureau, doing away with the senior research staff employed directly by the bureau, and changing the bureau into an organization designed to promote and coordinate research being conducted by university-based 'research associates' funded largely by National Science Foundation and other research grants. This rearrangement vastly increased the bureau's involvement with the larger academic community.

The focus has remained on empirical and policy-related research. Feldstein added programmes on issues such as aging, and asset pricing, and reinvigorated the NBER programmes on macroeconomics and on taxation. As of 2007, the NBER lists 17 major research programmes each involving 20 or more NBER research associates and each with its own director(s). These include aging, asset pricing, children, corporate finance, education, economic fluctuations and growth, health, industrial organization, international finance, labour, law and economics, monetary economics, productivity and public economics. In addition are smaller working groups working on another 16 topics from behavioural finance to the Chinese economy. The Conference on Income and Wealth also continues. Details of these programmes, those involved, and their publications can be found on the NBER website. The NBER's Research Associates now number about 600, and the NBER working paper series is a major research outlet. Links to the original NBER emphasis on measurement and business cycles are still to be found, however, notably in the NBER's data collection and in the Business Cycle Dating Committee.

See Also

- ▶ Burns, Arthur Frank (1904–1987)
- ▶ Business Cycle Measurement
- ▶ Kuznets, Simon (1901–1985)
- ▶ Mitchell, Wesley Clair (1874–1948)
- ▶ National Accounting, History of

Bibliography

Burns, A.F. 1946. Economic research and the Keynesian thinking of our times. *Twenty sixth annual report of the National Bureau of Economic Research*. New York: NBER.

Burns, A.F., and W.C. Mitchell. 1946. *Measuring business cycles*. New York: NBER.

Committee of the President's Conference on Unemployment. 1923. *Business cycles and unemployment*. New York: McGraw Hill.

Committee on Recent Economic Changes of the President's Conference on Unemployment. 1929. *Recent economic changes in the United States*. New York: McGraw Hill.

Fabricant, S. 1984. Toward a firmer basis of economic policy: The founding of the National Bureau of Economic Research. www.nber.org/nberhistory/sfabricantrev.pdf

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

Koopmans, T.C. 1947. Measurement without theory. *Review of Economic Statistics* 29: 161–172.

Kuznets, S. 1934. *National income, 1929–1932*. New York: NBER.

Kuznets, S., and M. Friedman. 1939. *Incomes from independent professional practice, 1919–1936*. New York: NBER.

Mitchell, W.C. 1927. *Business cycles: The problem and its setting*. New York: NBER.

Mitchell, W.C., W.I. King, F.R. Macaulay, and O.W. Knauth. 1921. *Income in the United States: Its amount and distribution, 1909–1919, part 1, summary*. New York: NBER.

Mitchell, W.C., W.I. King, F.R. Macaulay, and O.W. Knauth. 1922. *Income in the United States: Its amount and distribution, 1909–1919, part 2, detailed report*. New York: NBER.

National Bureau of Economic Research. Online. Available at: http://www.nber.org. Accessed 4 May 2007.

Rutherford, M. 2005. 'Who's afraid of Arthur Burns?' The NBER and the foundations. *Journal of the History of Economic Thought* 27: 109–139.

National Debt

Barry Gordon

In its modern sense, national debt emerged first in Florence and other Italian city-republics of the 15th century. Thereafter, the practice spread throughout Europe and was taken up by leading nation–states, including Spain, France and Holland. In England there were moves towards a more orderly system of public borrowings after the advent of William of Orange in 1688. The first permanent arrangements were introduced in 1715 (Dickson 1967). Assumption of state debts and establishment of related provisions for funding were undertaken by the Federal Government in the United States in 1790, with Alexander Hamilton the principal architect of the structure (Kimmel 1959).

Historically, the most common justification for incurring additional national debt is the sudden onset of fiscal emergency because of war. This has not been the only rationale, however. Additional debt has been undertaken in aid of national territorial expansion by peaceful means, as in the case of the Louisiana Purchase by the United States. As a device for financing public works it is sometimes claimed to have merits in terms of intergenerational equity. If government expenditures are used for projects which yield benefits for future generations, then it is appropriate that those generations help meet some of the costs involved.

Following the economic depression of the 1930s and the impact of the ideas of J.M. Keynes, other reasons were forthcoming for expansion of national debts. It was contended that public outlays derived from borrowing would create employment and stimulate growth in the private sector of the economy. Further, it was argued that whereas an external debt burdened a nation, its domestic debt might entail an internal redistribution of wealth but no necessary additional burden for the nation as a whole.

These latter grounds for larger national debts were revolutionary in terms of most of the popular and much of the professional opinion of the two preceding centuries. Over that period an array of arguments was marshalled in favour of a policy of national debt reduction at any and every available opportunity. Economists contributed to the array, their most influential contribution being the doctrine of the wages fund.

The leading economic argument for debt reduction was that such a measure would release additional funds for investment in productive activities in the private sector. As a result, wages and/or employment opportunities would increase and the rate of economic growth advance. Capital locked up in the public sector was capital wasted. It was also contended that debt reduction would improve the economic welfare of wage earners by creating scope for lower taxes in the wake of decreased governmental interest obligations. A related point was that reduction would result in a redistribution of income favouring the less affluent sections of the community.

In popular debate these arguments were sometimes supplemented by the contention that government borrowing placed an unjust burden of debt repayment on future generations. It was also affirmed that public confidence in a government, and hence public credit, was enhanced if that government was seen to be serious about a policy of reduction. Further, those who expressed alarm at the size of national debts sometimes reasoned as if there was a direct analogy between individual or family debt and government debt. If an individual or family went into debt, it was a sign of extravagance or mismanagement. The same was true of government. Those who reasoned in this fashion seem rarely to have extended the analogy to include business firms, especially if they were large ones.

As the foregoing survey suggests, if professional issues concerning financial techniques are put aside, then the subject of national debt is mainly of interest in terms of what Joseph Schumpeter called ‘economic sociology’. Judgement and advocacy have generally played greater roles in debate than has the application of systematic economic analysis (Schumpeter 1954, p. 327). However, the subject has been of deep concern to some prominent economists, particularly in the first half of the 19th century.

Almost all of the leading British classical economists were opposed to the maintenance of a national debt. In fact, ‘dismal’ predictions by political economists concerning the effects of such debt preceded the subsequent gloomy forecasts based on Malthusian population doctrine and the role of diminishing returns in agriculture. Adam Smith helped establish the mood when he prophesied that ‘the enormous debts’ of his time, ‘will in the long-run probably ruin all the great nations of Europe’ ([1776], 1937, p. 863).

David Ricardo shared Smith’s forebodings, and he seriously jeopardized the makings of a promising parliamentary career with a radical proposal for a once-and-for-all discharge of the existing British debt (Gordon 1976). According to Ricardo, the debt which had been accumulated in the wars with Napoleon, ‘destroyed the equilibrium of prices, occasioned many persons to emigrate to other countries in order to avoid the burden of taxation which it entailed, and hung like a mill-stone round the exertion and industry of the country’ (Hansard 1819, 1022–4). This sentiment was not shared by Thomas Robert Malthus and Lord Lauderdale. They were both concerned about the maintenance of an adequate level of demand in the economy and warned of the dangers inherent in too rapid a retirement of debt. However, Malthus and Lauderdale were in the minority, and most economists favoured at least some reduction in the debt as part of a programme of stringent budgetary economies. Particularly influential in this latter respect was Sir Henry Parnell, who became chairman of the Finance Committee of the Commons and published a work entitled *On Financial Reform* (1830) which had considerable impact (Hilton 1977; Gordon 1979).

Through the second half of the 19th century the British national debt was diminished gradually and the subject lost much of its significance for economists in that country. In America during the 1860s the situation was different in that the Civil War entailed the accumulation of a debt of almost $2.8 billion by 1866. This greatly alarmed some economists, including Amasa Walker, who displayed a Ricardo-like zeal in his protestations concerning the evil effects of this burden on the economy. By contrast, Henry C. Carey was not alarmist. Carey was opposed to rapid reduction of the debt because of the weight of taxation which this would involve (Kimmel 1959).

During that era of economic thought known as ‘neoclassical’ the study of public finance took shape as a distinct specialization within economics. Leading early treatises were C.F. Bastable, *Public Finance* (1892) and Henry Carter Adams, *The Science of Finance* (1898). This development created a new professional context for discussion of issues surrounding national debt. The context encouraged greater attention to the financial techniques involved and discouraged tendencies to adopt strong pro and anti stances on the principle of maintaining the device. This latter may help explain the relatively sober approach within professional ranks to the problem of increased indebtedness after World War I.

The subsequent influence of the revolutionary ideas of J.M. Keynes has been remarked above. However, it is important to appreciate that since the late 1950s there has been a notable revival of interest among economists in questions concerning the economic implications of national debt. That revival is attributable, in part, to community unease with the increasing absolute size in money terms of the public debts of many countries. Another factor has been a decreased confidence in the adequacy of assessments of the significance of national debt from the perspective of Keynesian macroeconomics. A third element is the renewal of attachment to atomistic, liberal ideology within sections of the economics profession.

See Also

- ▶ Burden of the Debt
- ▶ Public Debt
- ▶ Ricardian Equivalence Theorem

Bibliography

Dickson, P.G.M. 1967. *The financial revolution in England: A study in the development of public credit, 1688–1756*. London: Macmillan.

Gordon, B. 1976. *Political economy in parliament, 1819–1823*. London: Macmillan.

Gordon, B. 1979. *Economic Doctrine and Tory Liberalism, 1824–1830*. London: Macmillan.

Hansard, T.C. 1819. *Parliamentary debates*, vol. 40. London.

Hilton, B. 1977. *Corn, cash, commerce: The economic policies of the Tory Governments, 1815–1830*. Oxford: Oxford University Press.

Kimmel, L.H. 1959. *Federal budget and fiscal policy, 1789–1958*. Washington, DC: Brookings Institution.

Schumpeter, J.A. 1954. *History of economic analysis*. New York: Oxford University Press.

Smith, A. 1776. In *An inquiry into the nature and causes of the wealth of nations*, ed. E. Cannan. New York: Random House.

National Income

Thomas K. Rymes

Abstract

This article emphasizes how classical, neoclassical and real Keynesian economic theories are related to accounts of national income and its distribution. The more traditional parts of the analysis focus on rates of growth, capital accumulation and real net rates of return to capital, factoral distributions of income, and capital-theoretic problems in constructing matching national income accounts. More modern neo-Keynesian and monetary approaches are examined to account for theoretical roles played by money and banking in determining output, national income and technical progress. The effects of measures of banking output on modern national income accounts are stressed.

Keywords

Capital gains and losses; Capital theory; Central banking; Classical economists; Depreciation; Distribution of income and wealth; Factoral distribution of income; Friedman, M.; Hicks, J. R.; Hulten, C.; Human capital; Keynesian revolution; Kuznets, S.; Leontief, W.; Meade, J. E.; Measurement of capital; Monetary theory; National accounting; National income; National income accounting; Neo-Ricardianism; Net rates of return; Obsolescence; Output of banks; Quantity theory of money; Returns to human capital; Ricardian equivalence theorem; Stationary state; Stone, J. R. N.; Sustainable consumption; System of National Accounts; Technical change; Unemployment

JEL Classifications

D4; D10

Comprehensive systems of national accounts consist today of traditional national income, expenditure and product accounts, input output or production accounts, financial transactions and revaluation accounts (Rymes 1992) and national balance sheets. While many parts of this modern system are expressed in current and constant prices, national income, its factor and individual income distributions are meaningfully expressed only in current prices. Constant price, or 'quantity', indexes are used to measure 'real' expenditures over time and across nations, in productivity studies both partial and for all factors again over time and across industries and countries (see Erwin W. Diewert's contributions in ILO 2004 and IMF 2004). Indeed, much of modern economic history can now be written in terms of the nominal and real economic accounts over time.

Yet, to date, no one has put together a comprehensive examination of the whole accounting system seen from a particular set or sets of economic theory. Theorists, such as J.R. Hicks, Richard Stone, Wassily Leontief and James Meade, and quantitative economic historians such as Simon Kuznets have made notable contributions to national accounting and have been so recognized with Nobel Prizes. The general lack of emphasis on the connection with economic theory, however, causes the poor student of economics to find the structure of the official accounts a bewildering maze of 'uses and resources', which seem more the product of much worthwhile international compromise than the development of the accounts from basic principles of economic theory. Anyone who has tried to teach economics students with the assistance of the 1993 System of National Accounts (SNA 1993, Washington, DC.; Commission of the European Communities; International Monetary Fund; OECD; United Nations; and the World Bank (*sic*)) will not find in all the bureaucratic compromises of admittedly needed reconciliation and international comparisons those flashes of illumination which economic theories can give. A recent OEDC publication (Blades and Lequiller 2006) further illustrates dangers of the lack of economic theory. It never adequately explains the economic meaning behind consumers 'real' expenditures and producers 'real' outputs making up GDP, though such knowledge must be held if the reader is to understand the very useful warnings about ‘real shares’ and additivity problems associated with index numbers. Thus it is sad to read one of the best practitioners of national accounting today asserting ‘... the conceptual foundations of the present model of the national accounts are being progressively undermined by the shifting quicksand of economic theory...’ (Ward 2006, p. 327). Of course, Ward describes other eroding forces, but to give economic theory priority of place in conceptually undermining the accounts seems to me an error resulting from a despairing denigration of economic theory.

I concentrate here on how economic theory contributed to and conditioned national income accounting developments and to some extent how problems in constructing national accounts condition good economic theory. The central theme of this article, then, is the interplay between economic theory and national income accounting. Modern readers, especially students, once they see the interconnection between the accounts and economic theory, should find the national accounts as fascinating and exciting as I do and will each become, I hope, a ‘... passionate accountant’ (Lathen 1974, p. 183).

Classical and Neoclassical National Income Theories

David Ricardo argued the principal problem of political economy was the determination of the laws governing the distribution of national income among the classes of society (Ricardo 1971, vol. 1, p. 5). His question was a major concern of classical economic theorists and it has returned to some pre-eminence among economists today (Milanovic 2005). Consider the following set of extremely simple national income and expenditure accounts set out for a market economy to examine classical economic theory.

Incomes		Expenditures
WL		$P_C C$
$RP_K K$		$P_{K\Delta} K$
$RP_N N$		
$D_N P_N N$		
$D_K P_K K$		
Y	$\equiv$	E

where National Income (Y) is shown as identically equal to National Final Expenditures (E) or Product.

Examining the accounts for one country among many, one must distinguish between National Income and Domestic Product whereas, of course, World Income (WI) and World Expenditure or Product (WP) will be the same. Some economists regard the Domestic Product concept as more useful since it extracts from effects of the international redistribution of returns to capital. (For a contrary opinion, see Beckerman 1987.) More technical but telling objections can be raised against the Domestic Product concept when it is expressed in constant price terms in a world experiencing technical change in which international trade takes place in intermediate inputs of production.

Why however, does Y identically equal E? If we imagine the accounts were for an even simpler world where there was no capital, then the equality among the circular flows would be clear. Owners of labour would sell their time to producers and the value of their expenditures for the goods produced would cover the cost of the producers. For an extensive discussion of circular flows and the crucial capital-theoretic problems in national accounting, see Hulten (2006).

The notation involves the income of workers (WL), with W the set of money wage rates and L the corresponding set of the working times (hours, days, and so on) offered and demanded by the suppliers and demanders of labour; $RP_N N$ is the net rents earned by the natural agents of production, which, for illustrative purposes, we shall take mainly to be the inalienable and inexhaustible powers of the soil, where R is net rates of return, P_N is prices of the stocks of land so that RP_N is the net rents on the stocks of land (N); and RP_K is rentals earned by the stocks (K) of reproducible capital goods like machines, inventories and buildings. Inanimate things like land and capital goods earn nothing by themselves, and clearly what the classical economists had in mind when then they wrote of the factoral distribution of income was that the net rents on land were garnered by landowners for their husbandry, and the net rents being earned by capital were the net flow of income being earned by the owners of the capital goods, capitalists playing their rentier roles as savers and holders of the stock of capital in the economy. By the ‘factoral distribution of income’ classical economists meant the distribution of income among people, aggregated as the classes of society: labourers, landlords and capitalists. When it is borne in mind that the classical economists also saw labour, land and capital as factors of production, it can be clearly seen that classical theoretical economics was an immensely great scientific undertaking, one which still echoes throughout economics today.

The notation D_NP_NN and D_KP_KK refers to the rates of depletion or exhaustion of natural agents of production, such as the using up of pools of oil, which do not apply to our simple theoretical case of N being Ricardian land. Nor is there any discussion here of the rate of degradation of the environment capital (see Rymes 1991). Very importantly, D_KP_KK refers to the rates of depreciation or using up of capital in production.

On the Expenditure side of the accounts, P_CC is the values of the final consumption of the society, which, to many economists, is the be all and end all of economics. $P_{K\Delta}K$ represents the values of the gross capital formation taking place in the society. It is gross in that no allowance is taken of the fact that the new capital goods being produced may or may not be sufficient to replace the wear and tear on existing capital goods. Y and E refer then to Gross National Income and Expenditure respectively.

One of the major theoretical problems in contemporary theory and classical and contemporary national income accounting is the meaning of capital and the conception and measurement of ‘maintaining capital intact’. Even today, despite advances in accounting and economic theory, it is difficult if not almost impossible empirically to measure well the ‘wear and tear’ on capital in modern economic systems. Where depreciation arises from obsolescence, so severe are the problems of measurement that almost all economists today use Gross Domestic Income (Product) or Expenditure as the principal aggregate for economic analysis. National income analysis, then, is greatly hampered by the fact that good estimates of capital consumption and the depletion of natural agents of production, again to say nothing of the degradation of the environment, are generally not available.

If we did have such estimates, the National Accounts just set out could be revised further to appear as.

Incomes		Expenditures
WL		PcC
RP_KK		$P_K(G - D)K = P_KnK$
RP_NN		
Y*N	≡	E*N

where $P_K(G - D)K = P_KnK$ is net capital formation, with n being the rate of growth so that one would be able to see how important net returns were to capital in net national income, which also in this case is said to measure ‘sustainable’ consumption.

The importance of the capital problem extends to the measurement of labour income as well. Today, wages are paid not so much for the application of pure labour time but for the services of the human capital accumulated by the individuals through expenditures on education, health and even the raising of families. On such capital expenditures, though there is a direct link between the forgoing of present consumption and the accumulation of capital by the individuals, the difficulties of measuring the depreciation on intangible human capital in the so-called knowledge economies are as bad as, if not worse than, those for physical capital. Yet the problem of measuring the returns to human capital gripped the classical economists as well.

One could argue that the consumption of the workers was not final at all, but was perhaps just sufficient to maintain the labour force either at a particular level or at a certain growth rate. Suppose we could extend all of the capital measurement thinking previously outlined to the classical and modern neoclassical treatment of labour. We could write off the consumption of the workers as required inputs into the maintenance of the labour force. Much of PC would vanish along with

WL. The above accounts could be then even further dramatically reduced to

Incomes		Expenditures
RP_KK		PcC^*
$RP_N N$		$P_K(G-D)K = P_KnK$
$Y^{**}N$	≡	$E^{**}N$

where PcC* is the consumption of the capitalists (and landholders). The extreme classical Ricardian stationary state comes into focus, where the economy is said to have converged to a position where savings and accumulation have been pushed to the point where R, the net rates of return, *are* positive but so low that net savings and the rate of growth of net capital stock and national income, n, would be zero.

Though classical economists were aware that capital accumulation was unlikely to occur in given states of technology, the modern treatment of technical progress is to assume that it serendipitously occurs or, more interestingly, is an endogenous function of the rate of capital accumulation. If, however, technical progress were steadily occurring, then the long-period equilibrium of modern classical analysis and *theory* comes into view. If we ignore land and landholders, and if the consumption of the capitalists were some function of their income and the rate of return so that $PC^* = c((R), RPK)$, then national income for steady growth, the modern variant of the Ricardian stationary state, becomes

Incomes		Expenditures
$RPK- c((R)RPK)$	≡	$Pn'K$
or $(1- c((R)RPK))$	≡	$Pn'K$
$s(R)R$	≡	n'

that is, the economy may be said to have converged to an equilibrium where rates of return to capital exceeds the rate of growth of the income of the economy arising from technical progress, n′, if the fraction of returns to capital saved, s, is less than 1. If one assumes that the rate of technical progress is a function of R, then the whole structure of the classical and neoclassical national income accounts can be boiled down to reflect basic theories

$$S(R)R = n'(R(R^*(R)).$$

where the net rates of return to capital, the intertemporal prices in modern economies, are seen by the simplest accounts to be a function of the rates of saving, or intertemporal choice, and rates of technical change, itself the product of investing and expected rates of return, R*, themselves seen as some function of R. Thus, we see that, when asking questions about the distribution of national income, the national accounts can be set out to illuminate the forces of growth which play vital roles in determining national income. It can also be seen that Ricardo's question about the determinants of the factoral distribution of national income lies at the very heart of modern economic analysis, of both the neoclassical and neo-Ricardian growth varieties (see Barro and Sala-i-Martin 1995, in particular the chapter on growth accounting; and Pasinetti 1995). While economic theories may be said to generate the accounts designed to illuminate them, we have seen that they also illuminate the great theoretical difficulties and aggregation problems associated with Professor Hulten's questions about capital theory.

Readers should please note that I am largely by-passing the *severe* capital- theoretic difficulties alluded to by him. One of Hulten's observations that '... all aspects of capital ultimately are derived from the decision to defer current consumption in order to enhance or maintain expected future consumption' (2006, p. 195) means that capital is not a factor of production independently of the 'willingness to wait' and that multifactor productivity advance should be conceived as the improvement in the efficiency of working and waiting, n′, rather than an improvement in the efficiency of labour and capital. The deep theoretical questions involved in measuring capital, the growth of nations and the aggregation questions may be resolved to some extent by the application of Leontief's disaggregated production and capital accumulation accounts (see Cas and Rymes 1991; Rymes 1997).

Keynesian Theory

The Keynesian revolution clashed with classical and neoclassical theories and led to some of the modern ‘advances’ in national income accounting. Indeed, some national accountants argue that, partly as a result of Keynes and other theorists such as Jan Tinbergern, modern national accounting started in the 1930s (Bos 2003, 2006). At the same time economic theory started paying increased attention to institutional forms such as corporations and governments. Under these influences, our simplified national accounts now appear as

Incomes		Expenditures
WL		PCC
Ω		$P_{K\Delta}K$
Y	≡	E

where the net returns to capital and net rents on natural agents of production are largely replaced by corporate profits, Ω, which generally have measures of depreciation of limited economic meaning, and may or may not well reflect the distribution of interest to bondholders and dividends to shareholders with almost certainly no account being taken of capital gains and losses, and where the switch away from national income to gross national product reflects concern with unemployment rather than the level and the distribution of national income. When the revaluation accounts are added to the standard income accounts, theory again comes to the forefront.

Suppose that modern corporations distribute none of the profits or returns to capital they earn as dividends to their shareholders, ignoring for simplicity the payment of interest to bondholders, but reinvest their profits in the acquisition of capital goods for their firms. The value of the shares held by shareholders (and bought and sold among them) rise along with increases in the corporate stock of capital. It would appear from the national accounts as if the corporations did the saving whereas they may be used to test theories which have the corporations as mere intermediaries, whose investment decisions reflect the wishes of their shareholders.

The neo-Ricardian and Keynesian theories can be put together for the determination of not just the level but also the distribution of national income. If good estimates of the wear and tear on capital are available, one can revert from gross to net income and develop arguments addressed to the question of whether corporate firms and governments can affect the level and the distribution of national income. Here the national accounts can contribute to our knowledge of the extent to which individual households can be said to ‘see through’ corporate firms and governments in such matters as the Ricardian equivalence theorem (see Gillespie 1980, 1991). To do this, the accounts must be prepared with the various theories of institutional forms in mind; otherwise they may be dismissed with some derision by contemporary theorists (Prescott 2006).

When the personal distribution of national income is considered, national income accounts must be supplemented by longitudinal surveys of the distribution of income and wealth among individuals and families, the latter of which can be taken as representing constellations of individuals through time. Here again the theory of why certain families have such time preferences as to permit them to form dynasties requires much work if national income is to be so disaggregated so that those forces playing upon it may be extended to portray and explain individual and dynastic distributions of income and wealth.

Controversies Among Modern Monetary Theories and National Accounting

Recent developments in monetary theory present great challenges to national accounting. Some monetary theories, those based fundamentally on the quantity theory of money, assert that once-over changes in ‘costless’ fiat money cannot have effects on such real phenomenon as national income, whereas continuous changes in such monies, affecting continuous changes in price indexes, may have rather dramatic effects. Yet, as national balance sheets and wealth accounts show, outside fiat monies are becoming increasingly marginal. How is national income affected by these matters?

National income reflects differences in the underlying classical, neoclassical and Keynesian theories. Keynesian models of unemployment rest upon the empirical and theoretical unimportance of outside or fiat money. Friedman argues, against the Keynesian position, that with real capital gains (losses) accruing to holders of money because of Keynesian disequilibria, real national income will tend to equilibrate at classical economic levels. Thus, if money wage rates and prices are falling because of unemployment, then, according to Friedman, the real income of people, holding given amounts of outside fiat money, will be positive, and will rise faster and faster and become bigger and bigger the more quickly prices fall, thus causing the unemployment to vanish even if there were some adverse effects on expenditures while the deflations were going on (Friedman 1976, pp. 319–21). As monetary economies are characterized by less and less outside or fiat money, the less and less important is the Friedman counter to Keynes. The question which must be asked is this: is it meaningful to introduce capital gains and losses associated with deflations and inflations and the holding of fiat money into the revaluation accounts associated with national income estimates when, under modern monetary and central banking theory, such holdings, at least in the form of reserves with central banks, are vanishing?

The basic problem with the current national accounts is that we do not have meaningful measures of the output of private banks nor, even more importantly, of the output of central banks. If we applied the current method of imputation for the output of banks to modern central banks, their output would be seen to be zero (Rymes 2004). Since the banks are the principal producers of transactions services and affect monetary production technologies, it follows that the inability of the national accounts to arrive at satisfactory measures of the output of banks in general means that they cannot measure satisfactorily production in monetary economies (see Fixler and Reinsdorf 2006). Thus, though one of the central questions dividing Keynesian and neoclassical analyses and the effects of monetary developments on the concepts and measures of national income cannot be currently understood using the current national income accounts, even deeper questions emerge. Does the growth of banks and central bank policies affect capital accumulation, technical progress and national income? We simply do not know now!

Conclusion

The national income accounts have played central roles in the development of economic theory and analysis. Concepts and measures must be improved and developed to reflect better the fact that we live in monetary economies where we do not understand and do not accordingly measure well the outputs of banks and central banks, capital inputs, accumulation and technical progress, all which affect the distribution of national income. Ricardo's question still needs answers. Our current theories and measures of national income need work. Readers and students should therefore realize that there is much exciting and profitable theoretical and empirical study remaining to be done in national income accounting.

See Also

► National Accounting, History of

Acknowledgments I am indebted to Dr. Harry Postner, with whom I collaborated in preparing a paper given at a session we organized on 'National accounts and the teaching of economics' at the 1998 meetings of the International Association for Research in Income and Wealth (IARIW), for his continued criticisms of my work. The most dispiriting finding of that session was how few departments around the world maintained courses in national accounting as part of their curricula. I think the reason is that official national accounting places too little emphasis on connecting with modern economic theories and accordingly lose too much academic contact. I alone am responsible for this basic theme in this article. I also record with much gratitude the extremely critical and penetrating comments on the initial draft made by Professor Duncan McDowall of the history department at Carleton. Professor McDowall took valuable time away from writing a history of the Canadian national accounts to comment on my manuscript. I am also much obliged to Professor Mark Bils whose suggestion for tightening the monetary part of the argument was useful indeed. Finally, I thank Michael James for making my article far more readable than I could.

Bibliography

Barro, R., and X. Sala-i-Martin. 1995. *Economic growth*. Toronto: McGraw Hill.

Beckerman, W. 1987. National income. In *The New Palgrave: A Dictionary of Economics*, ed. J. Eatwell, M. Milgate, and P. Newman, vol. 3. Toronto: Macmillan.

Blades, D., and F. Lequiller. 2006. *Understanding national accounts*. Paris: OECD.

Bos, F. 2003. The national accounts as a tool for analysis and policy: Past, present and future. Ph.D. thesis, Twente University.

Bos, F. 2006. The development of the Dutch national accounts as a tool for analysis and policy. *Statistica Neerlandica* 60: 225–258.

Cas, A., and T.K. Rymes. 1991. *On concepts and measures of multifactor productivity in Canada, 1961–81*. Cambridge: Cambridge University Press, 2006.

Fixler, D., and M. Reinsdorf. 2006. Computing real bank services. A paper prepared for the NBER/CRIW Workshop, 18 July.

Friedman, M. 1976. *Price theory.* Chicago: Aldine.

Gillespie, I. 1980. *The redistribution of income in Canada*. Ottawa: Carleton Library.

Gillespie, I. 1991. *Tax, borrow and spend: Financing federal spending in Canada 1867–1990*. Ottawa: Carleton University Press.

Hulten, C. 2006. The 'architecture' of capital accounting: Basic design principles. In *A new architecture for the U.S. National Accounts*, ed. D.W. Jorgenson, J.S. Landefeld, and W.D. Nordhaus. Chicago: University of Chicago Press.

ILO (International Labour Organization). 2004. *Consumer price index manual: Theory and practice*. Geneva: ILO.

IMF (International Monetary Fund). 2004. *Producer price index manual: Theory and practice*. Washington, DC: IMF.

Lathen, E. 1974. *Accounting for murder.* Richmond Hill: Pocket Books.

Milanovic, B. 2005. *Worlds apart: Measuring international and global inequality.* Princeton: Princeton University Press.

Pasinetti, L. 1995. *Structural change and economic growth*. Cambridge: Cambridge University Press.

Prescott, E.C. 2006. The transformation of macroeconomic policy and research. *Journal of Political Economy* 114: 203–235.

Ricardo, D. 1971. On the principles of political economy and taxation. In *The works and correspondence of David Ricardo*, ed. P. Sraffa, vol. 1. Cambridge: Cambridge University Press.

Rymes, T.K. 1991. Some theoretical problems in accounting for sustainable consumption. In *Approaches to environmental accounting*, ed. A. Franz and C. Stahmer. Heidelberg: Physica.

Rymes, T.K. 1992. National accounting and financial flows. In *The new Palgrave dictionary of money and finance*, ed. P. Newman, M. Milgate, and J. Eatwell, vol. 3. Macmillan: London.

Rymes, T.K. 1997. The productivity of working and waiting. In *Capital controversy: Post-keynesian economics and the history of economic thought*, ed. P. Arestis, G. Palma, and M. Sawyer. London: Routledge.

Rymes, T.K. 2004. Modern central banks only have *real* effects. In *Central banking in the modern world: Alternative perspectives*, ed. M. Lavoie and M. Seccareccia. Cheltenham: Edward Elgar.

Ward, M. 2006. An intellectual history of national accounting: A review of André Vanoli. *A History of National Accounting. Review of Income and Wealth* 52: 327–340.

National Leadership and Economic Growth

Benjamin F. Jones

Abstract

Recent empirical analysis suggests that individual national leaders can have large impacts on economic growth. Leaders have the strongest effects in autocracies, where they appear to substantially influence both economic growth and the evolution of political institutions. These findings call for increased focus on national economic policies and the means of leadership selection, among other issues.

Keywords

Leadership; Growth; Institutions; Policy; Political economy; China

JEL Classifications

O11; O43; P16; F52

In the large literature on economic growth, the role of national leaders has received relatively little attention. Yet the imperative for such work is increasing: recent empirical evidence suggests substantial roles for individual leaders in explaining national economic growth as well as national institutional change, which can further influence the growth environment. This article considers the case for studying growth from a leadership perspective, reviews the primary econometric evidence, and discusses open questions.

Why Study Leadership?

To frame this question, first consider two opposing views of individual leaders in historical reasoning. At one extreme, the 'Great Man' view of history, classically associated with Carlyle (1837), interprets major events largely as consequences of the idiosyncratic actions of a few individuals. At the opposite extreme, classically associated with Tolstoy (1869) and Marx (1852), individual leaders play little or no role; rather, historical events are understood much more deterministically as the contest of broad social and technological forces. This latter view gained substantial traction in the 20th century throughout the social sciences. The apparent inevitability of the First World War and Butterfield's (1931) condemnation of earlier historical reasoning promoted the new paradigm, in which individual leaders would play muted roles. Modern theoretical implementations have provided potentially decisive constraints on leaders through median voter theory (Downs 1957). More broadly, the presence of 'veto players', through opposing political parties or the checks and balances of multiple institutions, can be seen to severely limit an individual leader's actions (Tsebelis 2002).

The literature on economic growth has progressed mostly within this 20th-century paradigm. Examinations of the fundamental causes of growth debate between institutions, culture, and geography, which typically operate without reference to the actions of particular personalities. While policy analysis also features in the growth literature, and some growth economists may imagine leaders indirectly as policymakers, leaders themselves are rarely the subject of focus. As one metric, the Web of Science shows that the keywords 'economic growth' intersected with 'property rights', 'international trade', or 'sub-Saharan Africa' produce hundreds of papers each since 1955, while the intersection of 'economic growth' with variants of 'national leadership' produces only three papers.

Nonetheless, there are several reasons that leadership may be an important object of study in a growth context.

Institutional Constraints are Incomplete

The constraints imposed on leaders from electoral pressures, opposition parties, independent legislatures and judiciaries all vary across countries. Autocracy, where these constraints are weak, is a common form of political organization. More generally, the modern growth literature has emphasized how the 'rules of the game' vary across countries, and that institutional differences can be powerful sources in explaining different development paths (see, e.g., Acemoglu et al. 2005). To the extent that the authority embedded in formal institutional rules and the authority embedded in individuals act as substitutes, the increasing visibility of institutional variation in explaining development paths may directly motivate leadership studies.

Classically, Weber's theory of leadership suggests just this point: leaders can have substantial influence, but only when other institutions are weak (Weber 1947). In a modern theoretical context, information asymmetries, commitment problems and limited liability all suggest agency for individuals that may be substantial depending on the local rules of the game. In a modern empirical context, several studies have demonstrated leader agency in sub-national political environments (e.g. Besley and Case 1995; Kalt and Zupan 1984; Levitt 1996), and in corporate environments (e.g. Johnson et al. 1985; Bertrand and Schoar 2003).

Theory Suggests Numerous Roles for a National Decision Maker

Theories of economic growth that emphasize public goods (such as infrastructure, education and health), national policies (such as international trade and monetary policy), or national-scale complementarities (for example, big push mechanisms) all suggest possibly important roles for a national leader. Furthermore, the capacity of leaders to make war or to pursue systematic corruption suggests other means of economy-wide influences.

Economic Growth has Substantial Medium-Run Volatility

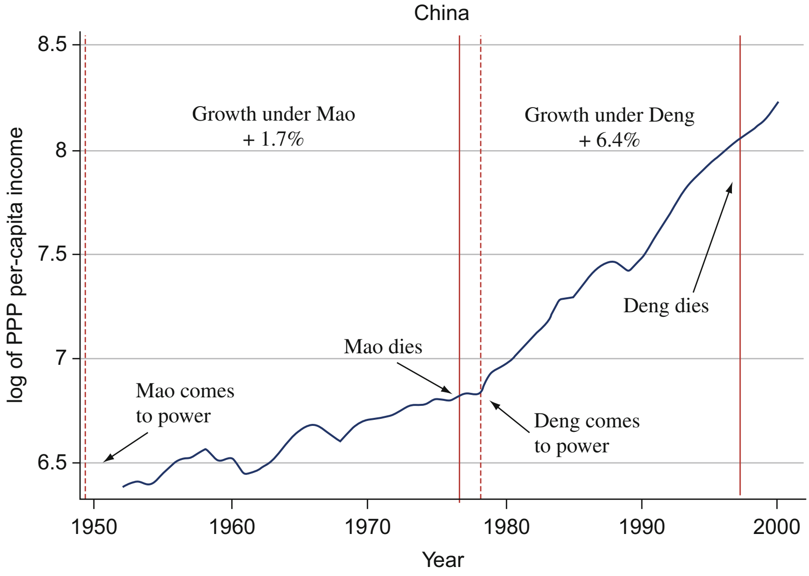
Empirically, economic growth within countries is extremely volatile, with one decade's growth rarely looking much like growth the decade before. The correlation in mean growth across consecutive decades within countries averages only 0.3 in the world sample (Easterly et al. 1993) with countries regularly experiencing substantial medium-run growth accelerations and growth collapses (Hausmann et al. 2005; Jones and Olken 2008). To explain such volatility, it is natural to look at influences that change at appropriate frequencies. National leaders, who change sharply and at relevant time scales, are one place to look.

The Empirical Evidence: Do Leaders Matter?

Identifying a causative effect of leaders on economic growth is challenging. Even if particular leaders and particular growth episodes are associated, it may be that growth changes drive leadership changes, without a causative effect of leaders. In fact, empirical evidence demonstrates that coups are less likely when growth is good (Londregan and Poole 1990) and that US presidents are less likely to be re-elected during recessions (Fair 1978).

Jones and Olken (2005) attempt to avoid this identification problem by examining cases where a leader's rule ends at death, through either natural causes or an accident. In these cases, the timing of the transfer from one leader to the next appears unrelated to underlying social and economic conditions. By examining all leader deaths since the Second World War, Jones and Olken (2005) test whether leaders have a causative impact on growth.

As one example, Fig. 1 presents the growth path for China from the Penn World Tables. The dashed vertical line indicates when a leader comes to power, and the solid vertical line indicates when the leader died. In China, we see that Mao's rule was closely associated with poor economic growth, averaging 1.7 per cent per year. After his death, growth averaged 5.9 per cent per year. The Cultural Revolution and the forced collectivization of agriculture were among many national policies that likely limited growth during Mao's rule, while Deng, who came to power in 1978, is often regarded as having moved China towards more market-oriented policies.



National Leadership and Economic Growth, Fig. 1 Growth in China under Mao and Deng

While the dramatic change in growth after Mao's death may suggest leader effects, this is one example and it could be a coincidence. Jones and Olken (2005) analyse all 57 cases of natural and accidental deaths in the world sample and test, on average, whether growth changes in an unusual fashion when leaders die. This approach rejects the hypothesis that leaders have no influence on growth. Moreover, the point estimates suggest substantial effects. Under the assumption that leader quality is independently drawn across leaders, one standard deviation of leader quality is associated with a 1.5 percentage point difference in the annual growth rate – a large effect.

An important additional finding is that leader effects are strongest in autocratic settings, especially in the absence of political parties or legislatures. Meanwhile, the hypothesis of no leader effects cannot be rejected in democratic settings. The findings are therefore quite consistent with Weber's theory of leadership, where leaders can matter substantially but only when they are unconstrained. These results point to an important intersection between institutions and individuals in understanding growth paths.

Further evidence about the relationship between individual leaders and political institutions is found in Jones and Olken (2009), which studies the effect of assassinations. That paper estimates the effect of assassination-induced leadership change by comparing cases where leaders were killed in assassination attempts with cases where leaders survived assassination attempts. The key identification assumption is that, conditional on a weapon being discharged in pursuit of killing a leader, whether the leader survives the attack can be treated as plausibly exogenous. The main finding with this approach is that the assassination of autocrats substantially increases the probability of democratization, with democratic transitions occurring at three times the background rate. Once again, the finding is limited to autocracies, with assassination of leaders in democracies provoking no institutional change.

Together, these findings suggest that institutions influence the impact of national leaders, and that national leaders can also influence the path of institutions. The constrained leader – the democrat – may have important degrees of agency, but at the level of national economic growth or the national political system, there is little evidence for an effect. The unconstrained leader – the autocrat – is seen as a powerful force in explaining the growth path, and a powerful force in the evolution of the political system.

Open Questions

If leaders matter to economic growth, then many further questions are raised. To close this article, I briefly consider some of the open issues.

Do leaders act merely to obstruct growth, or do they actively promote it? In one view, leaders are essentially destructive – highwaymen along the road to economic riches. Tendencies to steal, corrupt and make war are means through which leaders can adversely affect growth and may describe numerous leaders, such as Charles Taylor of Liberia and Mobutu Sese Seko of the former Zaire. In this view, economies would grow well in the absence of such interference. In another view, leaders can be actively good for growth – for instance by investing in public goods, choosing progrowth trade policies, or overcoming national-scale coordination problems. Lee Kwan Yew of Singapore might suggest such a view. Anecdotal assessments aside, whether leaders can be good, bad or both is an open empirical question.

Related questions of how leaders influence growth are intimately related to the role of national policies in explaining growth. Since leaders matter, the decisions they make – that is, their policies – appear to matter. (The converse is not true: policies might well matter even if leaders do not, if national policies are purely the expression of broader social forces.) While convincingly identifying key policies has proven difficult, and some authors doubt that national policy matters much (e.g. Easterly 2005), the findings of Jones and Olken (2005) motivate a renewed focus on policy choices. Put another way, the findings of substantial growth effects tied to individual leaders imply that growth is not purely deterministic but rather substantially within contemporary hands. While the empirical growth literature has had substantial success explaining worldwide income differences based on deep, historical determinants (such as institutional inheritances), the distant hand of history explains only a portion of the variance in modern incomes. When asking how to make poor countries rich, the unexplained, nondeterministic part of growth variation becomes especially relevant and, given the results about leadership, more within reach.

Additional questions surround the selection of leaders. Econometric studies have provided some lessons at the village and municipal level. Research in India (Chattopadhyay and Duflo 2004) exploits randomized reservations of village council seats for women to demonstrate that gender matters for the types of public goods provided. Research in Brazil (Ferraz and Finan 2008) employs regression discontinuity design across municipalities to demonstrate that higher wages attract greater numbers of candidates, more educated candidates, and electoral winners who fund more public goods. Much more work is needed along these lines, especially in autocratic settings. At the national level, it would be helpful to identify key observable characteristics that can separate good from bad leaders before their assumption of authority. A related subject is the design of institutional systems to produce the right kind of national leaders: in other words, institutional rules or other national features that attract well-intentioned, capable social planners rather than the simply vainglorious, or thieves. The door is open for creative empirical and theoretical explorations of these issues. Given the large effect that leaders appear to exert on economic growth, these more detailed questions become first-order subjects in understanding the growth process.

See Also

- ▶ Economic Growth
- ▶ Growth and Institutions
- ▶ Policy Reform, Political Economy of

Bibliography

Acemoglu, D., S. Johnson, and J. Robinson. 2005. Institutions as the fundamental cause of long-run growth. In *Handbook of economic growth*, ed. P. Aghion and S. Durlauf. Amsterdam: North Holland.

Bertrand, M., and S. Schoar. 2003. Managing with style: The effect of managers on firm policies. *Quarterly Journal of Economics* 98: 1169–1208.

Besley, T., and A. Case. 1995. Does political accountability affect economic policy choices? Evidence from gubernatorial term limits. *Quarterly Journal of Economics* 110: 769–798.

Butterfield, H. 1931. *The Whig interpretation of history*. London: G. Bell & Sons.

Carlyle, T. 1837. *The French revolution: A history*. London: Chapman & Hall.

Chattopadhyay, R., and E. Duflo. 2004. Women as policy makers: Evidence from a randomized policy experiment in India. *Econometrica* 72: 1409–1443.

Downs, A. 1957. *An economic theory of democracy*. New York: Harper & Row.

Easterly, W. 2005. National policies and economic growth: A reappraisal. In *Handbook of economic growth*, ed. P. Aghion and S. Durlauf. Amsterdam: North Holland.

Easterly, W., M. Kremer, L. Pritchett, and L.H. Summers. 1993. Good policy or good luck? Country growth performance and temporary shocks. *Journal of Monetary Economics* 32: 459–483.

Fair, R.C. 1978. The effect of economic events on votes for president. *Review of Economics and Statistics* 60: 159–173.

Ferraz, C., and F. Finan. 2008. *Motivating politicians: The impacts of monetary incentives on quality and performance*. University of California at Los Angeles, Mimeo.

Hausmann, R., L. Pritchett, and D. Rodrik. 2005. Growth accelerations. *Journal of Economic Growth* 10: 303–329.

Johnson, W.B., R. Magee, N. Nagarajan, and H. Newman. 1985. An analysis of the stock price reaction to sudden executive deaths. *Journal of Accounting and Economics* 7: 151–174.

Jones, B.F., and B.A. Olken. 2005. Do leaders matter? National leadership and growth since World War II. *Quarterly Journal of Economics* 120: 835–864.

Jones, B.F., and B.A. Olken. 2008. The anatomy of start-stop growth. *Review of Economics and Statistics* 90: 582–587.

Jones, B.F., and B.A. Olken. 2009. Hit or miss? The effect of assassinations on institutions and war. *American Economic Journal: Macroeconomics*, forthcoming.

Kalt, J.P., and M.A. Zupan. 1984. Capture and ideology in the economic theory of politics. *American Economic Review* 74: 279–300.

Levitt, S.D. 1996. How do senators vote? Disentangling the role of voter preferences, party affiliation, and senator ideology. *American Economic Review* 86: 425–441.

Londregan, J., and K. Poole. 1990. Poverty, the coup trap, and the seizure of executive power. *World Politics* 42: 151–183.

Marx, K. 1852. *The eighteenth Brumaire of Louis Napoleon*. New York: Die Revolution.

Tolstoy, L. 1869. *War and peace*. Moscow: Russkii Vestnik (first English edition: 1886, New York: William S. Gottsberger).

Tsebelis, G. 2002. *Veto players: How political institutions work*. New York: Russell Sage Foundation.

Weber, M. 1947. *The theory of social and economic organization*. New York: Free Press.

National System

Henry W. Spiegel

Keywords

Balanced growth; Bowen, F.; Carey, H. C.; Eucken, W.; Full employment; Hamilton, A.; Infant-industry protection; List, F.; National system; Protection; Social overhead; Sombart, W.; Tariffs

JEL Classifications

B1

The term ‘national system of political economy’ stems from a filiation of American and German ideas that arose in opposition to the universalist character of classical economics and were designed to promote public policies serving the economic development of the nation. The development was visualized as one that would yield a balance of agriculture and industry and make the most of a country’s potential economic strength. The term ‘American system’ occurs as early as 1787 in No. 11 of *The Federalist*, where Alexander Hamilton launches this appeal to his readers: ‘Let the thirteen states, bound together in a strict and indissoluble Union, concur in erecting one great American system, superior to the control of all transatlantic force or influence and able to dictate the terms of the connection between the old and the new world.’

Hamilton’s more detailed proposals regarding the ways and means to construct the American system can be found in his great state papers, written when he served as Secretary of the Treasury in President Washington’s cabinet, and dealing with manufactures, a national bank, and the public debt. With the help of these three instruments he wished to emancipate the new nation from the rural economy of its forefathers, one that Thomas Jefferson, Hamilton’s great antagonist, attempted to preserve. Among Hamilton’s specific devices to promote industrial development, bounties, or subsidies, stood out. Later writers emphasized protective tariffs rather than bounties.

These writers included Daniel Raymond, a Baltimore attorney, whose *Thoughts on Political Economy* of 1820, while not elaborating the notion of a national system in so many words, made a substantial contribution to the later interpretation of the term by introducing the concept of ‘capacity’ to produce goods, identified by him with national wealth. Raymond placed on government the duty of utilizing and enlarging this capacity by a policy of protection. His plea for protective tariffs was supported both by the infant-industry argument and the employment argument, in conjunction with which Raymond wrote explicitly of ‘full employment.’

The next step in elaborating the concept of a national system was taken by Frederick List, the German writer and promoter, who in 1827 during his residence in the United States published *Outlines of American Political Economy*. Like Hamilton, List writes of the ‘American system’, which was to realize its potential with the help of tariff protection. This work was written and distributed at the behest of a Pennsylvania manufacturers association whose members clamoured for tariff protection. Composed ostensibly in the form of letters addressed to a leading protectionist, the work appeared serially in the *National Gazette* of Philadelphia and was reprinted by more than 50 other newspapers. When published in pamphlet form, it was distributed in ‘many thousand’ copies, as List later reported. It was sent to the members of Congress and was apparently helpful in securing the adoption of the Tariff Act of 1828.

In an abortive attempt to win a prize, List wrote in French in 1837 an essay on *The Natural System of Political Economy*, which remained, however, unpublished until 1927, when it was printed in French and German. An English translation appeared only in 1983. This work anticipates in a number of respects List's principal work, *National System of Political Economy*, in which the national-system doctrine reached its full flowering. This work was published in German in 1841; an English translation, sponsored by protectionist interests in the United States, appeared in 1856, and another one, published in England, in 1885. The work, while substantial enough in itself, was intended to be the first part of a larger project, which, however, was never completed. Of the English translations, the earlier one omits the preface, while the later one contains extracts from the preface but omits the introductory chapter that provides a summary of the work.

In the *National System*, List finds fault with the classics for a variety of reasons. He takes them to task for having constructed a system of thought that is permeated by individualism and cosmopolitanism but neglects the nation. According to List, the community of nations is not a homogenous group but made up of members that find themselves at different stages of their development. List then goes on to construct a stage theory which visualizes progress from the agricultural stage to one in which agriculture is combined with industry, and to still another one in which agriculture, industry, and trade are joined together. List tends to equate agriculture with poverty and low level of culture, whereas industry and urbanization bring wealth and cultural achievement. The classics, with their homogenized picture of the world which neglected national differences, would tend to perpetuate the underdeveloped status of the United States and continental Europe vis-à-vis the highly developed Britain. According to List, each stage, or each nation at its respective stage, requires a different set of economic doctrines, whereas the classics claimed universal validity for their doctrines.

At heart, List wanted to improve on Providence by turning all people into Englishmen. To allow the underdeveloped countries of his time to participate in the march toward higher stages, attention would have to be paid to their productive capacities. The development and utilization of these was a task that List placed squarely on the national governments. In this connection List called for liberal political institutions, for the construction of what is now known as social overhead, especially in the form of transportation facilities, for balanced growth and for tariff protection for infant industries (not for agricultural products). The free-trade orientation of the classics List was willing to endorse as valid for the future, when all nations had utilized their potential and attained the most progressive stage. Then free trade would be combined with universal peace and a world federation.

There are a number of questions that List left unanswered. To begin with the most often heard objection to the infant-industry argument for protection, what tests are there to identify infant industries and to mark their eventual attainment of maturity, when protection presumably is to terminate? Moreover, List did not explain how the type of economic warfare that he envisaged would prepare the ground for universal peace. Nor did he show awareness of the likelihood that, once all nations had progressed to what he called the normal state one nation would again get ahead of the others, perhaps for reasons of technological advances, a matter treated with so much insight by Hume in his analysis of the migration of economic opportunities.

List had been a protectionist of sorts already in his young years in his native Germany. His protectionist leanings came to the fore in the United States, where he encountered an even richer potential for economic development and where changing economic conditions were more rapid and conspicuous. Here List's strictures on the classics fell on fertile ground because so many features of their dismal science did not seem to fit into the American environment, especially Malthus's population doctrine and Ricardo's theories of subsistence wages, diminishing returns, and free trade. Thus List's work coalesced with the works of native American critics of the classics, especially of Henry Carey, who developed theories of increasing rather than diminishing returns and of rising wages and profits and declared that each successive addition to the population brings a consumer and a producer. According to Samuelson, Carey's 'logic was often bad and his prolix style atrocious. But his fundamental empirical inferences seem correct for his time and place' (p. 1732). Beginning in 1848, Carey became an ardent exponent of protectionism. By this time List was dead and it is uncertain to what extent, if any, Carey was indebted to List's thought. Neither of the two developed his proposal for tariff protection in isolation but as parts of a wider system of thought, of a theory of economic development in the case of List and of a theory of a harmoniously ordered society in the case of Carey.

Among political leaders in the United States Henry Clay is often mentioned as an architect of the American system, in which the industrial east and the agrarian west were allied in a powerful union. He pleaded for such a system in a famous speech in 1824, in which he supported protective tariffs as instruments of industrial development. Later still, in 1870, Francis Bowen, an early teacher of economics at Harvard, would publish *American Political Economy*, in which he supported tariff protection and which caused him to lose his teaching job in economics, the president easing him into the presumably less controversial field of history.

In Germany, List's ideas had a profound and lasting influence. He promoted the customs union, which by 1844 covered almost all of Germany, and agitated for railroad construction and tariff protection. The very name of economics in Germany, Nationalökonomie, conveys associations with List. Some German interpreters of the history of economics have compared List with Marx. Both had utopian visions of a society to come in the fullness of time. Both made much of a fusion of theory and practice and of economics and politics. Both are linked by their reputation as rebels who opposed the established order. It is an interesting trivium that in 1841 List turned down an offer to serve as the editor of a newspaper that was to be published under the name of *Rheinische Zeitung*, a post that Marx filled the following year.

List's thought has an affinity with the historical schools and institutional economists, who had ideas of their own about the possibility of universally valid economic doctrines. The word 'system', cleansed of its protectionist implications, continued to play a key role in the writings of such twentieth-century German economists as Walter Eucken and Werner Sombart. An equally faint echo of the Hamiltonian idea can be discerned in the current usage of the word in conjunction with the study of comparative economic *systems*.

See Also

- ▶ Comparative Advantage
- ▶ Corn Laws, Free Trade and Protectionism
- ▶ Growth and International Trade
- ▶ Infant-Industry Protection

Bibliography

Bowen, F. 1870. *American political economy*. New York: Scribner.

Carey, H. 1858–59. *Principles of social science,* 3 vols. Philadelphia: Lipincott.

Conkin, P.K. 1980. *Prophets of prosperity: America's first political economists*. Bloomington: Indiana University Press.

Dorfman, J. 1946. *The economic mind in American civilization*. Vol. 1–2. New York: Viking Press.

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

Henderson, W.O. 1983. *Friedrich List: Economist and visionary 1789–1846*. London: Cass.

Hirst, M.E. 1909. *Life of Friedrich List and selections from his writings*. London: Smith, Elder.

List, F. 1827. *Outlines of American political economy*. Reprinted in *The life of Friedrich List and selection from his writings*, ed. M.E. Hirst. London: Smith, Elder, 1909.

List, F. 1837. *The Natural system of political economy*. Trans. and ed. W.O. Henderson. London: Cass, 1983.

List, F. 1956. *National system of political economy*. Philadelphia: Lippincott.

Samuelson, P.A. 1960. American economics. In *Postwar economic trends in the United States*, ed. R.E. Freeman. New York: Harper. Reprinted in *P.A. Samuelson, collected scientific papers*, ed. J.E. Stiglitz, vol. 2, 1732–1747. Cambridge, MA: MIT Press, 1966.

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

Spiegel, H.W. 1983. *The growth of economic thought*. Revised and expanded ed. Durham: Duke University Press.

Nationalism

Samir Amin

There is a certain ambiguity in words such as nationalism, let alone economic nationalism. Indeed, a distinction must be made between the social reality which determines a nation and the degree of autonomy of States in the world system. A distinction must equally be made between theories concerning the analysis of the world economic system and normative propositions that define strategies of insertion into or confrontation with this system.

The term 'nation' presupposes certain articulations between this reality, real or alleged, and other realities such as the State, the world system of States, the economy and social classes. We currently owe these concepts and their articulation into a system to the different social theories developed in the light of the 19th-century European historical experience. Within this framework the elaboration of two sets of theories took place – as it turned out, in counterpoint to one another: on the one hand, marxism and the theory of the class struggle; on the other, nationalism and the theory of class integration into the democratic bourgeois nation-state. Both theories take account of many aspects of the immediate reality which is marked both by social struggles ending in revolutions, and by struggles between nation-states ending in war. For protagonists of these theories, they have proven to be potent guides to action.

The efficacy of political strategies was, however, dependent upon specific circumstances defined by a coincidence – apparently limited in time and space – between elements: (i) coincidence between the State and another social reality i.e. the nation; (ii) the dominant position of bourgeois nation-states in the world capitalist system and their 'central' (as opposed to marginal) character in our conceptual system; (iii) a degree of worldwide application of the capitalist system which led central partners to form 'autocentred' interdependent economic units enjoying a high degree of autonomy vis-à-vis each other.

These circumstances define a possible field for 'national' economic policy. The instruments of this policy – the national centralized monetary system, customs laws, the network of material infrastructures in transport and communications, the unifying effect of a 'national' language, the unified administrative system and so on – enjoy a definite autonomy in relation to the 'constraints' imposed by an economy applied world-wide. Relations between classes, however wrought with conflict, are relegated to and by the national State. In this sense, there exists an average price for the national labour force which is determined by history and by internal social relationships i.e. a national price system that reflects decisive social relationships. In this sense, the 'law of value' assumes a national dimension. True, there is no Great Wall of China to separate these national systems from the world system that they constitute. Internal social relationships are partly dependant upon positions occupied by the national States in question in the world hierarchy. All these are 'central' capitalist economies but are not equally competitive. If social relations permit, these States can improve their position by pursuing coherent national policies. This effectiveness in turn facilitates social compromise and, without 'abolishing the class struggle', puts definite limits to conflicts.

In these circumstances, what is the role of the so-called 'national' reality? *A posteriori* ideology lends an autonomous dimension to the national reality by granting it pre-existence to the State. This in fact seems questionable. For the European bourgeoisie – from the Renaissance to the Enlightenment – appears cosmopolitan rather than narrowly national. This bourgeoisie shares its loyalty between several legitimacies, religious or philosophical convictions, feudal type friendships, but also in service to the State as absolutist monarchy when it appears reasonable to do so. It still remains generally mobile, at ease in the whole of Christendom. As to the peasantry, its loyalty focuses more on the soil and the locality than on the future nation in which it does not yet share culturally nor sometimes even linguistically. But the Nation is progressively created by the absolutist monarchical State, a task which is completed by bourgeois democracy. The regional ethnolinguistic conglomerates under the same King are not 'by nature' destined to become modern European nations: it is only a potentiality.

However, at closer inspection it appears that these circumstances, pervasive but limited in time to the 19th century, are even more limited in space. Around a few 'model' nation-States, the world of the capitalist system – structured by different pasts which in turn lose their legitimacy and efficacy – remains undefined in the light of an uncertain and obscure future.

The problem changes when we quit the limited framework imposed by the central bourgeois nation-states. For this forces us to examine 'regions' more closely whether they are organized into States or not. Regions are peripheral in relation to continuously expanding capitalist reproduction. On this level there is only a central State, i.e. a State which masters external relations and submits them to the logic of autocentred accumulation. On different levels, there are only 'countries', which are administered from outside as colonies or semi-colonies; these appear to be independent but incapable not only of moulding the outside according to their needs but also of avoiding their drift and shaping from outside.

So we are confronted with the problems relating to the specific future of these regions and peripheral States. This future is implied by the worldwide application of capitalism and is based on the thesis of worldwide application of the law of value as an expression of value in the productive system. This thesis implies that the labour force has only one value for the whole world system. If this value has to be related to the level of development of productive forces, it follows that this level will be characteristic of the whole world productive system and not of the different national productive systems which progressively lose their reality due to the worldwide application of this system. But the price of the labour force differs from country to country. This price depends on political and social conditions which characterize each national social formation. The more the reproduction of the labour force is partially ensured by a value transfer of non-capitalist market production and non-market production, the less is the price. The formal submission of peripheral non-capitalist modes of production to a global exploitation of capital allows for a higher rate of surplus-value in real capitalist production; this contributes to the heightening of the average level of the rate of surplus-value on a world scale.

Until the end of the 19th century, this worldwide application had led to the integration of only a certain number of basic products in an international rather than worldwide market. This first stage allowed for laws of value with a national content in the framework of constraints imposed by international competition by the embryonic world capitalist law of value. At this stage, social classes were still essentially national classes, defined by social relations formed within the limits of the State. There is thus a conjunction between class struggles and the play of politics which precisely takes place within the framework of the State. From the end of the 19th century until World War II, the internationalization of monopolistic capital went parallel to the international market in basic products. But this stage was characterized by the absence of world hegemony, and monopolies which were constituted on the basis of competing central States operated preferentially in peripheral regions cut out between colonial empires and zones of influence. Due to the absence of the State or its weakness in these peripheral regions, social relations contracted within central national States continued to define the dynamics of capitalist expansion. After World War II, the stage for the worldwide application of the productive processes was elaborated by an explosion of productive systems into segments which the so-called 'transnational' form of enterprise controlled and distributed all over the planet. The hegemony of the United States constituted an adequate framework for this transnationalization.

Henceforth the world dimension of the law of value dominates over its local dimensions. This reality is clearly reflected in economic discourse; the constraint imposed by competitiveness on a world scale is hauntingly evident in speeches by those in power; it is presented as unavoidable; to ignore it is synonymous with a denial of 'progress' and so on …. But by this very fact the State – whether national or not – also loses its efficacy as a place for elaborating strategies that command or modulate capitalist expansion. Since there is no planetary State, the coincidence between conflicts and class compromise on the one hand, and politics on the other hand has disappeared.

However, in general this crises does not affect the different components of the world system to the same extent. Developed capitalist centres such as the United States, Europe and Japan are in the main not threatened by this evolution. Here we must allow for certain differences, since the historical heritage in Europe – which is still divided into separate political States despite the unfinished construction of an economic community – places Europe in a more difficult position than the United States or Japan. This leads to the questioning of American hegemony and of its eventual end, but it does not question the very existence of the Nation-states considered.

The situation is very different at the periphery of the system. Here, at the end of World War II, once political independence had been regained, the bourgeoisie of the Third World nurtured a project for 'national construction in the cadre of global interdependence' which we will characterize here as the 'Bandung Project'. This project can be defined by the following elements: (i) the will to develop productive forces, to diversify production (i.e. to industrialize); (ii) the will to ensure that it is the national State that assumes direction and control of the process; (iii) the belief that 'technical' models constitute given 'neutrals' which can only be reproduced even if they have to be mastered first; (iv) the belief that the process involves no initial popular initiative, but only popular support for State action; (v) the belief that the process is not fundamentally contradictory to participation in trade within the world capitalist system, even if this leads to short-lived conflicts.

The realization of this national bourgeois project by implication meant bringing under control through the State and by the hegemonic national bourgeois class, at least the following processes: (i) control of the reproduction of the labour force; this implies a relatively complete and balanced development so that, for example, local agriculture is capable of delivering products in reasonable quantity and at prices that ensure the valorization of capital essential to this reproduction; (ii) control over national resources; (iii) control over local markets and the capacity to penetrate the world market under competitive conditions; (iv) control over financial circuits thus enabling the centralization of surplus and the orientation of its productive use; (v) control over current technologies at a level of development reached by productive forces. The circumstances surrounding capitalist expansion in the years 1955–1970 have to a certain point favoured the crystallization of this project.

Today it is no longer possible to ignore the shortcomings of such attempts, which have not been able to resist a reversal in favourable circumstances. Agricultural and food crises, external financial debt, mounting technological dependency, fragility in the capacity to resist any future military aggressions, creeping waste in the manner of consumer capitalist models, and their influence in the areas of ideology and culture, are signs of historical limitations to these attempts. Even before the present crises opened the occasion for a 'Western offensive' which could reverse these developments, these shortcomings had already reached an impasse.

This period is now over and the focus in the new world circumstances is centered around the offensive by the capitalist West against the people and nations of the Third World. Here the objective is to subordinate their future evolution to the particularities of a redeployment of transnational capital.

Are these only temporary circumstances which will necessarily be followed by a new dawn of 'national bourgeois' advances? Or are we seeing an historical turning point which will exclude the pursuit of successive national bourgeois attempts such as those that characterized at least a century of our past history?

Our hypothesis is that the contemporary crises marks the end of an epoch; an epoch which in the case of Asia, Africa and Latin America can be called the century of the National Bourgeoisie, in the sense that it has precisely been marked by successive attempts at national bourgeois edification. Our hypothesis is that the Third World bourgeoisie now finally sees its own development in terms of the Comprador subordination imposed upon it by the expansion of transnational capitalism.

The nationalist populist political strategy known as deconnection appears at this junction as a credible future alternative. For the restoration of the Comprador system on a Third World scale is bound to be hampered by the rise of populist movements. In the initial stage, the populist form is not a surprising development since it is undefined and characterized by ambiguous ideologies. It reflects the broad character of a class alliance, in which classes are in turn uncertain of their determination and deprived of autonomy and class consciousness. But this does not exclude it as a potent world disintegrating force which under certain conditions can evolve towards positive crystallizations.

We suggest that these positive crystallizations involve a merging of three conditions. These are, first, a deconnection in the sense of a strict submission of external relations in all areas to the logic of internal choices taken without consideration of criteria relating to world capitalist rationality; second, a political capacity to operate social reforms in an egalitarian sense. This political capacity is both a condition of deconnection – since existing hegemonic classes have no interest in it – and as possible consequence of deconnection, since this obviously implies a transfer of political hegemony. A deconnection without reform has little chance of emerging. If it did emerge under certain economic conditions, it would lead to an impasse; third, a capacity for absorption and technological invention, without which the autonomy of decision making acquired could not be realized.

Thus defined, the conditions for a positive response to the challenge of history appear severe, and any merging of such conditions seems improbable. In the immediate future, such a possibility seems remote; it may nevertheless appear to be the only reasonable solution.

See Also

- ▶ Autarky
- ▶ Colonialism
- ▶ Dependency
- ▶ Imperialism
- ▶ Inequality Between Nations
- ▶ Periphery
- ▶ Structuralism
- ▶ Terms of Trade and Economic Development
- ▶ Unequal Exchange
- ▶ Uneven Development
- ▶ Vent for Surplus

Nationalization

M. V. Posner

In socialist economies most enterprises, and all large enterprises, are publicly owned and controlled. In most capitalist economies, at most times, there have been some examples of enterprises owned, controlled, or managed by agents of the government. Naval shipyards, munitions factories, post offices, early telecommunication systems, and the water cycle – these are some of the earliest and continuing examples of public enterprise in an essentially private, capitalist setting.

In the first half of the 20th century there were two large and powerful bursts of expansion from this small base. First, in the wake of inflation and slump, between the end of World War I and the opening of the World War II major takeovers of commercial concerns in the private sector occurred in several countries, particularly Italy; but it is important to note that many of these concerns were facing insolvency, or feared insolvency. The second wave of nationalization came immediately after World War II in the whole of Western Europe, usually based on some form of political doctrine, sometimes reinforced by political anger against, for instance, individual capitalists who had ‘collaborated’ with the enemy during the war, but also often as the culmination of a long period of state involvement in the rationalization, amalgamation, or regulation of natural monopolies.

Thus, in Italy, in the fascist period, several major investment banks were taken over to avoid failure, and with them were taken their industrial affiliates and customers. In France, after World War II, automobile firms and banks joined electricity, gas and the railways. In Britain, the postwar Labour government enforced nationalization of the railways, after thirty years of publicly influenced amalgamations and mergers; of the coal industry, much in the spirit of committees of inquiry that had reported before the War; and of electricity and gas, in a way which effectively turned a number of small regional public utilities, already largely under public ownership, into large national monolithic monopolies.

The study of nationalized industries in Western Europe then became an amalgam of what was known in North America as ‘public utility regulation’, and of the more complex task of managing enterprises whose activities were directly competitive with private enterprise. Both in Western Europe (particularly in Norway and Italy) and even more so in the developing world, the nationalization of depletable resources was widely practised from the 1950s onwards. Reserves of oil and natural gas in Western Europe, and of other minerals elsewhere, were used in part to control rates of depletion, in part to limit the power of expatriate companies who were otherwise likely to exploit the natural resource, but in large part to appropriate the ‘rent’ to the host government. Whether the alternative method of extracting rent – a suitable tax system, or a system of auctioning production licences – works better or worse than national ownership is still a matter for controversy. In the UK, it is widely believed that the petroleum revenue tax, as introduced by a Labour government and reformed by a Conservative government, has worked very well; but in Norway a state oil company played a major part in appropriating the rent from the Norwegian part of the North Sea.

In many ways, the distinction between the publicly owned ‘natural monopolies’ and the publicly owned ‘competitive enterprises’ is a false one. Already in the 1950s, and certainly in the 1980s, most of the important business of the railways in most West European countries was directly competitive with road or air transport, even though in some countries the state-owned railways were cushioned against the rigours of competition from the roads for a mixture of political and environmental reasons, or more often through historical inertia. In many countries, space-heating by the electricity and gas public utilities were highly competitive with each other, and the private sector oil companies took a vigorous share of both domestic and industrial heating markets in most countries for most of the postwar period. The monopoly position of the post, telephone and telegram companies did persist until the development of electronics in the 1960s and 1970s made the old system impossible to maintain, and the postal monopolies were broken more and more by private express delivery services.

Nevertheless, most observers do see a necessary distinction between a ‘natural monopoly’ and other potential candidates for public ownership. Even those who see the clearest and strongest arguments for unfettered free enterprise in a capitalist economy are apt to accept the case for the regulation of natural monopolies; and if an enterprise is to be regulated by a government agency, why should it not be owned by a government agency? Until the process of technical change began to erode the monopoly position of the energy and transport industries as traditionally organized, they tended to fall within the ambit of public ownership in most countries outside North America.

Therefore there was a continuum of public enterprise, from the highly monopolized and regulated water services at one extreme, to the highly competitive engineering companies in the Italian public sector or the French banks. Nevertheless, the public image of nationalization in most countries was the image of the large, monolithic, public utility, protected in most of its markets from the rigours of competition, with its customers at its mercy, and fairly evident collusion between management and workforce to maintain an easy and undisturbed life. Nobody relied much on customers to influence these giants, ‘consumerism’ (the various bodies that banded individual customers together, sometimes with the support of government funds, sometimes without that support) was not very successful either. The responsibility for control therefore was necessarily seen to rest in the hands of the government, who usually by constitutional provision and always in practice, were the ‘owners’ of the enterprises, with powers to appoint and dismiss managers.

In most countries, and at most times, governments that owned nationalized industries also acted as their bankers. At some times in some countries individual enterprises have been allowed to ‘go to the market’ to borrow, nominally on their own account; but there has most often been an explicit or implicit state guarantee of that borrowing, and it has rarely been possible for a publicly owned enterprise to borrow substantially without permission of the government.

Most governments in Western Europe, and many throughout the world (for instance, Japan) have therefore had ministries, or parts of ministries, devoted entirely to the control of nationalized industries. Cabinet meetings have frequently had on their agenda the financial or labour-relations problems of these industries; the selection of the top management or supervisory boards for the nationalized industries has preoccupied ministers, prime ministers and heads of state; trade union organization has, through the postwar decades, often been rearranged so as to parallel and match at every level the structure of management in the nationalized monoliths. In some countries, notably Italy, the affairs of the nationalized industries have been of even wider political importance: individual ministers have built or destroyed their careers in the course of battles with the independent barons of the nationalized industries; often newspapers and radio stations, and even political parties, have been influenced by the activities of, and sometimes by financial appropriations from, publicly owned industry.

If we take, as a measure of the financial involvement of government in the affairs of nationalized industries, the sum of their net current account subsidy of all nationalized enterprises taken together, plus capital account loans to those entities, then in peak years this has amounted to ten per cent of the government budget in some industrialized countries, and far larger proportions in developing countries. In some Western European countries, the public sector (excluding public administration as conventionally defined) at its peak, accounted for nearly 25 per cent of employment, output, and capital investment.

Rules for the behaviour of the managers (‘agents’) of public enterprise attracted the attention of many economic writers from the 1930s onwards. The combination of a marginal cost pricing rule, investment decisions determined by an internal rate of return requirement (‘test discount rate’), and a constraint on the overall financial deficit or surplus of an enterprise, was used or recommended in most countries.

It is simple to show that a system which has a pricing rule, an investment rule, and a constrained profit and loss account, is over-determined, in the sense that one or other of these three rules will under normal circumstances either be redundant or have to be overridden. But if the scale of output, or the size of the enterprise, can itself be varied, then the number of targets can be adjusted precisely to the number of variables, and the system is determinant. In simple language, an enterprise with marginal costs that are low compared with total costs, but which is not permitted by its parent ministry to exceed a certain fixed annual maximum ‘loss’ on revenue account, will contract its scale of operations so as to fit in with its financial constraint. For instance, throughout the world, railway systems have, with greater or lesser speed and smoothness, adjusted themselves to this combination of principles by reducing the number of trains. Equally, those enterprises which have found surpluses easy to earn, have been tempted to expand the scale of their operations.

The application of the investment rule in a loss-making enterprise is a vexed and, in practice, still unsolved problem, although formal theoretical solutions are easy to define: essentially, the test is whether the investment will increase or decrease the expected deficit. The practical difficulty with those solutions is that neither governments nor their citizens have proved willing to 'pour good money after bad' in the way theoretical prescription would suggest.

Working out of the details of marginal cost pricing, and the avoidance of some simple absurdities (the marginal cost of an additional passenger on most trains is zero – should therefore the price of *all* rail tickets be zero?), has stimulated quite a lot of good theoretical economics. The distinction between long and short-run marginal costs, coping with technological progress, and other complexities has been taken furthest in the theory of electricity pricing. Price discrimination between different customers – the extent to which it should be permitted, the rules which should control it, and the political acceptability of the theoretical solutions – have all caused considerable difficulties.

The net result of the refinement of these financial rules, and the drive towards greater efficiency and managerial independence of the large nationalized industries, has made them indistinguishable, for many purposes, from the large corporations of the private sector. The public complaint about nationalized monopolies – their size, insensitivity to customers, excessive political power, their tendency to act as 'states within a state' – have been not dissimilar to the complaints levied against private sector giants. Indeed, disaffection, for instance, in the United States with 'the telephone company' or their electricity utility company is not at all unlike disaffection in Western Europe with their nationalized counterparts. And despite the political attractions to socialistically minded labour unions of nationalization in the 1940s and 1950s, organized labour has had as many disputes with employers in the public sector as in the private sector; and, indeed, these disputes have become more politicized, and therefore harder to settle, when public sector employers were involved. From the point of view of the unions, nationalized industries have become part of 'state capitalism'; from the point of view of laissez-faire economists or politicians, large public corporations have strengthened large labour unions and these bilateral monopolies have been tempted to collude together against the interests of consumers and the interests of the public generally.

So, when in the 1970s and 1980s the fashion began to grow for splitting up, 'hiving off', or loosening centralized control of huge *private* sector enterprises, this spirit was very readily transferred by right-wing political parties into 'denationalization' or 'privatization' campaigns for the *public* sector. Ironically, although these campaigns started with the notion of turning big public enterprises into small private enterprises, quite rapidly the impetus was redirected, and what was changed became more and more the form of ownership rather than the form of organization. What nationalization had put together, denationalization has not always torn asunder.

One hypothesis that would explain the swing in fashion from favouring giant public corporations in the 1940s to the other extreme of favouring small-scale enterprise in the 1980s is that the underlying technological facts have themselves changed over the 40-year period. For instance, in the UK, the electricity supergrid transmission system required a single nationwide switching and control room, under centralized control; the development of local load-balancing devices, and of smaller top-up gas turbine plant could, in the 1980s, perhaps enable the marked diminution of the importance of the transmission interconnection, and allow greater autonomy for a number of regional generating utility companies. But this hypothesis is not very plausible, either in the particular case cited or more generally. At the same time as the wave of denationalization, privatization, and 'hiving off' in the 1980s, amalgamations and mergers were proceeding apace in many other parts of the economy, in sectors as far apart as financial services and food processing. The twilight of nationalization in the OECD world cannot be substantially attributed to technological change.

It seems unlikely that the present swing of the pendulum against public ownership will return the situation in Western Europe to where it was before

World War II, but doubtless the swing still has further to go. Some economists are quite clear that in the 30 years during which the pro-nationalization fashion lasted, technical process was delayed, resources were misallocated, market opportunities were missed. My own judgement would be that in the first decade – broadly to about 1960 – public ownership did quite well; in a second phase – through the decade of the 1960s – it did become stuck, labour union strength did prevent change far more than in the private sector, managerial competence was not high, political interference was sometimes cripplingly great. By the early 1970s these difficulties had been taken in hand, and many of the giant corporations were doing really rather well for their owners and their customers, but by that time the long-term shift of opinion against these dinosaurs had become unchallengeable – the public had come to think of them as slow-moving relics of the past.

As the English poet Pope should have written: 'Of forms of ownership let fools contest, what e'er is best managéd is best.'

See Also

- ▶ Marginal and Average Cost Pricing
- ▶ Privatization
- ▶ Project Evaluation
- ▶ Public Utility Pricing
- ▶ Socialism

Bibliography

Chenot, B. 1956. *Les entreprises nationalisées*. Paris: Presses Universitaires de France.

Einaudi, M., M. Byé, and E. Rossi. 1955. *Nationalization in France and Italy*. Ithaca: Cornell University Press.

Holland, S. (ed.). 1972. *The state as entrepreneur*. London: Weidenfeld & Nicolson.

Posner, M.V. 1973. *Fuel policy: A study in applied economics*. London: Macmillan.

Posner, M.V., and J.S. Woolf. 1967. *Italian public enterprise*. London: Duckworth.

Pryke, R. 1971. *Public enterprise in practice*. London: MacGibbon & Kee.

Robson, W.A. 1960. *Nationalized industry and public ownership*. London: George Allen & Unwin.

Shepherd, W.G. 1965. *Economic performance under public ownership*. New Haven/London: Yale University Press.

Votaw, D. 1964. *The six-legged dog: Mattei and ENI*. Los Angeles: University of California Press.

Natural and Normal Conditions

John Eatwell

For economic science to begin, the object which that science is to investigate must be defined unambiguously. This definition was first provided by Adam Smith in Chapter 7 of Book 1 of the *Wealth of Nations*. There Smith defined the object of the analysis of value and distribution to be the mode of determination of 'natural' prices. Marshall replaced the evocative label 'natural' with the more prosaic 'normal'.

In both cases 'natural' and 'normal' must refer not only to prices, but also to the outputs and means of production, and the levels of overall activity, associated with those prices, since the object of investigation, the market economy, must be expressed in coherent form.

A primary issue in the development of theoretical knowledge in the social sciences (or, indeed, in any science) is the problem of abstraction and the definition of abstract categories. This problem has two dimensions: first, the *object* on which the enquiry is to be focused must be defined in terms that will permit statements of general validity; secondly, the *theory* which is to explain the magnitude or state of the object must itself be constructed at a particular level of abstraction. Although these two dimensions are not unrelated they are essentially sequential. If they were to be simultaneous (as they are in 'intertemporal equilibrium', e.g. Debreu 1959) the object might be defined to fit the theory, and the theory would in consequence reveal little other than its own structure.

In defining the object of the analysis and identifying the forces which determine it, the assumption is made, implicitly, that the forces of which the theory is constituted are the more dominant, systematic and persistent. Transitory and arbitrary phenomena are abstracted from intentionally; as are those forces which are related to specific circumstances as opposed to the general case. In quantitative analyses, the dominant forces are expressed in algebraic form, as functions and constants, and constitute the *data* of the theory. The model may then (if it has been specified correctly) be solved to determine the magnitude of the object. It is known that, except by a fluke, the magnitude determined as a solution will not be exactly that observed in reality. It cannot be, since a variety of transitory forces, known and unknown, have been excluded. Nonetheless, since the theory is constructed on the basis of dominant and persistent forces, the magnitude determined by the analysis is the *centre of gravity* of the actual magnitude of the object. Whether this centre of gravity is a temporal constant, or takes different values through time, does not affect the essence of the method.

The development of abstract categories, in particular the sequential formulation of object and theory, may be traced in the evolution of economic thought.

The 17th and 18th centuries saw the progressive development of the social division of labour and the emergence of wage labour as the idea emerged that prices – the parameters of markets – and hence the entire economic system, might be subject to the influence of systematic 'laws'. On the basis of this insight Adam Smith constructed the abstraction of an economy organized entirely through competitive markets, and isolated the problem of price formation as a necessary element in the search for an understanding of the laws determining the operations of the economy. To distinguish the dominant from the transitory, Smith characterized the competitive market as establishing 'natural or average' rates of wages, profits and rents. When the price of a commodity is just that which provides for the payment of the land, labour and 'stock' used in its production at their natural rates, then the commodity sells at its *natural price* (Smith 1961, Book 1, ch. 7).

While natural prices were held to be the outcome of the persistent forces in the economy, *market prices*, the prices which actually rule at any one time, are influenced by a variety of transitory or specific phenomena, elements which may be excluded from the analysis of the more permanent forces in the economy.

The natural price is characterized not only as a single price for each commodity, but also by a uniform rate of profit on the value of capital invested in each particular line. Indeed, as Ricardo argued, it is the active role played in the organization of production by the capitalists seeking the maximum return on the finance they have invested in means of production that is the basis of the tendency toward natural prices (Ricardo 1951a, p. 91). Marx elaborated this point by emphasizing that the tendency toward the equalization of the general rate of profit and the exchange of commodities at their prices of production (as the called natural prices) 'requires a definite level of capitalist development' (Marx 1967, p. 177). So the associated categories of natural price and of the general rate of profit were an integral part of the characterization of a capitalist economy.

The fundamental change in economic theory which occurred in the final quarter of the 19th century did not, with respect to prices, lead to any significant change in the definition of the object. The new neoclassical theory was an alternative to the classical theory. As an alternative it necessarily offered a new and different explanation of the same object. This continuity in the object which accompanied the great discontinuity in the theory is particularly evident in Marshall, who devoted considerable attention to the specification of short-period *normal* prices and long-period *normal* prices, the concepts he substituted for the market prices and natural prices of Smith and Ricardo (Marshall 1961, Book 5, chs 3, 5). But the same continuity may be found in the work of Walras (1954, pp. 224, 380), Jevons (1970, pp. 36, 135–6), Böhm-Bawerk (1959, p. 380) and Wicksell (1934, p. 97).

Two important aspects of the specification of this familiar framework for the analysis of capitalist economies should, perhaps be clarified.

First, the notion of the tendency towards a uniform general rate of profit on the supply price of capital goods derives from the two-fold character of capital in a market system: money-capital and commodity-capital. In a system in which production and distribution are organized by means of a generalized process of exchange money assumes the form of the general equivalent of value, and ownership of money or access to finance endows the ability to own and control the production and distribution processes. Hence the accumulation of monetary wealth becomes, but the nature of the competitive system, the ultimate objective of each individual capitalist, leading him to attempt to maximize the return on the value of the means of production in which he invests his money. But the production of surplus (profits) in the economy as a whole is not a financial phenomenon, it takes place in the process of production. The realization of a financial return and the organization of the process of production are two dimensions of the same phenomenon, two phases in the circuit of capital, which find their conceptual unity in the general rate of profit.

Second, the determination of natural prices and the general rate of profit is associated with the 'socially necessary' or 'dominant' technique of production. At any one time a given commodity may be produced by means of a variety of techniques: some 'fossils' embodying out-of-date methods, which are not being reproduced since at existing prices they would yield a rate of return on their supply price lower than the general rate of profit, but which nonetheless do yield positive quasi-rents; some 'superior' techniques which are used only by a limited number of producers and yield super-profits. The various theories of value and distribution are not concerned with these, but with 'the conditions of production normal for a given society' (Marx 1976, p. 129), the 'normality' being defined by dominance throughout the competitive market.

These considerations amount to the proposition that satisfactory analysis of value and distribution in a capitalist economy should endeavour to explain and determine the normal or long-period position of the system – whereby long-period is meant not that which occurs in a long period of time, but rather that which is determined by the dominant forces of the system within a period in which those forces are constant or changing but slowly. Hence if we are to present a coherent analysis of the relationship between prices, distribution and the general level of output, then the *object*, the determination of which is to be explained by the theory of output, must be the natural, or normal, level of output, itself the centre of gravity of the transitory forces which affect output at any given time. Thus a long-period *normal* analysis of the formation of natural prices must be accompanied by a long-period *normal* analysis of output.

The abandonment in modern neoclassical theory of the method of analysing natural and normal conditions as centres of gravitation and its replacement by the framework of intertemporal equilibrium – a framework within which the ideas of 'long-run' and 'short-run', or of 'gravitation toward' have no meaning – marks a major shift in the content of economic theorizing (Garegnani 1976; Milgate 1979). The content and significance of this shift has been little noticed and less discussed. Yet in the application of economic theory it is as significant as the change in the theory itself which occurred at the end of the 19th century.

See Also

- ▶ British Classical Economics
- ▶ Centre of Gravitation
- ▶ Natural Price
- ▶ Prices of Production
- ▶ Stationary State

Natural and Warranted Rates of Growth

J. A. Kregel

JEL Classifications
B1

The concepts of the natural and warranted rates of growth of national income, associated with the work of R.F. Harrod and E.D. Domar, were first developed in the 1930s and 1940s as part of the rethinking of the theory of economic fluctuations generated by Keynes's *General Theory*. Somewhat paradoxically, they formed an initial impetus for the theories for long-run steady growth elaborated in the 1950s and 1960s.

In the early 1930s Harrod criticized the static nature of economic analysis, suggesting that it be supplemented by a 'dynamic' theory: static theory determined the levels of variables, dynamic theory should explain the 'rates of change' of the variables taken at a point in time. Harrod's first attempt at dynamic theory, *The Trade Cycle* (1936), appeared almost simultaneously with Keynes's book, which Harrod considered limited to statics, even though it argued that the system could achieve equilibrium at less than full employment, because it dealt with the equilibrium *levels* of output and employment. After a lengthy correspondence with Keynes (cf. Keynes 1973, pp. 151ff), Harrod published a new version of his theory, 'An Essay on Dynamic Theory', (1939) in which he formulated a 'dynamic equilibrium' for income, Y, defined as the 'warranted rate of growth' $g_w = \mathrm{d}Y/\mathrm{d}t)/Y$, to complement Keynes's static equilibrium. Due to the outbreak of war the theory did not attract attention until he presented it in a series of popular lectures (Harrod 1948) after the war.

In Keynes's theory any level of output and employment, including full employment as a special case, was a potential equilibrium; the actual equilibrium was determined by the point of effective demand given the general state of expectations expressed in the propensity to consume, the marginal efficiency of capital and liquidity preference. Harrod was thus led to analyse a 'dynamised version of Keynes' … effective demand' (Harrod 1959), defined as the rate of growth produced by the rate of investment chosen by entrepreneurs which is warranted in the sense of maintaining a rate of expansion of effective demand which is consistent with entrepreneurial expectation and with individuals' autonomous decisions to save. The level of income, Y_0, prevailing at any point in time in the actual development of the economy will be determined by the entrepreneurs' expectations of the rate of growth of income $(\mathrm{d}Y/\mathrm{d}t)/Y_0$. On the basis of the expected $\mathrm{d}Y/\mathrm{d}t$ they will decide the investment necessary to satisfy this expected expansion in demand. This decision is made on the basis of the 'capital coefficient' (which Harrod called C, but is now generally written as v), $(I_0 = v(\mathrm{d}Y/\mathrm{d}t)$, defined as the total money expenditure that must be made on new investment projects to create an additional £ of output. The public's decisions to spend and save expressed as $\mathrm{S} = sY_0$ will then determine the actual increase in income via the multiplier $(\mathrm{d}I/\mathrm{d}t)/s = \mathrm{d}Y/\mathrm{d}t$. Entrepreneurs' expectations will only be confirmed if $v(\mathrm{d}Y/\mathrm{d}t) = sY_0$ which when rearranged produces Harrod's famous growth equation $g_w = (\mathrm{d}Y/\mathrm{d}t)/Y_0 = s/v$, with $S/Y_0 = I/Y_0$ which is Keynes's equilibrium. The rate of expansion of income is thus warranted and since entrepreneurs' expectations have been confirmed they are presumed to expect income to continue to expand at that rate. Thus, given Y_0 and s there is a set of expectations which produces a dynamic equilibrium rate which will describe an expansion of income through time of $Y = Y_t\exp(g_w t)$.

For Harrod, the analytical importance of his dynamics was to be found in the proposition that while in static analysis any departure from equilibrium produced centripetal forces driving the variable back to its equilibrium value, in dynamic analysis any movement away from equilibrium (in this case the warranted rate of growth of income) would set up centrifugal forces which would move the system further away from its equilibrium position. For example, if income were growing at the warranted rate and investment rose above the warranted rate, $I_{\mathrm{t}} > I_0\exp(g_w t)$, income would expand at a higher rate, inventories would be drawn down and additional investment would be required to restore them to normal; the expectations which produced the warranted rate would be revised upwards as investment would appear insufficient relative to the expansion in sales, leading to further increases which would eventually surpass available labour and resources. Thus, instead of returning to the equilibrium rate, g_w, an inflationary boom in which expectations would eventually be disappointed by shortages of supply, leads to a collapse of investment and expectations. Since the dynamic equilibrium is unstable, Harrod thus concludes that the warranted rate of growth is inherently unstable.

Just as in Keynes's theory, there is no reason for the warranted rate to be associated with full employment, nor is there any reason for a disturbance of the system from a dynamic equilibrium to lead to a full employment rate. Disturbances will in general lead to a series of erratic booms and slumps of variable duration with respect to the warranted rate. The full employment rate of growth does however play a role in this cyclical process by setting a limit beyond which it is impossible for the economy permanently to grow, either in equilibrium or disequilibrium. If the rate of growth of potentially employable labour, given by the rate of population growth, is $n = (\mathrm{d}N/\mathrm{d}t)/N$, the full employment rate of growth representing the maximum sustainable growth rate would be $g = n = s/v$ unless technical progress expanded output per man employed. When available technical progress is used to increase labour productivity by $\tau = (\mathrm{d}(Y/N)\mathrm{d}t)/(Y/N)$ the maximum sustainable rate, which Harrod called the 'natural' rate, would be $g = n + \tau$ The natural rate will only be an equilibrium position, i.e. a warranted rate, if households save the required proportion of income s_r which given the optimal introduction of new production techniques producing v_r, is required to produce $g_n = s_r/v_r = \mathrm{n} + \tau$. Since there is no economic mechanism that links s and v to n and$T$the natural rate is unstable, but for different reasons than if it happened by chance to be a warranted rate.

Thus, for any actual state of the economy there will be a value for $g_w \leq g_n$ which is given by the values of Y_0, v and s determined by the past history of the economy. There can, of course, be only one value for g_w since there cannot be more than one value of Y_0, s or v for any given point in time. If the economy grows at some other rate, say g_a, then Y will not expand along the warranted path $Y_t = Y_0\exp(g_w t)$, so that the rate which would be required to produce warranted growth from any subsequent point in time*,* t, would depend on the actual values of Y_t, *s and v*.

For example, if $g_a = s_a/v > g_w/s = v$, then $s_a > s$ and investment will continue to increase $g_a$until the upper limit of g_n is surpassed. This may be conceivable, for example in the period after a deep slump, but physical bottlenecks and increases in money wages due to labour market shortages will eventually lead to inflationary boom and a subsequent collapse back into a slump which will cause incomes and investment to fall, causing s_a to fall. At any time in this process it would be possible to calculate on the basis of the level of income, Y_t, and associated s and v, the rate of investment which, if adopted would produce warranted equilibrium growth from that time onwards. Although it is highly unlikely that the economy would adopt this rate, it serves as a benchmark with which to compare the actual behaviour of the economy and thus to predict the direction of its subsequent cyclical movements.

There will thus be a different, but unique, value of g_w for every actual position of the system as it develops through time. Only if g_w is in fact attained will the economy exhibit stable, non-cyclical growth, while departures from the rate will not set up self-correcting movements to instantly restore it.

These two aspects of Harrod's theory have caused much misunderstanding. The fact that there is only one 'unique' or 'knife-edge' equilibrium growth path for any given t and condition of the system has led some economists to consider this as the main cause of instability. Yet Harrod himself considered 'instability' to be an inherent property of the general concept of dynamic equilibrium as represented by the warranted growth rate. Since there would be only one warranted rate for any given condition of the economy it could be used to explain the cyclical behaviour of the economy if g_a diverged from g_w. But in Harrod's theory there would be a new warranted rate for every new combination of Y, s and v thrown up by the actual growth of the economy; g_w was only unique because each point in time was characterized by unique conditions. The role of the instability property of the warranted rate, given the natural rate, was to explain how the system would move when it was not growing at its dynamic equilibrium rate.

Domar (1946, 1947), writing after the publication of the *General Theory*, reacted to a specific problem in Keynes's theory, pointing out that the very investment expenditure that provides the demand for the output of existing productive capacity implies increased productive capacity in future periods. Investment as a means of increasing aggregate demand is thus a 'mixed blessing', for if the investment sufficient to prevent unemployment today creates excess capacity tomorrow then even more investment will be required tomorrow. Long-run unemployment could be avoided only by increasing investment at an increasing rate. To analyse this problem it was inevitable that Domar recast Keynes's analysis in terms of rates of change.

Domar approached the problem by separating the influence of investment on aggregate demand and on productive capacity or supply. Keynes had already provided the analysis of demand in terms of the multiplier ($k = 1/s$) giving the expansion in demand resulting from increasing investment as $\mathrm{d}Y_d/\mathrm{d}t = k(\mathrm{d}I/\mathrm{d}t)$. On the supply side, however, since all of net investment, and not only the increase, expands productive capacity Domar amends Keynes's approach and considers the fraction of the labour force employed as a function of the ratio of income to *potential* productive capacity rather than as a simple function of income. Defining α as the net value added produced by a £ of net investment, potential productive capacity will then increase by αI where I is the aggregate cost of new investment projects. On the micro level, however, some new capacity will be competing with older capacity, and since some investment projects will be carried out on the basis of expectations which will not be realized, Domar defines σ as the 'potential social average productivity of investment' for the economy. The divergence between α and σ (as well as the assumption that $\alpha < \sigma$) thus represents errors in investment decisions, investment outpacing the growth in the labour force or investments incorporating inappropriate technology. The supply-side effect is thus $\mathrm{d}Y_s/\mathrm{d}t = \sigma I$. The answer to Domar's question of whether there is a constant rate of growth of investment at which the demand will rise sufficiently rapidly to offset the effect of investment on supply is thus found where $\mathrm{d}Y_d/\mathrm{d}t = \mathrm{d}Y_s/\mathrm{d}t$ or where $k(\mathrm{d}I/\mathrm{d}t) = \sigma I$. This equality can be rewritten as $(\mathrm{d}I/\mathrm{d}t\ /I = \sigma/k$ which Domar calls the 'required' rate of growth of investment.

Domar's assumption that unemployment is determined by the relation of income to potential capacity means that the 'required' rate implies full capacity utilization and thus full employment. The failure of the economy to grow at this rate implies excess capacity. If productive potential arising from net investment σI is defined as P, $\sigma = (\mathrm{d}p/\mathrm{d}t)/I$, then a coefficient of utilization determined by the relative expansion of demand and capacity can be defined as, $\theta = (\mathrm{d}Y_d/\mathrm{d}t)/(\mathrm{d}P/\mathrm{d}t)$ Since $\mathrm{d}Y_d/\mathrm{d}t = \mathrm{k}(\mathrm{d}I = \mathrm{d}t)$ and $\mathrm{d}P/\mathrm{d}t = \sigma I$, θ can be written as $(\mathrm{d}I/\mathrm{d}t)/I.\ k/\sigma$ assuming that $\alpha = \sigma$ If investment is expanding at the required rate

$$(\mathrm{d}I/\mathrm{d}t)/I = \sigma/k, \mathrm{d}Y_d/\mathrm{d}t = \mathrm{d}P/\mathrm{d}t \text{ and } \theta = 100 \text{ per cent capacity utilization.}$$

Domar's required rate is thus equivalent to Harrod's natural rate of growth (*sr/vr*)

When $\alpha = \sigma$ since $k = 1/s$ and $\sigma = (\mathrm{d}Y/\mathrm{d}t)/I = 1/vr$, $\sigma/k = s/vr$.

Domar's analysis of divergence of the actual growth rate from the 'required' rate also produces an analysis of instability, for when $(\mathrm{d}I/\mathrm{d}t)/I$ is below k/σ the required rate, $\mathrm{d}Y_d/\mathrm{d}t$ is less than $\mathrm{d}P/\mathrm{d}t$, so part $(1-\theta)$ of new productive potential is unused. This excess capacity thus implies the existence of unemployment. A higher rate of growth of investment would be required to eliminate the excess capacity and unemployment, but since current productive capacity is already excess to needs, entrepreneurs are more likely to try to reduce than to increase their desired capacity by lowering $(\mathrm{d}I/\mathrm{d}t)/I$, which will increase rather than decrease both unemployment and excess capacity, producing a slump. Thus, in difference from Harrod's analysis, the natural rate is a unique equilibrium or 'knife edge' rate as well as being unstable. For Domar instability is not linked to the conceptual definition of dynamic equilibrium by means of a warranted rate, but rather to the 'paradox' that is the dynamic equivalent to the Keynesian paradox of saving: given s, the elimination of excess capacity, whether it is caused by the effects of investment on the expansion of demand or productive capacity, requires more capital to be built, while a shortage of productive capacity requires a reduction in the rate of growth of investment. This result is parallel to Harrod's statement to the effect that a general glut of commodities is due to entrepreneurs producing too little rather than too much.

While both Harrod and Domar sought to use the concepts of warranted and natural or required rates as an aid to understanding the cyclical implications of Keynes's analysis, and despite the differences in their approach, their work served to form the basis of what came to be known as the 'Harrod–Domar' theory of steady growth. By interpreting the variables *s* and *v* as being given exogenously the theory produced what Kaldor (1951) called 'Harrod's problem', or as Joan Robinson (1965, p. 52) put it:

> Given *s*,. . . and *v*,. . . *g* is determined. There is only one value of *g* which (provided it does not exceed *n*) is not impossible. The uniqueness of *g*, not anyquestion about the stability of the corresponding growth path, created the problem of the 'knife edge'.

This 'problem' was 'resolved' by introducing differential savings propensities from wages and profits to make *s* a variable determined by the distribution of income, which would allow multiple long-period unemployment growth equilibria, as in the post-Keynesian theories of growth and distribution. Alternatively (cf. e.g. Solow 1970, ch. 2), if movements in relative prices of capital and labour services are allowed to produce substitution of capital for labour, as in an aggregate production function, then *v* would become variable over time and lead to the full employment of both factors, despite Domar's (1952, pp. 23–6) explicit warning that the introduction of a Cobb–Douglas production function to solve this problem would lead directly to this traditional pre-Keynesian result.

These two conflicting interpretations of the applicability of Keynes's unemployment equilibrium in the long period, soon enlarged to include the wider question of capital theory, created a debate in which steady state theories overwhelmed the interests of both Harrod and Domar in the implications of Keynes's theory for the problem of economic fluctuations and dynamics.

See Also

- ▶ Aggregate Demand Theory

Bibliography

Domar, E.D. 1946. Capital expansion, rate of growth, and employment. *Econometrica* 14 (April): 137–147. Reprinted in Domar (1957), 70–82.

Domar, E.D. 1947. Expansion and employment. *American Economic Review* 37 (March): 34–55. Reprinted in Domar (1957).

Domar, E.D. 1952. Economic growth: An econometric approach. *American Economic Review, Papers and Proceedings* 42 (May): 479–495. Reprinted as 'A theoretical analysis of economic growth', in Domar (1957).

Domar, E.D. 1957. *Essays in the theory of economic growth*. New York/Oxford: Oxford University Press.

Harrod, R.F. 1936. *The trade cycle*. Oxford: Clarendon Press.

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

Harrod, R.F. 1948. *Towards a dynamic economics*. London: Macmillan.

Harrod, R.F. 1959. Domar and dynamic economics. *Economic Journal* 69 (September): 451–464.

Kaldor, N. 1951. Mr Hicks on the trade cycle. *Economic Journal* 61 (December): 833–847.

Keynes, J.M. 1973. In *The collected writings of J.M. Keynes*, Vol. 14: The general theory and after – Part II, defence and development, ed. D. Moggridge. London: Macmillan.

Robinson, J. 1965. Harrod's knife edge. In *Collected economic papers*, ed. J. Robinson, vol. 3. Oxford: Basil Blackwell.

Solow, R.M. 1970. *Growth theory: An exposition*. Oxford: Clarendon Press.

Natural Experiments and Quasi-Natural Experiments

J. DiNardo

Abstract

Natural experiments or quasi-natural experiments in economics are serendipitous situations in which persons are assigned randomly to a treatment (or multiple treatments) and a control group, and outcomes are analysed for the purposes of putting a hypothesis to a severe test; they are also serendipitous situations where assignment to treatment 'approximates' randomized design or a well-controlled experiment.

Keywords

Experiment; Natural experiments; Quasi-natural experiments; Randomization; Regression discontinuity design; Returns to schooling; Social experiments; Treatment effect

JEL Classifications

C42

The term 'natural experiment' has been used in many, often, contradictory, ways. It is not unfair to say that the term is frequently employed to describe situations that are neither 'natural' nor 'experiments' or situations which are 'natural, but not experiments' or vice versa.

It will serve the interests of clarity to initially direct most of our attention to the second term – experiment. A useful, albeit philosophically charged definition of an experiment 'is a set of actions and observations, performed in the context of solving a particular problem or question, to support or falsify a hypothesis or research concerning phenomena' (Wikipedia 2006).

With such a broad definition in hand, it may not be surprising to observe a wide range of views among economists about whether or not they perform experiments. Vernon Smith, for example, in experimental methods in economics, begins with the premise that 'historically, the method and subject matter of economics have *presupposed* that it was a *non–experimental … science more like astronomy or meteorology than physics or chemistry*' (emphasis added). As he makes clear, his observation implies that *today*, economics is an experimental science. Bastable's article on the same subject in the first edition of *The New Palgrave* overlaps only superficially with Smith's and divides experiments along the lines suggested by Bacon: *experimenta lucifera*, in which 'theoretical' concerns dominate, and *experimenta fructifera*, which concern themselves with 'practical' matters. In sharp contrast to Smith, Bastable concludes that *experimenta lucifera* are 'a very slight resource' (1987, p. 240) in economics.

These two views of experiment, however, do not seem helpful in understanding the controversy regarding natural experiments. 'Experiment' in our context is merely the notion of putting one's view to the most 'severe' test possible. A good summary of the the spirit of experiment (natural or otherwise) comes from the American philosopher Charles Sanders Peirce (and see Mayo 1996 for a nice exposition of this and related points):

> [After posing a question or theory], the next business in order is to commence deducing from it whatever experimental predictions are extremest and most unlikely . . . in order to subject them to the *test of experiment*.
>
> The process of testing it will consist, not in examining the facts, in order to see how well they accord with the hypothesis, but on the contrary in examining such of the probable consequences of the hypothesis as would be capable of direct verification, especially those consequences which would be very unlikely or surprising in case the hypothesis were not true.
>
> When the hypothesis has sustained a testing as severe as the present state of our knowledge ... renders imperative, it will be admitted provisionally ... subject of course to reconsideration. (Peirce 1958, 7.182 (emphasis added) and 7.231 as cited in Mayo 1996)

The Philosophy of Experimentation in Natural Science

In the emergence of modern natural science during the 16th century, experiments represented an important break with a long historical tradition in which observation of phenomenon was used *in* theories as a way to justify or support a priori reasoning. In Drake's (1981) view: 'The Aristotelian principle of appealing to experience had degenerated among philosophers into dependence on reasoning supported by casual examples among philosophers and the refutation of opponents by pointing to apparent exceptions not carefully examined.' In the useful historical account provided by Shadish et al. (2002) it is suggested that this 'break' was twofold: first, experiments were frequently employed to correct or refute theories. This naturally led to conflict with political and religious authorities: Galileo Galilei's conflict with the Church and his fate at the hands of the Inquisition is among the best-known examples of this conflict. Second, experiments increasingly involved 'manipulation' to learn about 'causes'. Passive observation was not sufficient. As Hacking (1983, p. 149) says of early experimenter Sir Francis Bacon: 'He taught that not only must we observe nature in the raw, but that we must also "twist the lion's tale", that is, manipulate our world in order to learn its secrets.'

Indeed, at some level in the natural sciences there has been comparatively little debate about the centrality of experiment – ironically, it has typically been only philosophers of science who have downplayed the importance of experiment. Hacking (1983) makes a strong case that philosophers typically have exhibited a remarkably high degree of bias in minimizing their importance in favour of 'theory'. Until the 19th century, the term experiment was typically reserved for studies in the natural sciences.

In the low sciences such as economics and medicine, the role of experiment is been the subject of extensive debate, much tied up with the debate on whether all the types of experiments possible in real science are possible in economics as well as with debates about the many meanings of the word 'cause'.

A key distinction between much real science and economics involves the centrality of 'randomization'. No randomization is required, for example, to study whether certain actions will produce nuclear fission, since 'control' is possible: if a set of procedures applied to a piece of plutonium – under certain pre-specified experimental conditions – regularly produces nuclear fission, as long as agreement exists on the pre-specified conditions and on what constitutes plutonium, and so on, it is possible to put the implied propositions to the type of severe test that would gain widespread assent – all without randomization. Put in a different way, randomization is required only when it is difficult to put a proposition to a severe test without it.

A related issue is whether a study of 'causes' requires some notion of 'manipulation'. Most definitions of 'cause' in social science involve some notion of 'manipulation' (Heckman 2005) – Bacon's 'twisting of the tail', so to speak. In physics, by way of contrast, some important 'causes' do not involve manipulation per se. One might argue that Newton's law of gravitation was an example of a mere empirical regularity that became a 'cause'. Indeed, when proposed by Newton, Leibnitz objected to this new 'law': in the prevailing intellectual and scientific climate where the world was understood in terms of 'mechanical pushes and pulls', this new law seemed to require the invocation of 'occult powers' (Hacking 1983). (There is an element of irony in Leibnitz's objection. Leibnitz is believed by some to be the object of Voltaire's satire as the character Dr. Pangloss in *Candide* of whom it is said that he 'proved admirably that there is no effect without a cause ... in this the best of all possible worlds' – a very different notion of causation! Voltaire 1759, ch. 1.)

In this article, we take the view that, even if manipulation were not necessary to *define* causality, manipulation is central to whether it is possible to discuss the idea intelligibly in social sciences and whether some kind of 'severe test' is possible (DiNardo 2007). Some philosophers have sought to *define* science around issues related to 'control', arguing that the phenomena economists try to investigate are impossible to study scientifically at all. Philosophers have articulated numerous reasons for the difference between social and natural science. A few examples may be helpful: Nelson (1990, pp. 102–6) argues, for example, that the objects of enquiry by the economist do not constitute 'a natural kind'. Put very crudely, the issue is the extent to which all the phenomena that we lump into the category 'commodity', for example, can be refined to some essence that is sufficiently 'similar' so that a scientific theory about commodities is possible in the same way as a 'body' is in Newtonian mechanics. This is often discussed as the issue of whether the relevant taxonomy results in 'carving nature at the joints'. Hacking (2000) introduces the notions of 'indifferent kinds' – the objects in the physical science – atoms, quarks, and so on with 'interactive' kinds – the objects of study in medicine or the social sciences. We might interact with plutonium or bacteria, but neither the plutonium nor the bacteria are aware of how we are classifying them or what we are doing to them. This can be contrasted with 'interactive kinds' that are aware and for which 'looping' is possible. For example, mental retardation might lead to segregation of those so designated. This segregation might lead to new behaviours which then might not fall under the old label, and so on. Consequently, investigation of such phenomena might be likened to 'trying to hit a moving target'. Searle (1995) on the other hand, notes that the objects of interest in social science while epistemologically objective, are ontologically subjective. While the loss of 100 dollars may be very 'real' to someone, the notion of money requires groups of individual to accept money as a medium of exchange. Again the existence of atoms does not require us to recognize their existence.

Randomization: An Attempt to Evade the Problems of Imperfect 'Control'

If one accepts the centrality of manipulation (or something like it), it will not be surprising that the application of principles of experimentation to humans who have free will, make choices, and so on entails a host of issues that, inter alia, sharply constrain what might be reasonable to expect of experiments, natural, or otherwise.

If it is not possible, desirable, or ethical to 'control' humans or their 'environment' as it sometimes is in the natural sciences, is it possible to learn anything at all from experiment broadly construed? *Randomization* in experiments developed in part to try to evade the usual problems of isolating the role of the single phenomenon in situations. In the 19th century, it was discovered that by the use of 'artificial randomizers' (such as a coin toss) it was possible, in principle, to create two groups of individuals which were the same 'on average' apart from a single 'treatment' (cause) which was under (at least partial) control of the experimenter. Hacking (1988, p. 427) has observed that their use began primarily in contexts 'marked by complete ignorance': the economist F. Y. Edgeworth was early to apply the mathematical logic of both Bayesian and 'classical' statistics to a randomized trial of the existence of 'telepathy'.

Although economists played an important role in the development of randomization, economists as a whole were quite slow to embrace the new tools. In an echo of debates that faced natural sciences in the 1600s, this was due in part 'because the theory [of economics] was not in doubt, applied workers sought neither to verify nor to disprove' (Morgan 1987, pp. 171–2).

Over time, the term 'experiment' evolved to include both experiments of the 'hard sciences' where a measure of control was possible as well as situations in which artificial randomizers were used to assign individuals (or plots of land, and so on) to different 'treatments'. A key role was played by R. A. Fisher (1935) and his seminal *Design of Experiments* as well subsequent publications which discussed the theory and practice of using artificial randomizers to learn about causes.

There are at least two key limitations of randomized experiments relative to experiments where 'scientific' control is possible:

- Without real control, one only has a weak understanding of the 'cause' in question. For instance, one can do a randomized controlled trial of the effect of aspirin on heart failure while understanding nothing of the mechanism by which aspirin affects the outcome. Moreover, it is clear that the experiment is 'context specific'. One's generalization about atoms in a laboratory often extends to atoms in other contexts in a way not possible in social science.
- Any single experiment – even under the ideal situation – does not always reveal the true answer. In the logic of randomized design, the usual inference procedure is merely one that *would* give the right answer on average *if* the experiment were repeated. At best, the true answer is just a 'long-run tendency' in repeated identical experiments.

Social Experiments: Why not do a 'Real' Randomized Trial?

Even without these limitations, there is a long list of reasons why economists frequently have little interest in randomized trials. The most important reason is that many of the real randomized experiments (often called 'Social experiments') of which one could conceive (or have been implemented), are immoral or unethical. At a most basic level, the decision as to who 'performs an experiment' and who 'decides' or is recruited to be experimented upon often reflects deep-seated social injustice. Even Brandeisian (see below) experiments can take on a sinister cast – state governments surely do not consider the interests of all their citizens equally.

Indeed, historically the conduct of experiments on persons has told us as much or more about the structure of society than anything else: one well-known example is the series of 'experiments' conducted by the US Public Health Service from 1932 to 1972 on about 400 poor black men who had advanced syphilis. One aim of the experiment was to determine the effect of untreated syphilis. To this end, the medical doctors misrepresented themselves to the subjects (the sons and grandsons of slaves), claiming to provide free medical care. For example, when penicillin became the standard of care, the subjects were deliberately not provided with the medication: rather, the doctors were content to observe the horrific progress of the disease as some went blind or insane.

Another set of reasons is practical – experiments are costly to administer. Another reason is attrition: often people drop out of such experiments (often in non-random ways), greatly complicating the problem of inference. A distinct, although sometimes related, reason is that the results of social experiments involving randomization are sometimes difficult to interpret. One often cited reason is that those recruited to participate in such experiments may be different from those for whom the policy is ultimately intended. In even the simplest experiments, 'compliance' is imperfect. Not everyone assigned to a treatment takes it up – indeed, it is often the case that analysis is made on an 'intent to treat' basis. That is, those 'assigned' to treatment are compared to those assigned to the control whether or not those assigned to treatment actually 'took' the treatment. Another often cited reason is that what is likely when a social experiment is conducted with a small number of persons might be very different when applied to much larger numbers of persons. Persons, unlike atomic particles, enjoy free will. In the world of persons, the 'experiment' does not necessarily stop after the experimenters have made their observations. For example, even in the context of a true randomized experiment, those denied treatment often have the opportunity to find it elsewhere (see Heckman and Smith 1995, with references, for one discussion of the merits of randomized trials in the social science).

Types of Natural Experiments

Thus far we have seen that the word 'experiment' can be used in two very different senses: one to denote situations where real 'control' is possible and second involving artificial randomizers. As a consequence, the term 'natural experiment' has been used in very different senses. I now turn to the origins of the term and the different ways the term has been used, although we focus on natural experiments most frequently arising in economics.

Natural Experiments in Natural Science

An early use of the term 'natural experiment' in English describes an investigation into the functioning of 'nature'. The term comes from a translation *Saggi di naturali esperienze fatte nell'Accademia del Cimento* published in Italian in 1667 which appeared in an English translation by Richard Waller in 1684 as *Essayes of natural experiments made in the Academie del Cimento* (Waller 1684). The short-lived Accademia del Cimento was founded in Florence in 1657 by the Medici brothers, Prince Leopold and Grand Duke Ferdinand II, and the *Saggi* record a small subset of the large number of experiments by the Cimento that involved such issues as 'smells do not traverse Glass', and 'the failure to confirm Existence of Atoms of cold' (1684, p. xx). Although the experiments of the Academy included trials involving humans, they did not involve randomization. Indeed, the legacy of these investigations into humans is more relevant to the study of 16th-century culture and authority relations than 16th-century science. (Tribby 1994, for example, discusses an investigation into a 'gentler' laxative that could 'satisfy' the needs of Grand Duke Ferdinand II as well as those of the many 'delicate persons' who visited or had dealings with the court that involved experimentation on individuals described variously as 'a mercenary', 'a vagrant', 'the Little Moor', and so on.)

Over time, in the hard sciences, the term natural experiment has also come to describe both cases where 'nature' provides an experiment that resembles the controlled situation that scientists would like observe but are unable to create themselves. An unsuccessful experiment may help make the point clear: in a famous quote by Albert Einstein to Erwin Findlay Freundlich (who was attempting to assess the whether path of a ray of light was affected by gravity), Einstein wrote: 'If only we had a considerably larger planet than Jupiter! But nature has not made it a priority to make it easy for us to discover its laws.' ('Wenn wir nur einen ordentlich grösseren Planeten als Jupiter hätten! Aber die Natur hat es sich nicht angelegen sein lassen, uns die Auffindung ihrer Gesetze bequem zu machen', (as cited in Ashtekar et al. 2003; translation from the *New York Times*, 24 March 1992).

Natural Experiments as Serendipitous Randomized Trials

In contrast to the natural experiment of the hard sciences, the term natural experiment is often used by economists to denote a situation where real randomization was employed, without the intent of providing a randomized experiment. For example, between 1970 and 1972 men from specific birth cohorts were conscripted into the US military by way of a draft lottery. Each day of the year was randomly assigned a number which (in part) determined whether or not one was at risk of being inducted into the military service to fight in the US war on Indochina. As a consequence, men of specific birth cohorts born only a day apart, for example, had very different risks of serving in the military. In Hearst, Newman and Hulley (1986), the authors asked whether the war continued to kill after the warrior returned home. The authors compared, among other things, the suicide rates among individuals who on average were *ex ante* similar, but who had very different probabilities of having completed military service.

The example is sufficiently simple to make a number of points about the limitations of natural experiments. *If* one can assume that the mere fact of having such a birth date put one at high risk of military duty, and that having a birth date raised (or did not lower) any person's risk of serving in the military, then it is possible to use something akin to two stage least squares (2SLS) to estimate an 'average' effect of military service for those who were induced to serve in the military by the draft lottery. However Hearst et al. (1986) are quick to observe that *whether or not* one actually served in the military, the mere fact of having been put at risk of the lottery might have had an effect on delayed mortality. In econometric terms, this would be a violation of the 'exclusion restriction' of 2SLS. If such is the case, it is apparent that a comparison of men with high-risk birthdays to those with low-risk birthdays will be an admixture of the effect of the military service on later mortality *and* any direct effect of the lottery itself. An additional problem is the possibility of non-random selection induced by men dying while at war. This was judged to be small due since the fraction of US soldiers who died while serving in action was a small fraction of the total.

Returning to how one might go from an estimate generated in this way to more general inference, one has a number of other obstacles. For example, the delayed mortality effects of military service on those *induced* to serve by an unlucky birth date might be different from the effect on those who *volunteered* to fight in the war. If the effects are very different, it would obviously be incorrect to use estimates generated by those induced to serve to extrapolate to the broader population of interest.

More generally, our ability to generalize the valid results of an experiment is much more limited when we can only manipulate the cause indirectly (as in the example above) than when we can manipulate the cause directly: there is often the possibility of important differences between persons who take up the treatment as a result of having been encouraged to participate and those who were similarly encouraged but did not take up the treatment.

The Regression Discontinuity Design as a Natural Experiment

One research design that involves the 'serendipitous' randomization of individuals into a treatment is called the regression discontinuity design. Since it is a relatively 'clean' example of something that approaches a truly randomized experiment without involving explicit randomization, it provides a good illustration of the strengths and weaknesses of natural experiments. (For an analysis of the relationship between the regression discontinuity design and randomized controlled trials see Lee, 2007.) For illustration, let us consider DiNardo and Lee's (2004) analysis of the causal effect of 'unionization' on firms in the United States. The naive approach would be to compare unionized firms to non-unionized firms.

The basis of the regression discontinuity design is the existence of a 'score' or a 'vote' which assigns persons to one treatment or another. In the US context, workers at a firm can win the right to form a labour union by means of a secret ballot election. If 50 per cent plus one of the workers votes in favour of the union, the workers win the right to be represented by a union; less than that, and they are denied such rights.

To understand how this works, consider elections at two different sets of work sites that employ large numbers of workers. In one set, $0.5 + \Delta$ of the workers vote in favour of the union and win the right to bargain collectively where Δ is some small number. In another set, slightly less than 50 per cent vote in favour of the union, and are denied the right to bargain collectively. The vote share in these sites is $0.5 - \Delta$. Suppose we have large amounts of data on such elections and can accurately estimate the average outcome (say the fraction of firms that continue to exist 15 years after the vote).

Using almost exactly the same set-up as before, we compare those places where the union wins with those where the union loses:

$$E[\bar{y}_{\text{Union}} - \bar{y}_{\text{No Union}}] = E[y|\text{vote} = 0.5 + \Delta] - E[y|\text{vote} = 0.5 - \Delta]$$

If firm survival is described by the same 'model' as in å above, where now $T = 1$ denotes winning the right to bargain collectively, we get:

$$E[\bar{y}_{\text{Union}} - \bar{y}_{\text{No Union}}] = \beta + \left(\underbrace{E[f(X)|\ \text{vote} = 0.5)\Delta] - E[f(X)|\ \text{vote} = 0.5 - \Delta]}_{\text{Observable Differences}}\right) + \left(\underbrace{E[\varepsilon|\ \text{vote} = 0.5 + \Delta] - E[\varepsilon|\ \text{vote} = 0.5 - \Delta]}_{\text{Unobservable Differences}}\right)$$

The 'trick' is that if we choose Δ to be small enough (that is, close to zero), then

$$E[f(X)|\text{vote} = 0.5 + \Delta] \approx E[f(X)|\text{vote} = 0.5 - \Delta] \text{ and } E[\varepsilon|\text{vote} = 0.5 + \Delta] \approx E[\varepsilon|\text{vote} = 0.5 - \Delta]$$

and we get a 'good' estimate of the 'effect of unions' in the same sense that we get a good estimate of the effect of a treatment in a randomized controlled trial. That is, if we focus our attention on the difference in outcomes between 'near winners' and 'near losers' such a contrast is formally equivalent to a randomized controlled trial if there is at least some 'random' component to the vote share. For example, sometimes people take ill on the day of the vote – if that happens randomly in some sites, two sites that would have had the same final vote tally had everyone shown up are now different. When such differences are the difference between recognition or not, one has the practical equivalent of a randomized controlled trial. The mere existence of a 'score' that discontinuously exposes one to a treatment is not enough. This design would not be appropriate, for example, to analyse the causal effects of US Congressional votes on various issues. Substantial 'manipulation' – that is, through negotiation, and so on – of the final vote tally is common and suggests that individuals near but on opposite sides of the threshold are not otherwise similar (see regression-discontinuity analysis).

A few moments' reflection will make clear both the appeal of such experiments and their limits. Advocates of a natural experiment approach point to the fact that the implicit randomization involved in this design means that we can be more confident with such a comparison than a naive comparison that merely compares unionized to non-unionized firms. This would almost certainly confound the true 'effect' with pre-existing differences in unionized and non-unionized firms with 'unionization'. Advocates will also point to the fact that the experiment is relevant to a potential policy – say lowering the threshold required to win representation rights by a small amount.

Detractors will observe many limitations. Is the effect of a union that is set into a place by a 51 per cent vote the same as the effect of a union where the workers vote unanimously? Possibly not. Stipulating the validity of the estimate, is it reasonable to suggest that the effect of unionization would be the same if all workplaces were allowed to vote on a union? Probably not. Is it possible that a union at one work site affects other work sites? What about the effect on the firm's competitors? Indeed, it is even possible to question the premise that a union is a 'treatment' at all. Does it make sense to talk of a single effect of a labour union when there is such heterogeneity in what the notion 'labour union' represents? While the anarcho-syndicalist Industrial Workers of the World (IWW) of Joe Hill (a famous militant IWW member and subject of a well-known folksong) and the American Federation of Labor and Congress of Industrial Organizations (AFL-CIO) of George Meany (a conservative 'anti-communist' who was its president for many years) were both labour unions, they had virtually contrary aims and wildly different political structures.

More generally, 'causes', 'treatments', and so on are much more fragile objects for the types of things usually interesting to economists than the types of things interesting to natural science. The concepts of natural science are often capable of quite substantial refinement in a way that concepts in the human sciences rarely are.

'Natural Natural Experiments'?

As I have already mentioned, the term 'natural experiment' has been used in several different ways inconsistent with our definition. It seems pointless, however, to claim that our definition is the 'true' or correct one. We shall therefore consider some cases that use the term which do not obviously involve randomization of a treatment or something that approximates such randomization.

Rosenzweig and Wolpin (2000) for instance, have coined the expression 'natural natural experiments' to denote a wide range of studies involving the use of twins. The emphasis on the word 'natural' is intended to highlight the role of nature in providing the variation. Twins have been of inordinate interest to the social scientists since they seem to offer the possibility of 'controlling' for 'genetics'. Consider one case of interest to economists, 'returns to schooling'. Does acquiring an additional year of school result in higher wages in the labour market? How much higher? To fix ideas consider a simple model of the sort:

$$y_{ij} = \beta S_{ij} + a_j \ + \varepsilon_{ij}.$$

We are interested in some outcome, say hourly wages, and the causal effect of years of schooling S. It will greatly simplify the discussion if we assume that all persons 'treated' with 'schooling' experience the same increase in their wages – that is, the treatment effect is a constant across individuals. We have gathered a random sample of $j = 1,..., J$ 'identical' (monozygotic) twins ($i = 1, 2$). The term a_j is not directly observable but includes everything that the twins have in common – genetics, environment, and so on. The error term ε_{ij} includes everything that the twins do not have in common and cannot be observed as well as the effects of misspecification, and so on. Though this simple set-up can be greatly elaborated (see Ashenfelter and Krueger 1994, for a clear exposition) the essential idea is that the *difference* between the twins purges the outcome of the a_j term so that an ordinary least squares regression of the difference in wages Δy_{ij} on ΔS_{ij} yields a good estimate of

$$\hat{\beta} \text{ is a good estimate of } \beta + \frac{\text{Var}\Delta\varepsilon, \Delta S}{\text{Var}(\Delta S)}.$$

The first term is the goal of such studies. The second term points to the possibility that there are other influences which might be correlated both with schooling and that affect the outcome. The second term can be interpreted as the slope coefficient from the following hypothetical ordinary least squares (OLS) regression, where δ is the slope of the 'best-fitting' line in this expression:

$$\varepsilon = \text{constant} + S\delta + \text{error}.$$

When will $\hat{\beta}$ to be a good estimate of the returns to schooling β? The conditions are essentially the same as for the randomized controlled trial: if we can treat the assignment of schooling to the two twins as if it were determined by a random coin toss then differences in the level of schooling between the two twins – ΔS_{ij} – will be independent of differences between the two twins in unobserved influences on wages – $\Delta\varepsilon_{ij}$. Detractors of this approach doubt that such an assumption is plausible. In simple language, if the twins are so 'identical' why do they have different levels of schooling? Perhaps the parents noticed that one twin was more interested or had more 'aptitude' for schoolwork than another. If that were the case, estimates of the returns to schooling would be confounded with differences in the aptitude for schooling despite the fact that we had 'controlled' for a large number of other factors. The key difference between this case and what I have identified as a natural experiment is the lack of an obvious approximation to randomization. Bound and Solon (1999) discuss, inter alia, a host of difficulties in treating twin differences as experimental variation. I do not discuss twins studies that utilize twins as a 'surprise' to family size which have some element of randomization.

Other Research Designs. Quasi-Experiments

Finally, I should make note of the fact that some authors use the term natural experiment more broadly than I have construed it here. Meyer (1995, p. 151) for instance, considers natural experiments the broad class of research designs 'patterned after randomized experiments' but not (generally) involving actual randomization. One term often used for such situations is 'quasi-experiment'. The relationship between these quasi-experiments and the natural experiments I have been describing is quite varied and ranges from those whose difference from the standard of randomized assignment is merely a matter of 'degree' to those in which assignment to treatment differs so much from the standard of randomization that it is really a difference in 'kind'.

Most of these quasi-experiments are variants of a 'before and after' where an observation is made before and after a treatment. Often a before–after comparison for one set of observations (the treatment – *T*) is compared to another set (the control – *C*). A typical set-up might compute a treatment effect by taking the difference in two differences:

$$\begin{aligned}\text{Treatment Effect} = &\left\{\bar{y}_{T,\,\text{after}} - \bar{y}_{T,\,\text{before}}\right\}\\ &- \left\{\bar{y}_{C,\,\text{after}} - \bar{y}_{C,\,\text{before}}\right\}.\end{aligned}$$

For this reason, such quasi-experiments are described as using 'difference-in-differences' approach to identifying a causal relationship.

In the United States, the fact that the state (or city) governments have some liberty to enact laws independently of the federal government, for example, has led to a great deal of research using 'Brandeisian' experiments. The term comes by way of US Supreme Court Justice Louis Brandeis, in the case *New State Ice* v. *Liebmann*:

> There must be power in the States and the Nation to remould, through experimentation, our economic practices and institutions to meet changing social

> and economic needs. ... It is one of the happy incidents of the federal system that a single courageous State may, if its citizens choose, serve as a laboratory; and try novel social and economic experiments without risk to the rest of the country. (U.S. Supreme Court *New State Ice Co.* v. *Liebmann*, 285 U.S. 262 (1932))

To give one such example, consider DiNardo and Lemieux's (2001) evaluation of the effect of changing the age at which it is legal to purchase alcohol or the consumption of marijuana. At the beginning of the 1980s states generally enforced two types of legal regimes. In one set, alcohol could not be legally sold to those under the age of 21. In another, the legal minimum drinking age (LMDA) was 18. In the mid-1980s, the federal government put a great deal of pressure on those states with LMDA of 18 to raise them to 21 and by the end of the 1980s, in all states drinking age was 21.

The assignment of drinking age statutes to the states at the beginning of the 1980s could not be considered 'approximately' random. Utah, for example, which is home to a large number of adherents to the Mormon religion – which proscribes alcohol use – had a 21-year drinking age at the beginning of the 1980s. However, due to a federal policy implemented in the mid-1980s of eventually denying federal highway funds to states with legal minimums less than 21 years old, something perhaps approximating an 'experiment' can be arrived at by comparing *changes* in alcohol or marijuana consumption during the 1980s in those states which were forced to change (and changed early) with those who were forced to but raised their drinking age later.

Let Δy_t denote the change in the fraction of 18–21 year olds who reported smoking marijuana in the previous 30 days from 1980 to 1990 in states that had 18-year-old drinking ages that were increased, and Δy_c denote the similar change in states whose drinking age was always 21. Then an estimate of the effect of the drinking age might be:

$$\Delta y_t - \Delta y_c = \text{Effect of LMDA}.$$

Although randomization is not employed per se, the credibility of these exercises can be at least partially evaluated. For instance, if the outcome of interest has been approximately constant in both the treatment and control groups for a long time preceding the change in legal regime, the estimate is generally more credible. Less credible is the case in which the outcomes in the control group and the treatment group are quite variable over time, the control group and the treatment group do not follow similar patterns *before* the proposed experiment, or when both are true.

Controversies: Concluding Remarks

Natural experiments and their like have been at the heart of much work in economics. Nonetheless, they are the subject of considerable debate. One of the most cited limitations of natural experiments – by both supporters and detractors – is that such experiments are context specific. Indeed, one frequently encountered 'strength' of natural experiments is that it often concerns the evaluation of an actual policy. There are limitations, however. If we assume that the experiment is 'internally valid' we still have to ask: how do we generalize from one experiment to the broader questions of policy? The foregoing has suggested that it is difficult. There are at least three broad classes of reasons:

1. While a natural experiment might provide a credible estimate of some particular serendipitous 'intervention', this may have only a weak relation to the type of interventions being contemplated as policies. Many of the potential reasons for a weak relationship are similar to those encountered in social experiments (among other things, for example, the effect of a treatment in a demonstration programme might be quite different from the outcome that would obtain if the treatment were applied more broadly or to different persons).
2. Some interesting questions are unanswerable with such an approach because serendipitous randomized experiments are few and far between. The extent to which this criticism is warranted, of course, depends on the availability of alternative ways of putting our views to a severe test.

3. More generally, without a 'theory', estimates from natural experiments are uninterpretable.

I am sympathetic with all three criticisms although (3) deserves some qualification. While it has been argued that even in the natural sciences it is impossible to have 'pre-theoretical' observations or experiments, Hacking (1983) makes a strong case that experimentation has a life of its own, sometimes suggesting ideas in advance of theory, other times the consequence of theory, and sometimes testing theories. Much of this debate in the natural sciences revolves around the notion of what constitutes a 'theory'. Whatever the validity of the view that one cannot experiment in advance of 'theory' in the natural sciences, in the social sciences, it is clear that no theory has the same standing as, say, general relativity in physics. This is the sense in which Noam Chomsky observes that 'as soon as questions of will or decision or reason or choice of action arise, human science is pretty much at a loss' (Magee 2001, 184). Indeed, the standing of randomized experiments – in some fields of enquiry regarded as 'the gold standard' of evidence – is a great deal lower than the best experiments of natural science; they are most often useful in situations otherwise marked by 'complete ignorance' (Hacking 1988). In short, while the human sciences might have the same ambition as natural science, the status of what we know will almost surely be quite limited.

Nonetheless, one does not need a 'correct' theory to hand, nor an understanding as rich as that found in some of the natural sciences to find an experiment useful. At the risk of over-using such metaphors, the fact that the Michelson–Morley experiments were in part about testing for the existence of 'ether' did not make them uninteresting. Experiments are just ways to use things we (think we) understand to learn about something we do not. And while the sorts of 'natural' experiments 'serendipitously' provided by society may be very limited and are often the product of unhappy social realities, they can sometimes perhaps serve a small role in enhancing our understanding.

Any assessment of the usefulness of natural experiments depends on how one judges the power of other methods of enquiry. Such a discussion is well beyond the scope of this article. Nonetheless, not discounting their many limitations, one benefit of natural experiments I have tried to highlight is that for some they might open up the possibility of revising their beliefs in light of evidence or suggest new ways to think about old problems, however limited. A key aspect of experiments (natural or otherwise) is the willingness to put one's ideas 'to the test'. Often, careful study of a natural experiment, however limited, may also make one aware of how complicated and difficult are the problems we call 'economics'. Even if the success we might have in generalizing natural experiments more broadly may be quite limited, if they bring nothing but humility to the claims social scientists make about much we actually understand, that alone would justify an interest in natural experiments.

See Also

- ▶ Difference-in-Difference Estimators
- ▶ Experimental Economics
- ▶ Experimental Economics, History of
- ▶ Experimental Labour Economics
- ▶ Experimental Methods in Economics
- ▶ Experiments and Econometrics
- ▶ Fisher, Ronald Aylmer (1890–1962)
- ▶ Regression-Discontinuity Analysis

Bibliography

Ashenfelter, O., and A.B. Krueger. 1994. Estimates of the economic returns to schooling from a new sample of identical twins. *American Economic Review* 84: 1157–1173.

Ashtekar, A., R.S. Cohen, D. Howard, J. Renn, S. Sarkear, and A. Shimony. 2003. *Revisiting the foundations of relativistic physics: Festschrift in honor of john Stachel*, Boston studies in the philosophy of science. Vol. 234. Dordrecht: Kluwer Academic.

Bastable, C.F. 1987. Experimental methods in economics (i). In *The New Palgrave: A Dictionary of Economics*, ed. J. Eatwell, M. Milgate, and P. Newman, Vol. 2. London: Macmillan.

Bound, J., and G. Solon. 1999. Double trouble: On the value of twins-based estimation of the return to schooling. *Economics of Education Review* 18: 169–182.

DiNardo, J. 2007. Interesting questions in freakonomics. *Journal of Economic Literature*.

DiNardo, J. and Lee, D.S.. 2002. The impact of unionization on establishment closure: A regression discontinuity analysis of representation elections. Working Paper No. 8993. Cambridge, MA: NBER.

DiNardo, J., and D.S. Lee. 2004. Economic impacts of new unionization on private sector employers: 1984–2001. *Quarterly Journal of Economics* 119: 1383–1441.

DiNardo, J., and T. Lemieux. 2001. Alcohol, marijuana, and American youth: The unintended consequences of government regulation. *Journal of Health Economics* 20: 991–1010.

Drake, S. 1981. *Cause, experiment, and science: A galilean dialogue, incorporating a new english translation of Galileo's bodies that Stay atop water, or move in it*. Chicago: University of Chicago Press.

Fisher, R.A. 1935. *Design of experiments*. Edinburgh/London: Oliver & Boyd.

Hacking, I. 1983. *Representing and intervening: Introductory topics in the philosophy of natural science*. Cambridge: Cambridge University Press.

Hacking, I. 1988. Telepathy: Origins of randomization in experimental design. *Isis* 79: 427–451.

Hacking, I. 2000. *The social construction of what?* Cambridge, MA: Harvard University Press.

Hearst, N., T.B. Newman, and S.B. Hulley. 1986. Delayed effects of the military draft on mortality: A randomized natural experiment. *New England Journal of Medicine* 314: 620–624.

Heckman, J.J. 2005. The scientific model of causality. *Sociological Methodology* 35: 1–97.

Heckman, J.J., and J.A. Smith. 1995. Assessing the case for social experiments. *Journal of Economic Perspectives* 9(2): 85–110.

Lee, D.S. 2008. Randomized experiments from non-random selection in U.S. house elections. *Journal of Econometrics*.

Magee, B. 2001. *Talking philosophy: dialogues with fifteen leading philosphers*. Oxford: Oxford University Press.

Mayo, D.G. 1996. *Error and the growth of experimental knowledge science and its conceptual foundations*. Chicago: University of Chicago Press.

Meyer, B. 1995. Natural and quasi-experiments in economics. *Journal of Business and Economic Statistics* 13: 151–161.

Morgan, M.S. 1987. Statistics without probability and Haavelmo's revolution in econometrics. In *The probabilistic revolution: Ideas in the sciences*, ed. L. Krüger, G. Gigerenzer, and M.S. Morgan, Vol. 2. Cambridge, MA: MIT Press.

Nelson, A. 1990. Are economic kinds natural? In *In Scientific Theories of Minnesota Studies in the Philosophy of Science*, ed. C. Wade Savage, Vol. 14. Minneapolis: University of Minnesota Press.

Peirce, C.S.. 1958. In Collected Papers, vols. 7–8, ed. A. Burks. Cambridge, MA: Harvard University Press.

Rosenzweig, M.R., and K.I. Wolpin. 2000. Natural 'natural experiments' in economics. *Journal of Economic Literature* 38: 827–874.

Searle, J. 1995. *The construction of social reality*. New York: Free Press.

Shadish, W.R., T.D. Cook, and D.T. Campbell. 2002. *Experimental and Quasi–Experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Tribby, J. 1994. Club Medici: Natural experiment and the imagineering of 'Tuscany'. *Configurations* 2: 215–235.

Voltaire. 1759. *The history of candide; or all for the best*, ed. C. Cooke. London, 1796.

Waller, R. 1684. *Essayes of natural experiments made in the academie del cimento, under the protection of the most serene Prince Leopold of Tuscany*. Facsimile edn, ed. R. Hall, trans. R. Waller. New York/London, 1964.

Wikipedia. 2006. Experiment. http://en.wikipedia.org. Accessed 28 Sept 2006.

Natural Law

N. E. Simmonds

It is not uncommon to find the term 'natural law' being applied to any philosophical theory that espouses a belief in the 'objectivity' of moral standards, or the possibility of moral knowledge. If we avoid this inflated usage, however, and seek to identify a natural law tradition that is to some extent distinct from other cognitivist moral theories, it is probably best to identify such a tradition in terms of three basic features. First, natural law theories regard morality as, in some sense, a body of precepts. Even if the theory has a broadly teleological character, it will not have a nakedly maximizing structure: rather, the teleology will serve to justify a body of rules or standards. Secondly, natural law theories take juridical equality as a fundamental assumption: men are assumed to be of equal standing before the law of nature. Even when the theory serves to justify unequal rights in the real circumstances of society, those unequal rights are justified by reference to principles that treat everyone equally. The tension between natural rights and positively established rights which is therefore implicit in the idea of juridical equality finds expression in the third basic feature of natural law theories: the way in which they approach the relationship between natural law and the positive law enacted by men. Natural law represents the ultimate objective foundation by reference to which positive laws must be evaluated. But positive law is nevertheless necessary, and is far more than just an imperfect reflection of natural law. Positive laws are required in part to induce compliance with standards that would not otherwise receive the obedience of weak or evil men; but they are required also to give concrete detail to the general requirements of natural law. Natural law may require, for example, that conduct in certain areas of social life should be co-ordinated, but it will not necessarily specify the precise form that that co-ordination should take: natural law therefore requires the existence of positive law as a body of publicly ascertainable rules making co-ordination possible.

Perhaps the most significant metamorphosis of the natural law tradition is to be found in the shift from the position of Aquinas, which achieved pre-eminence in the later Middle Ages, to the theories of Grotius and Pufendorf in the 17th century. Most commentators have been struck by the change in character that natural law theory undergoes over this period, but there has been less agreement about what features actually mark the essential difference. On one view, the 17th-century writers put forward a theory of natural rights rather than a theory of natural law. But, although the 17th-century theories certainly display a more individualistic character, this is not invariably associated with the development of a rights-based theory: Pufendorf, for example, takes 'duty' as his basic concept rather than 'right'. On another view, the 17th-century writers offer a secular theory which can be contrasted with the theocentric approach of Aquinas. For reasons that will be explained, this view must be rejected. A better way of comprehending the change of tone and approach that separates Aquinas from Grotius is by reference to the role that notions of 'good' play in their theories. For Aquinas, an account of what is good for man forms the central pillar around which an understanding of natural law must be constructed. The role of positive law is to provide for the good, thus considered. For Grotius and Pufendorf, on the other hand, the role of law is to provide a framework within which men who are self-seeking and who live in conditions of scarcity may live together in a social order that enables each to pursue his own good as he conceives it. Although it would clearly be absurd to portray writers such as Grotius and Pufendorf in the guise of fully fledged liberals making a dramatic break with the past, it is nevertheless some such change of emphasis and orientation that marks the distinctive character of the theories that emerged in this period.

Given the way in which natural law theories depend upon some deep notion of human equality, and yet frequently adopt a conservative standpoint towards the material inequalities of social life, various stratagems have been adopted in order to bridge the gap between ideal and reality. Thus, in 17th-century thought, a basic right to appropriate and enjoy the resources of the natural world is possessed by men equally, yet it serves to justify the unequal division of wealth and resources in established society. In Aquinas the tension appears and is resolved in a different form, within his central notion of the good. The Aristotelian view, that the best life for man is a life of philosophic contemplation accessible only to a leisured elite, is replaced in Aquinas by the idea that man's ultimate good lies in a beatific vision of God that is potentially accessible to everyone, but only in a life after death: the postulate of equality is preserved by moving its centre of gravity to another world. It is in this recurring tension between the ideal realm of equality and the material world of inequality that we find the basis for Marxist critiques of natural law theory and, indeed, of bourgeois legal thought more generally.

The orthodox position for the natural lawyers of the 17th century was that the content of natural law could be determined by reason, but that it derived its binding force from the divine will. The role of the notion of divine will within such theories was, in effect, to preserve a deontological character for natural law within a basically teleological form of argument. According to both Grotius and Pufendorf, reason shows us that human nature and circumstances being what they are, man can live in society only if certain basic rules are observed, e.g. rules defining and protecting rights of property. But this establishes only that such rules are requirements of utility: it does not show that they are requirements of natural law. Thus Pufendorf is careful to point out that, considered apart from the divine will, the precepts of natural law are merely ‘like the prescriptions of physicians for the regimen of health’ but are not laws (*De Officio Hominis et Civis*, 1682, 1.3.10). Actions are right and wrong (as opposed to wise and foolish) only in relation to a law: and a law, Pufendorf holds, presupposes the will of a superior. Natural law binds by virtue of the divine will. Given that we know certain rules to be necessary for social life, we know that such rules must be willed by God. Since God created our nature and fitted us with the capacities that make social life possible, it must be his will that we should live in society and observe those rules that are necessary for the existence of social life.

Grotius is often regarded as denying the role of the divine will in natural law, and he was so interpreted by Pufendorf, who attacked him on precisely this point. It is in fact unlikely that Grotius intended any such radical move away from the theo-centric approach. He says that natural law arguments would have a degree of validity even if God did not exist: but this may simply mean that the rules of natural law are not arbitrary but are founded on the nature of man and of his circumstances. In fact the idea of the divine will could not be so easily discarded, since it was employed in these theories to solve a number of fundamental problems. First was the question of how an action being obligatory differs from an action being one that we merely have good reason to perform. Second was the question of how moral reasons are related to prudential reasons: a problem that became particularly acute once morality was conceived of as a body of rules rather than as based on certain virtues as aspects of character. Lastly, and most significantly for our purposes, the notion of the divine will preserved a deontological character for natural law even while the reasoned arguments being offered were arguments of a basically utilitarian character. As we shall see, it was this feature of natural law thought that was later to bring about a dramatic transformation that some have seen as the death of natural law.

It might at first be thought that Hobbes represents an exception to the argument that 17th-century natural law theories ascribed a vital role to the divine will. There are of course large questions about whether Hobbes forms part of the natural law tradition at all. But it should be noted that, on the *concept* of natural law, Hobbes puts forward the orthodox view that precepts of reason can only be thought of as laws if they are considered to be products of the divine will (see ch. 15 of *Leviathan*, 1651).

As we have seen, the theo-centric framework of natural law theory preserved a deontological form for the precepts of natural law while allowing the substantive arguments (the need for certain rules given the known features of human nature, etc.) to take on a basically utilitarian character. What is often described as the ‘critique’ of natural law produced by David Hume in the 18th century is really best understood as a removal of the deontological framework, leaving only the utilitarian arguments in place. Hume removed God from the picture and offered a justification for rules of justice and property that appealed straightforwardly to arguments of ‘convenience’ or utility. Once this move was made, however, a dramatic sea-change was in process, for if the rules of justice and property are not prescribed by God, they are simply justified by utility. Of course, when Hume spoke of utility he did not have in mind a simple maximizing structure with a clearly defined maximand. But in the hands of Bentham, the notion of utility was developed in precisely that way.

Hume’s removal of God from the picture of natural law was undoubtedly a decisive move. Yet the underlying utilitarian cast of much natural law writing meant that there was a good deal of continuity between Hume’s predecessors and his immediate heirs. There had always been a tendency for the separate precepts of natural law to collapse into a general injunction to maximize utility, so that natural law ideas could continue to live a ghostly afterlife in the writings of utilitarians. Moreover, the reliance on speculative histories of, for example, the rise of private property, which had characterized the writings of Grotius and Pufendorf, was to take on a more descriptive and naturalistic character in the work of Adam Smith and the writers of the Scottish Enlightenment.

See Also

- ▶ Common Law
- ▶ Entitlements
- ▶ Invisible Hand
- ▶ Jurisprudence

Bibliography

Cairns, H. 1949. *Legal philosophy from Plato to Hegel*. Baltimore: Johns Hopkins Press.

Crowe, M.B. 1977. *The changing profile of the natural law*. The Hague: Nijhoff.

d'Entreves, A.P. 1970. *Natural law*, 2nd ed. London: Hutchinson.

Finnis, J. 1980. *Natural law and natural rights*. Oxford: Clarendon Press.

Forbes, D. 1975. *Hume's philosophical politics*. Cambridge: Cambridge University Press.

von Gierke, O. *Natural law and the theory of society*. Trans. E. Barker. Cambridge: Cambridge University Press, 1958.

Haakonssen, K. 1981. *The science of a legislator: The natural jurisprudence of David Hume and Adam Smith*. Cambridge: Cambridge University Press.

Jones, J.W. 1940. *Historical introduction to the theory of law*. Oxford: Clarendon Press.

O'Connor, D.J. 1967. *Aquinas and natural law*. London: Macmillan.

Simmonds, N.E. 1984. *The decline of juridical reason*. Manchester: Manchester University Press.

Strauss, L. 1953. *Natural right and history*. Chicago: Chicago University Press.

Tuck, R. 1979. *Natural rights theories*. Cambridge: Cambridge University Press.

Natural Monopoly

William W. Sharkey

An industry is a natural monopoly if total costs of production are lower when a single firm produces the entire industry output than when any collection of two or more firms divide the total among themselves. An industry can be a natural monopoly if production by a single firm is the outcome of unrestricted competition, or a natural monopoly may exist if competitive forces lead to a different industry structure. Generally a natural monopoly is characterized by subadditivity of a representative firm's cost function. A cost function c is subadditive at an output x if $c(x) \leq c(x^1) + c(x^2) + \cdots + c(x^k)$ for all non-negative $x^1, \ldots, x^k$ such that $\sum_{i=1}^{k} x^i = x$. If all prospective firms in the industry have the same cost function, or if one firm has a uniformly better technology, then subadditivity implies that industry costs are minimized if only one firm is active in the market. While subadditivity is a purely technical condition, it is also possible for natural monopoly to arise from purely economic forces if the imperfectly competitive outcome is inefficient. However, competition in a market with a small number of firms is inherently the domain of game theory and a unique equilibrium outcome is rarely found. Therefore it is generally acceptable to adopt the technical criterion of subadditivity as the defining characteristic of natural monopoly.

The concept of natural monopoly predates its definition in terms of subadditivity. Economists of the 19th and early 20th centuries spoke of natural monopoly conditions arising both from the superior efficiency of single-firm production and the undesirable consequences of excessive or 'destructive' competition. Often both forces were present at the same time, as for example was the case when competing telephone companies fought for subscribers during the early growth of the industry. Alfred Marshall was one of the first to identify formally the technology, in the form of the representative firm's cost function, as the fundamental determinant of industry structure. Industries with increasing average cost of production were generally competitive, while decreasing cost industries were imperfectly competitive or monopolistic. J.M. Clark (1923) contributed to the understanding of natural monopoly through his careful analysis of the economics of overhead costs, or in more recent terminology in the economics of 'non-convexities'. Clark recognized that in many manufacturing industries overhead costs are a significant fraction of total costs and that competition among firms in such an industry is far from perfect. In periods of slack demand there is a tendency for price to fall to marginal cost of production which may be less than average cost. At other times there may be quantity discounts or overt price discrimination among customers or across markets as firms strive to make up for earlier shortfalls. Thus the equilibrium in such a market is one in which variability and complexity replace the simplicity of a competitive equilibrium price. In extreme cases there may be no equilibrium unless firms in the market establish a standard of behaviour in which minimal cooperation replaced 'cut-throat competition'. Clark was also a pioneer in the empirical study of declining average cost industries. He correctly noted that most costs which appear fixed in the short run are variable in the long run. His estimates of long-run economies of scale in the railroad industry were significantly different from earlier results and remarkably similar to more recent estimates. Clark also recognized that product differentiation must be accounted for in testing for economies of scale.

By the middle of the 20th century it was recognized that railroads, telecommunications, and local public utilities all possessed to some degree the characteristics of a natural monopoly. To the extent that it was precisely defined, a natural monopoly was assumed to be an industry with significant long-run economies of scale. With increasing sophistication economists measured the actual scale economies in the above industries and others with similar characteristics. However, during this period it became increasingly apparent that the degree of scale economies was not the only relevant attribute of a natural monopoly. Most industries thought to be natural monopolies were regulated in the United States and publically owned elsewhere. In the regulatory climate in the USA there arose a set of new and persistent questions concerning the permissible grounds for competition at the boundaries of a natural monopoly. For example, regulated railroads faced increasing competition from regulated and unregulated trucking, and regulated telephone companies faced increasing competition from private networks and speciality carriers. Regulators were increasingly called upon to set standards for this form of competition, generally by means of complex methodologies for cost allocation. In effect, regulators were asked to determine in what way, if any, a regulated firm should be allowed to complete with an intermodal rival or an entrant. Scale economies provide little guidance in questions of this sort. At best scale economies describe the cost characteristics of a single product firm or a multiproduct firm which always increases outputs in the same proportion. The competitors of a regulated firm, however, are not required to produce outputs in the same proportion as the regulated firm. Instead they may choose to enter only the most lucrative markets. When such entry occurred, it was attacked as 'cream skimming' by the regulated firms and portrayed as innovative competition by the entrants. The concept of subadditivity arises naturally in such a context. If all firms share the same technology and the cost function is strictly subadditive then entry necessarily raises total industry cost. Therefore it is the degree of subadditivity rather than the degree of scale economies that is relevant in determining the minimum cost industry structure.

Although subadditivity is a simple concept to define mathematically, it is difficult to verify in practice. Unlike scale economies, which can be defined using local information about the cost function in the neighbourhood of an output x, subadditivity requires global information about the cost function for all values $x' \leq x$ If the representative cost function c is U-shaped with a unique minimum average cost at x' then there are scale economies for all outputs $x \leq x'$. Moreover it can be shown that there exists an output x'', with $x' < x'' < 2x'$ such that c is subadditive for all outputs $x \leq x''$ Thus in the single output case, scale economies are sufficient but not necessary for subadditivity. For a multiproduct cost function scale economies are neither necessary nor sufficient for subadditivity. For example the cost function $c(x_1, x_2) = x_1 + x_2 + (x_1 x_2)^{\frac{1}{3}}$ exhibits scale economies for all non-negative outputs but is nowhere subadditive. For this function there are 'diseconomies of scope', which means that the subadditivity condition fails to hold for orthogonal output vectors. For many cost functions it can be shown that economies of scale and scope are together sufficient for subadditivity. However, this is not true in general as can be seen from simple counterexamples (Sharkey 1982). Since direct tests for subadditivity are difficult to arrange it is of interest to determine sufficient conditions which may be easier to verify in certain contexts. The most useful sufficient condition is known as 'cost complementarity' which exists if the second partial derivatives of the cost function are everywhere nonpositive. Roughly speaking, cost complementarity occurs if there are 'increasing returns to scale and scope'.

Once it is known or thought likely that an industry is a natural monopoly, there remain difficult questions concerning the proper form of regulation. In the definition of natural monopoly a single firm must have a subadditive cost function using the best available technology at a given point in time. However, rival firms might at any future time discover new technologies that justify their entry into the industry. Entry may also be attractive to a firm with the same or inferior technology. Whenever the incumbent monopolist's prices are chosen in such a way that the revenue collected from an identifiable submarket exceeds the cost of serving that submarket, entry is potentially attractive in the submarket. For example, if a natural monopoly firm serves a geographically dispersed market, and either chooses to or is required to set prices on the basis of average cost per customer, then entry may be attractive to a relatively inefficient firm that specializes in serving the low-cost customers. More surprisingly, it is possible that entry in at least one submarket is possible for any conceivable set of prices of the incumbent. That is, assuming that the incumbent firm's cost function is subadditive, that all potential entrants have the same or higher costs at all outputs, and that the incumbent is allowed complete freedom to choose and maintain a set of prices, there may be no prices such that the incumbent can break even and simultaneously deter entry.

Prices which do deter entry by rivals with the same (or inferior) technology are known as 'sustainable prices'. Let the market be characterized by a demand function $D(p)$ and cost function $c(x)$. Then a price vector p is sustainable at x if $x = D(p)$, $\sum_{i=1}^{n} p_i x_i = c(x)$ and there do not exist alternative prices $p' \leq p$ and outputs $x' \leq D(p')$ such that $\sum_{i=1}^{n} p_i x_i' > c(x')$. A market in which there are no barriers to entry is known as a 'contestable' market. If there exist sustainable prices (and the natural monopolist is allowed to choose prices without regulator interference) then entry will not raise total industry cost. Actual entry will occur only if there is a technological innovation which reduces industry cost. In addition, if the market is contestable, the threat of potential entry will force the monopolist to choose from the set of sustainable prices (and therefore earn zero profits) provided that the monopolist behaves as assumed in the definition of sustainability. Sustainable prices can be proven to exist if various sets of assumptions are made about costs and demands. For example, if cross elasticities of demand are zero and all second partial derivatives of the cost function nonpositive, in which case there is 'cost complementarity', then sustainable prices are known to exist.

There are several serious objections to the behavioural assumptions implicit in the definition of sustainability. If sustainable prices exist and the market is contestable then a passive pricing strategy by the natural monopolist can guarantee nonnegative profits and minimum total industry costs. However, the monopolist might earn strictly positive profits by following a different strategy, such as committing to maintain outputs rather than prices if entry occurs. Furthermore, if sustainable prices fail to exist, the monopolist is even less likely to follow a passive pricing strategy, since it is possible that after entry occurs the monopolist's revenues are less than the cost of producing the reduced output. Even a regulated monopolist might be required to raise prices and thereby give an incentive for additional entry. A less constrained monopolist would be likely to pursue either a cooperative strategy to accommodate some entry, or a more threatening strategy to deter it.

In addition to the formulation of a definition of natural monopoly and the investigation of entry behaviour in natural monopoly markets, a number of subsidiary themes have been pursued in the natural monopoly literature. For example, a definition of 'subsidy free prices' has been found which takes account of the ability of subsets or coalitions of a regulated firm's customers to obtain service on their own (Faulhaber 1975). The investigation of optimal pricing subject to a budgetary constraint has also received considerable attention in papers by Ramsey (1927), Boiteux (1956), and Baumol and Bradford (1970). A paper by Baumol et al. (1977) demonstrated conditions under which the 'Ramsey optimal' prices are also sustainable. Numerous papers on the cost allocation problem have also appeared, including both axiomatic methods and more explicit game theoretic solution concepts.

See Also

- ▶ Contestable Markets
- ▶ Monopoly
- ▶ Ramsey Pricing
- ▶ Subaddivity

Bibliography

Baumol, W.J., and D.F. Bradford. 1970. Optimal departures from marginal cost pricing. *American Economic Review* 60(3): 265–283.

Baumol, W.J., E.E. Bailey, and R.D. Willing. 1977. Weak invisible hand theorems on the sustainability of prices in a multiproduct natural monopoly. *American Economic Review* 67(3): 350–365.

Boiteux, M. 1956. Sur la gestion des monopoles publics astreints à l'équilibre budgétaire. *Econometrica* 24(1): 22–40.

Clark, J.M. 1923. *Studies in the economics of overhead costs*. Chicago: University of Chicago Press.

Faulhaber, G.R. 1975. Cross-subsidization: Pricing in public enterprise. *American Economic Review* 65: 966–977.

Ramsey, F. 1927. A contribution to the theory of taxation. *Economic Journal* 37: 47–61.

Sharkey, W.W. 1982. *The theory of natural monopoly*. Cambridge: Cambridge University Press.

Natural Price

G. Vaggi

Keywords

Bon prix; Cantillon, R.; Competition; Cost of production; Effectual demand; Exchange value; Fundamental price; Hutcheson, F.; Locke, J.; Market price; Marshall, A.; Marx, K. H.; Natural price; Net product; Normal price; North, D.; Petty, W.; Physiocracy; Price theory; Pufendorf, S.; Quesnay, F.; Rent; Smith, A.; Sraffa, P.; Steuart, Sir J.; Turgot, A. R. J.; Uniform rate of profit; Value theory

JEL Classifications

B1

In the *Wealth of Nations* Smith says that

> when the price of any commodity is neither more nor less than what is sufficient to pay the rent of the land, the wages of labour, and the profits of the stock employed in the raising, preparing and bringing it to market, according to their natural rates, the commodity is then sold for what may be called its natural price. (Smith 1776, p. 72)

In the same chapter he explains that in economic theory this particular price level is important because it is a sort of benchmark for the actual price of the commodity, its market price (p. 73). The market price is different from the natural price but tends to move towards it all the time because of competition between producers. 'The natural price, therefore, is, as it were, the central price, to which the prices of all commodities are continuously gravitating' (p. 73). Smith's concept of natural price and his description of the competitive mechanism which guarantees that the market prices tend to move towards it became an important element in classical political economy. Smith's analysis was entirely subscribed to by Ricardo (Ricardo 1821, pp. 88–91), and was a central point in the classical theory of value and in the price theories of some neoclassical economists.

Smith's notion of natural price is part of a more general analysis of the normal and regular causes which determine the value of commodities. Smith's theory can be divided into three main aspects. First of all, there is the definition of natural price, which is made up of three component parts, wages, profits, and rent. In Chapter 6 of the *Wealth of Nations,* Smith explains that the price of all commodities resolves itself into wages, profits and rent, as soon as we abandon the 'early and rude state of society which precedes both the accumulation of stock and the appropriation of land' (Smith 1776, p. 65). The price must also repay the raw materials and the capital equipment consumed in production, but the prices of these commodities are also made up of the wages, profits and rent required in their own production (p. 68). Thus ultimately the price of each product is entirely made up of those three parts, which include the incomes of workers, landlords and capitalists who take part in the final production of the good and also the incomes of all those who have indirectly contributed to produce it in previous years. The techniques of production of a commodity have an important influence on its natural price, because they determine the relative shares of profits, rent, and wages. But the natural price also depends on the distribution of income, that is to say, on the level of the natural rates at which wages, rent and profits must be paid.

According to Smith, each rate is determined on a different market and this depends on several circumstances. Therefore the natural price of each commodity is determined by the methods of production and by the exogenously given values of the rates which remunerate the three classes which take part in production. It is worth noticing that for Smith, society is made up of different classes, labourers, landlords and capitalist entrepreneurs, whose economic functions are clearly separated. When all the commodities that make up the output of society are assessed according to their natural prices, the part of this value given by wages is the capital stock of society (p. 110), while rent and profits make up the net product, or surplus.

The second feature of Smith's price theory is the description of the reasons why the natural price is the price level which prevails in the long run, and around which market prices gravitate. This price mechanism is an important element in the notion of natural price because it guarantees that the permanent causes of value are those which influence the natural price, while market price deviations are due to temporary circumstances. The market price fluctuates and may differ from the natural price, but there are forces which compel it towards the natural price.

The factors affecting natural prices must be regarded as the permanent and fundamental forces that determine the value of produced commodities, quite independently from the day-to-day changes in their market prices. This second part of Smith's analysis of natural prices contains several concepts. First, there is the notion of effectual demand which is used to explain the differences between natural and market prices. Effectual demand is the 'demand of those who are willing to pay the natural price of a commodity' (p. 73). Of course a change in this price affects the effectual demand. The quantity produced and brought to the market may be lower (or higher) than the effectual demand, in which case the market price of the commodity will be higher (or lower) than the natural one. This mechanism explains why there are differences between natural and market prices.

The second step in Smith's analysis of the gravitation of market prices around natural prices consists in the competitive mechanism itself. Here, too, several logical stages may be distinguished. (a) For Smith the fact that the market price is higher than the natural one implies that at least one of the three parts which make up the price of a product is higher than it would have been if its contribution to production was remunerated according to its natural rate; it seems reasonable to assume that profits are the share which takes advantage of the favourable market conditions (but the process works in the same way if wages and rent are higher than their natural rates). (b) Entrepreneurs are aware of the existence of these different rates of profit in the different sectors of the economy. (c) There are no barriers to the free circulation of capital, thus entrepreneurs move towards the most remunerative sectors; this is the crucial aspect of Smith's analysis of competition (see Sylos-Labini 1976). (d) These capital movements lead to an increase in the output of the products which yield the highest rates of profit. (e) Since the quantity produced and brought to the market of these products increases while the effectual demand in unchanged, the market price falls. This does not mean that there is a downward-sloping demand schedule. In Smith's price theory there is no continuous differentiable inverse relationship between quantities and prices, as is found in neoclassical economics (Garegnani 1983).

Free competition tends to bring about a uniform rate of profit throughout the economy. Hence the concept of natural price is related to the existence of a single rate of profit on the capital invested in all sectors, and is regarded by Smith as 'a centre of repose and continuance' for the actual market price (Smith 1776, p. 75).

The view that it is possible and useful to separate the day-to-day fluctuations in market prices from the stable and permanent causes of the value of commodities can be traced back to the 16th century. It was part of Scholastic tradition to believe that there was a logical distinction between the actual price of a product and its *true* value. The former price can vary quite a lot according to the state of trade, while the value is always the same. Von Pufendorf believed that the value, or just price, of a commodity depended mostly on the difficulty of acquiring and producing it (Pufendorf 1688, pp. 684–9). Theoreticians of the just price regarded it as the level to which actual prices ought to conform. They gave no indication of any spontaneous mechanism which should guarantee that market values would adapt to these just levels.

As a student, Adam Smith read the works of von Pufendorf, and his teacher, Francis Hutcheson, wrote a book entitled *A System of Moral Philosophy* in which the distinction between value and price was restated along very similar lines (Hutcheson 1754–5, pp. 53–5). At the end of the 17th century, Dudley North and John Locke maintained that regulations and government interventions could not affect the price of commodities, which depended on market conditions (North 1691, Preface; Locke 1691, pp. 4, 11, 13).

Some years before the publication of the works of Locke and North, Sir William Petty regarded the cost of production of commodities as the main cause determining their true value. Ultimately all commodities are produced by two common denominators, land and labour, and their exchange values are in proportion to the quantities of these non-produced goods which have been employed in their production (Petty 1662, p. 44). The value of goods is regulated by the physical cost of production, which is regarded as the true measure of the difficulty of acquiring them. For Petty, the natural price depends upon the amount of labour required to produce a commodity with the best available technique (pp. 50–1).

Richard Cantillon developed Petty's analysis of land and labour as the original components of the value of each commodity. He transformed the amount of labour employed in production into an equivalent quantity of land. Thus, the value of each commodity is given by the quantity of land which has been directly and indirectly used in its production (Cantillon 1755, p. 29). This is the intrinsic value of the products, and their market price fluctuates around it (pp. 28–30). Moreover, Cantillon presented the well-known theory of the 'three rents'; the farmer receives two thirds of the products of land, one third is required to pay workers' wages and other expenses, the second third is the profit from his enterprise; the final third accrues to landlords as rent (p. 43).

Quesnay and the Physiocrats also distinguish the permanent value of commodities from their market price. For Quesnay, the fundamental price is the lowest level of the selling price for the producer. This value is the minimum level of the market price: it is the sum of all the expenses incurred by the cultivator in the production of a commodity, and there is a loss when the market price is lower than this value (Quesnay 1757, p. 555). The fundamental value of commodities is stable and varies quite slowly, on the other hand market prices change rapidly. Quesnay concentrated his attention on the fundamental price of primary commodities, which included the technical costs of production plus the annual rent paid to the landlords (1757, p. 555; Quesnay 1756, p. 443). Quesnay believed that two elements contribute to determining the fundamental value of agricultural products: farming techniques, which determine the physical cost of production, and the rule which fixes the distribution of income, at least in the form of rent. The inclusion of an element, rent (which is part of the country's surplus), in the fundamental value of a commodity is an important step towards Smith's concept of natural price. Now the permanent value of commodities is not only the result of technical conditions but also of the social rules and customs which determine the distribution of the net product.

Quesnay used the term ‘natural price’ to indicate the state of prices when free and unobstructed competition in all the markets regulates the exchanges between buyers and sellers (Quesnay 1766, pp. 829–30). In this case the actual exchange value of the products of land is a *bon prix,* it exceeds the fundamental price and leaves the farmer with a profit (Quesnay 1757, p. 529). Quesnay provided a good explanation of the reasons why the market price cannot be lower than the fundamental one, but there is no indication of the existence of market forces which lead the actual price towards the *bon prix.* In Quesnay’s value theory the notion of fundamental price is only a sort of threshold which fixes the lowest market price, but profits are still not part of the fundamental price.

In 1767 Sir James Steuart published *An Inquiry into the Principles of Political Oeconomy* in which he made at least two important contributions to the classical theory of value. The first was the notion of the real, or intrinsic, value of the goods. He says that two things make up the price of a product, ‘the real value of a commodity and the profit upon alienation’ (Steuart 1767, p. 159). The real value is the cost of production, which depends upon the average techniques which have been adopted and which establishes the amount of time needed to produce a commodity. The ‘profit upon alienation’ is the positive difference between the actual price and the real value (1767, p. 159). Thus profits are not part of the value of commodities, but according to Steuart ‘such profits subsisting for a long time, they insensibly become *consolidated*, or as it were, transformed into the intrinsic value of the goods’ (1767, p. 193, Steuart’s italics). Thus, in the normal condition of the market, the value of commodities must also include entrepreneurs’ profits, which are a permanent feature of the exchange value of goods. Steuart’s second contribution to price theory is the concept of effectual demand; this notion indicates the demand of consumers who can actually pay for a product and is clearly distinguished from wants and desires (1767, pp. 151–3). Steuart’s analysis does not provide a theory of profit capable of explaining the level which becomes consolidated in the intrinsic value of commodities. The normal value is not yet defined in a way which explains the existence of a regular element of profit in the exchange value of commodities.

In the *Obsérvations sur le mémoire de Saint Péravy*, Turgot distinguished the fundamental and market price of commodities. The first concept is defined as the cost of production, which includes wages, raw materials and interests on the capital advanced. The fundamental value is fairly stable, while the exchange value is ruled by supply and demand and ‘it has a tendency to approach it (the fundamental price) continually, and can never move away from it permanently’ (Turgot 1767, p. 120, n. 16). There is an important difference between Quesnay’s and Turgot’s use of the term ‘fundamental price’. Turgot’s notion does not simply indicate the lowest level of the market price, but is the value to which this price must tend. Turgot included a regular profit among the necessary expenses of production (Meek 1973, p. 17). Turgot’s interest on the capital advanced is not only a depreciation allowance but includes profit for the entrepreneur. In *Réflexions sur la formation et la distribution des richesses* (1766) Turgot clearly says that the return to the capitalist entrepreneur must be divided into three main categories: ‘depreciation of the capital’, ‘wages of superintendence and direction as well as the risk premium’ and ‘pure return on his capital which he could have earned if he had not employed it in industry’ (Groenewegen 1971, p. 333; see Turgot 1766, pp. 152, 154). Now profits are an essential part of the permanent value of commodities, but above all Turgot’s notion of profit is different from those of Steuart and Quesnay. Profit is defined as a rate on the capital invested. This definition of profits is quite different from that of profit upon alienation, according to which profits are influenced by market conditions where the products are sold. For Turgot, on the contrary, the rate of profit depends mainly on competition between capitalist producers who act with a view to obtaining the highest possible rate of profit. This mechanism explains the existence of a continuous tendency towards the equalization of rates of return in all of the capital.

In the *Lectures on Jurisprudence* which Adam Smith gave at Glasgow in the academic year 1762–3, we already find the distinction between natural and market price, together with the description of the mechanism by which the latter price gravitates around the natural value (Smith 1762–3, pp. 353ff.). Smith's analysis of competition among producers explains that natural prices are bound with the existence of a uniform rate of profit in all the sectors of the economy. The existence of this uniform rate has been traditionally adopted to describe the prices which prevail in the long run, when it is possible to abstract from all the accidental causes which influence market prices. In Smith's economics, technology and income distribution are the permanent forces which determine the value of natural prices.

In classical economics, the notion of natural price is necessary to build up an abstract analysis of the main features of the economy. This notion helps to single out the main characteristics of the capitalistic process of development and their relationships to changes in the distribution of income. Thus the concept of natural price is part of the study of the long-term changes in economic systems, which derive from capital accumulation. Natural price is an essential element of the classical method of analysis, which investigates the features of the long-term positions of the economy, when demand does not affect prices and income distribution (Garegnani 1976, section 1).

In Chapter 4 of *On the Principles of Political Economy and Taxation*, Ricardo subscribes to Smith's theory of natural prices (1821, pp. 88–92). He was interested in the analysis of the permanent changes in income distribution, and was not interested in the temporary deviation of market prices from their natural value.

However, there is a major difference between Smith's and Ricardo's theories of profit. Smith says that profits and wages are determined on separate markets and that the natural price is the sum of these shares plus rent, while Ricardo says that the rate of profit and the real wage are inversely related.

Marx's notion of prices of production shares many of the features of Smith's natural price; both concepts are associated with the existence of a uniform rate of profit in all sectors of the economy (see Marx 1894, pp. 153–8). Moreover, Marx accepted Ricardo's analysis of the reasons why market prices fluctuate around natural ones (1894, p. 179). Like Ricardo, he believed that real wages and the rate of profit vary in opposite directions. In his 1951 Introduction to *The Works and Correspondence of David Ricardo,* Sraffa clearly singled out the implications of Ricardo's theory of profit determining commodities natural value. Sraffa explicitly mentioned the concepts of natural price and prices of production in presenting his theory of price determination and retained the notion of a uniform rate of profit throughout the economy (Sraffa 1960, pp. 9, 6).

In the *Principles of Economics* (1920), Alfred Marshall referred to Smith's natural price, for which he substituted the notion of normal price (Marshall 1920, p. 289). In his discussion of the causes which influence the value of commodities he said that in general, market values are deeply affected by demand, while normal prices depend on the cost of production of commodities. The former price prevails in the short run, but 'the longer the period, the more important becomes the influence of cost of production on value' (1920, p. 291). Normal prices are determined by the persistent causes of value, and are not influenced by fitful and irregular events (1920, pp. 304–5). It should be pointed out that Marshall's notion of cost of production is not the same as the notion put forward by Ricardo and Marx. Moreover, he was sceptical about the existence of a tendency towards the equalization of the rates of profit in all economic activities (1920, pp. 506–7, 512). Nevertheless inside each branch of trade there can be a fair rate of profit which must be reckoned as a component element of the normal price (1920, pp. 513–14).

See Also

- ▶ British Classical Economics
- ▶ Market Price

Bibliography

Cantillon, R. 1755. *Essai sur la nature du commerce en général*. Edited by H. Higgs. London: Cass, 1959.

Garegnani, P. 1976. On a change in the notion of equilibrium in recent work on value and distribution. In *Essays in modern capital theory*, ed. M. Brown, K. Sato, and P. Zarembka. Amsterdam: North-Holland.

Garegnani, P. 1983. The classical theory of wages and the role of demand schedules in the determination of relative prices. *American Economic Review* 73: 309–313.

Groenewegen, P.D. 1971. A reinterpretation of Turgot's theory of capital and interest. *Economic Journal* 81: 327–340.

Hutcheson, F. 1754–5. *A system of moral philosophy*. Glasgow: Robert and Andrew Foulis.

Locke, J. 1691. Some considerations of the consequences of the lowering of interest and raising the value of money. In *The works of John Locke,* vol. 5. London, 1823.

Marshall, A. 1920. *Principles of economics,* 8th ed. Reprinted, London: Macmillan, 1972.

Marx, K. 1894. *Capital*. Vol. 3. London: Lawrence & Wishart, 1977.

Meek, R.L. 1962. *The economics of physiocracy*. London: George Allen & Unwin.

Meek, R.L., ed. 1973. *Turgot on progress, sociology and economics*. Cambridge: Cambridge University Press.

North, D. 1691. Discourses upon trade. In *Early English tracts on commerce.* London: The Political Economy Club. Reprinted, Edited by J.R. McCulloch. Cambridge: Cambridge University Press, 1954.

Petty, W. 1662. A treatise of taxes and contributions. In *The economic writings of Sir William Petty,* ed. C.H. Hull. Cambridge: Cambridge University Press, 1899.

Quesnay, F. 1756. Fermiers. In *François Quesnay et la Physiocratie.* Paris: INED, 1958.

Quesnay, F. 1757. Hommes. In *François Quesnay et la Physiocratie.* Paris: INED, 1958.

Quesnay, F. 1766. Du commerce. In *François Quesnay et la Physiocratie.* Paris: INED, 1958.

Ricardo, D. 1821. On the principles of political economy and taxation. In *The works and correspondence of David Ricardo*, ed. P. Sraffa with the collaboration of M.H. Dobb. Cambridge: Cambridge University Press, 1951.

Smith, A. 1762–3. *Lectures on jurisprudence.* Edited by R.L. Meek, D.D. Raphael, and P.G. Stein. Oxford: Oxford University Press, 1978.

Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*. Oxford: Oxford University Press, 1976.

Sraffa, P. 1951. *Introduction to the works and correspondence of David Ricardo*. Vol. 1. Cambridge: Cambridge University Press.

Sraffa, P. 1960. *Production of commodities by means of commodities*. Cambridge: Cambridge University Press.

Steuart, J. 1767. *An inquiry into the principles of political oeconomy.* Edited by A. Skinner. Edinburgh/London: Oliver & Boyd, 1966.

Sylos-Labini, P. 1976. Competition: The product market. In *The market and the state*, ed. T. Wilson and A. Skinner. Oxford: Clarendon Press.

Turgot, A.R.J. 1766. *Réflexions sur la formation et la distribution des richesses*. Edited by Meek. 1973.

Turgot, A.R.J. 1767. Observations sur le mémoire de Saint-Péravy. In *The economics of A.R.J. Turgot,* ed. P.-D. Groenewegen. The Hague: Martinus Nijhoff, 1977.

von Pufendorf, S. 1688. *De jure naturae et gentium – libri octo*. Oxford: Clarendon Press, 1934.

Natural Rate and Market Rate of Interest

Axel Leijonhufvud

Abstract

The terms 'natural rate' and 'market rate' of interest were introduced by Wicksell (1898, 1906) to denote an equilibrium value and the actual value of the real rate of interest. Wicksell applied these concepts to explain the inter-equilibrium movement of money and prices using the hypothesis of maladjustments in the interest rate. Wicksell's work made the nexus between money creation, intertemporal resource allocation disequilibrium and movements in money income the dominant theme in macroeconomics for three decades. However, Keynes's conclusions over the saving–investment problem in the *General Theory* led to the abandonment of the concept of 'natural' rate of interest.

Keywords

Credit; Cumulative processes; Fiat money; Forced saving; Hayek, F. A. von; Inflation; Inflation targeting; Inside money; Intertemporal resource allocation; IS–LM model; Keynes, J. M.; Liquidity preference; Loanable funds; Monetarism; Money supply; National income; Natural rate and market rate of interest; Neutrality of money; New classical macroeconomics; Overinvestment; Quantity theory of money; Saving–investment coordination; Stockholm School; Unemployment equilibrium; Wage rigidity; Wicksell, J. G. K.

JEL Classifications

B2

The main analytical elements of Knut Wicksell's *Interest and Prices* can be found in the works of earlier writers. Wicksell was familiar with Ricardo's distinction between the direct and indirect transmission of monetary impulses. Although unknown to Wicksell in 1898, Henry Thornton had provided a clear account of the cumulative process in 1802, as had Thomas Joplin of the saving–investment analysis somewhat later (cf. Humphrey 1986).

Yet Wicksell did not just coin the terms 'natural rate' and 'market rate of interest'. His development (1898; 1906) of these ideas made the nexus between money creation, intertemporal resource allocation disequilibrium and movements in money income the dominant theme in macroeconomics for three decades until it was submerged in Keynesian economics. His starting point was the quantity theory, understood as the proposition that in the long run the price level will tend to be proportional to the money stock. His objective was to explain how both money and prices come to move from one equilibrium level to another. This inter-equilibrium movement became his famous 'cumulative process'. The maladjustment of the interest rate was the key hypothesis in Wicksell's explanation.

The 'market rate' denotes the actual value of the real rate of interest while the 'natural rate' refers to an equilibrium value of the same variable. The latter term by itself divulges Wicksell's engagement in the ancient quest for a 'neutral' monetary system, that is, a system neutral in the original sense that all relative prices develop as they would in a hypothetical world without paper money. Wicksell asserted three equilibrium conditions that the interest rate should satisfy; the first of these was that the market rate should equal the rate that would prevail if capital goods were lent and borrowed in kind *(in natura)*. This criterion was later shown by Myrdal, Sraffa and others not to have an unambiguous meaning outside the single input–single output world of Wicksell's example. The further development of Wicksellian theory, therefore, centred around the two remaining criteria: saving–investment coordination and price level stability.

The interest rate has two jobs to do. It should coordinate household saving decisions with entrepreneurial investment decisions and it should balance the supply and demand for credit. If the supply of credit were always to equal saving and the demand for credit investment, the two conditions could always be met simultaneously. But there is no such necessary relationship between saving and investment on the one hand and credit supply and demand on the other. In Wicksell's system the banks make the market for credit; they may, for instance, go beyond the mere intermediation of saving and finance additional investment by creating money; the injection of money drives a wedge between saving and investment; this could only be so if the banks set the market rate below the 'natural' value required for the intertemporal coordination of real activities. The resulting inflation and endogenous growth of the money supply would continue as long as the banking system maintained the market rate below the natural rate. Wicksell analysed the case of a 'pure credit' economy in which the cumulative process could go on indefinitely, but he also pointed out that, in a gold standard world, the banks would eventually be checked by the need to maintain precautionary balances of reserve media in some proportion to their demand obligations.

Wicksell used the model to explain long-term trends in the price level and was critical of those who, like Gustav Cassel, used it to explain the business cycle. Nonetheless, subsequent developments of his ideas went altogether in the direction of shorter-run macroeconomic theory. In Sweden, Erik Lindahl (1939) and Gunnar Myrdal (1939) refined the conceptual apparatus, in particular by introducing the distinction between *ex ante* plans and *ex post* realizations and thereby clarifying the relationship between Wicksellian theory and national income analysis. The attempts by the Stockholm School to improve on Wicksell's treatment of expectations were less successful, however, producing a brand of generalized process-analysis in which almost 'everything could happen'.

In Austria, Ludwig von Mises and Friedrich von Hayek focused on the allocational consequences of the Wicksellian inflation story. The Austrian overinvestment theory of the business cycle became known to English-speaking economists primarily through Hayek's *Prices and Production* (1931). In expanding the money supply, the banks hold market rate below natural rate. At this disequilibrium interest rate, the business sector will plan to accumulate capital at a rate higher than the planned saving of the household sector. If the banks lend only to business, the entrepreneurs are able to realize their investment plans whereas households will be unable to realize their consumption plans ('forced saving'). The too rapid accumulation of capital (which also has the wrong temporal structure) cannot be sustained indefinitely. The eventual collapse of the boom may then be exacerbated by a credit crisis as some entrepreneurs are unable to repay their bank loans.

The Austrian 'monetary' theory of the cycle has been overshadowed first by Keynesian 'real' macrotheory and later by monetarist theory. One problem with it is the firm association of inflation with overinvestment. The US stagflation in the 1970s, for example, will not fit. The reasons lie largely in the changes that the monetary system has undergone. Most obviously, commercial banks now lend to all sectors and not only to business. More importantly, however, inflation in a pure fiat regime does not tend to distort intertemporal values in any particular direction (although it may destroy the system's capacity for coordinating activities over time): it simply blows up the nominal scale of real magnitudes at a more or less steady or predictable rate. In contrast, the Austrian situation that preoccupied Mises and Hayek in the late 1920s was one of credit expansion by a small open economy on the gold standard. Given the inelastic nominal expectations appropriate to this regime, the growth of inside money would be associated with the distortion of relative prices and misallocation effects predicted by the Austrian theory.

In England, Dennis Robertson and J. Maynard Keynes both worked along Wicksellian lines in the 1920s. The novel and complicated terminology of Robertson's *Banking Policy and the Price Level* (1926) may have made the work less influential than it deserved. Keynes's *Treatise on Money* (1930), although also remembered as a flawed work, nonetheless remains important as a link in the development of macroeconomics from Wicksell to the *General Theory.*

In the *Treatise,* Keynes, like Wicksell, assumes that the process starts with a real impulse, that is, a change in investment expectations. Unlike Wicksell, he focuses on deflation rather than inflation. For Keynes with his City experience, *the* interest rate was determined on the Exchange rather than set by the banks. Consequently, a deflationary situation with the market rate exceeding the natural rate can only arise when bearish speculation keeps the rate from declining. When saving exceeds investment, therefore, money leaks out of the circular spending flow into the idle balances of bear-speculators. Thus the analysis stresses declining velocity rather than endogenously declining money stock. At this stage of the development of Keynesian economics, the banks are already edging out of the theoretical field of vision and the original connection of natural rate theorizing with criteria for neutral money is by and large severed.

The model of the *Treatise still* assumes that, when market rate exceeds the natural rate, the resulting excess supply of present goods will cause falling spot prices but not unemployment of present resources. Although the focus is on a disequilibrium process, at a deeper level the theory is still comfortingly classical. As long as the economy remains at full employment, the bear-speculators who are maintaining the disequilibrium are forced, period after period, to sell income-earning securities and accumulate cash at a rate corresponding to the difference between household saving and business sector investment. Automatic market forces, therefore, are seen to put those responsible for the undervaluation of physical capital under inexorably mounting pressure to allow correction of the market rate. And the longer those agents acting on incorrect expectations persist in obstructing the intertemporal coordination of activities, the larger the losses that they will eventually suffer.

In the *General Theory,* Keynes starts the story in the same way: investment expectations take a turn for the worse – 'the marginal efficiency of capital declines'; the speculative demand for money prevents the interest rate from falling sufficiently to equate *ex ante* saving with investment. But at this point the *General Theory* takes a different tack: the excess supply of present resources, which is the immediate result of the failure of intertemporal price adjustments to bring intertemporal coordination, is eliminated through falling output and employment. Real income falls until saving has been reduced to the new lower investment level.

This change in the lag-structure of Keynes's theory ('quantities reacting before prices') is not necessarily revolutionary by itself. But Keynes combines it with the assumption that the subsequent price adjustments will be governed, in Clower's terminology, not by 'notional' but by 'effective' excess demands. For the economy to reach a new general equilibrium, on a lower growth path, interest rates should fall but money wages stay what they are. Following the real income response, however, saving no longer exceeds investment so there is no accumulating pressure on the interest rate from this quarter; at the same time, unemployment does put effective pressure on wage rates. Interest rates, which should fall, do not; wages, which should not, do. From this point, Keynes went on to argue that nominal wage reductions would not eliminate unemployment unless, in the process, they happened to produce a correction of relative prices (an eventuality that he considered unlikely). This argument was the basis for his 'revolutionary' claim that a failure of saving–investment coordination could end with the economy in 'unemployment equilibrium'.

Prior to the *General Theory*, writers in the Wicksellian tradition had generally treated 'saving exceeds investment' and 'market rate exceeds natural rate' as interchangeable characterizations of the same intertemporal disequilibrium. The basic proposition could be couched equally well in terms of quantities as in terms of prices. In the *General Theory,* Keynes moved away from this language. Constructing a model with output and employment variable in the short run was a novel task and Keynes, as the pioneer, was unsure in his handling of expected, intended and realized magnitudes. Thus his preoccupation with the 'necessary equality' of saving and investment *(ex post)* was to produce endless confusion over interest theory. If saving and investment are always equal, the interest rate cannot be governed by the difference between them; nor can the interest rate mechanism possibly coordinate saving and investment decisions. To Keynes, two things seemed to follow. One was the substitution of the liquidity preference theory of the interest rate for the loanable funds theory; the other was the abandonment of the concept of a 'natural' rate of interest (Leijonhufvud 1981, pp. 169 ff.)

These were not innocent terminological adjustments. The brand of Keynesian economics that developed on the basis of the IS–LM model had only a shaky grasp at the best of times of the intertemporal coordination problem originally at the heart of Keynes's theory. The Keynesian position shifted already at an early stage back to the pre-Keynesian hypothesis of money wage 'rigidity' as the cause of unemployment. This switched the focus of analytical attention away from the role of intertemporal relative prices (the market rate) in the coordination of saving and investment to the relationship between aggregate money expenditures and money wages. This brand of 'Keynesian' theory which excludes the saving–investment problem (that is, excludes the market-natural rate problem) could hardly be distinguished from Monetarism in any theoretically significant way.

Monetarism gained enormously in influence during the inflationary 1970s. But its period of dominance was brief. This was so in part because, in its New Classical form, it was both theoretically implausible and empirically weak. In part, however, it was swept aside by a wave of innovations in payments technology and in forms of short-term credit that undermined the stability of the relationship between the money stock and income which had been the very linchpin of monetarist doctrine.

Most recently, this has led to a return to a basically Wicksellian doctrine of what monetary policy should aim to accomplish and how it should be conducted. Leading central banks are now committed to targeting the inflation rate (rather than the price level) and use the interest rate as their primary instrument for pursuing that goal. This policy doctrine has been elaborated in the book by Woodford (2003) which borrows its title from Wicksell.

See Also

- ▶ Stockholm School
- ▶ Wicksell, Johan Gustav Knut (1851–1926)

Bibliography

Cassel, G. 1928. The rate of interest, the bank rate, and the stabilization of prices. *Quarterly Journal of Economics* 42: 511–529.

Hayek, F.A. 1931. *Prices and production*. London: Routledge & Kegan Paul.

Humphrey, T.M. 1986. Cumulative process models from Thornton to Wicksell. *Federal Reserve Bank of Richmond Economic Review* 18–25.

Keynes, J.M. 1930. *A treatise on money*, 2 vols. London: Macmillan.

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

Leijonhufvud, A. 1981. The Wicksell connection. In *Information and coordination*, ed. A. Leijonhufvud. New York: Oxford University Press.

Lindahl, E. 1939. *Studies in the theory of money and capital*. New York: Holt, Rinehart & Winston.

Myrdal, G. 1939. *Monetary equilibrium*. Edinburgh: William Hodge.

Palander, T. 1941. On the concepts and methods of the Stockholm school. In *International economic papers*, vol. 3. London: Macmillan, 1953.

Robertson, D.H. 1926. *Banking policy and the price level*. New York: Augustus M. Kelley, 1949.

Taylor, J.B. (ed.). 1999. *Monetary policy rules*. Chicago: University of Chicago Press.

Wicksell, K. 1898. *Interest and prices*. New York: Augustus M. Kelley, 1962.

Wicksell, K. 1906. *Lectures on political economy*, vol. 2. London: Routledge & Kegan Paul, 1934.

Woodford, M. 2003. *Interest and prices: Foundations of a theory of monetary policy*. Princeton: Princeton University Press.

Natural Rate of Unemployment

Michael J. Pries

Abstract

Milton Friedman defined the natural rate of unemployment as the level of unemployment that resulted from real economic forces, the long-run level of which could not be altered by monetary policy. Macroeconomic policymakers continue to view the natural rate as a key benchmark due to the belief that monetary policy can counter short-run deviations of the unemployment rate from the natural rate. It is important, however, that policymakers focus as much attention on understanding the real determinants of the natural rate, and the policies that can affect it, as they do on trying to identify and counteract deviations from it.

Keywords

American Economics Association; Demography; Friedman, M.; Inflationary expectations; Labour supply; Natural rate of unemployment; Phillips curve; Rational expectations; Real business cycles; Search models of unemployment; Taylor rule; Unemployment insurance; Unemployment–inflation tradeoff; Wage rigidity

JEL Classifications

D4; D10

In his 1968 presidential address to the American Economics Association, Milton Friedman famously defined the natural rate of unemployment as

> . . . the level that would be ground out by the Walrasian system of general equilibrium equations, provided there is imbedded in them the actual structural characteristics of the labor and commodity markets, including market imperfections, stochastic variability in demands and supplies, the cost of gathering information about job vacancies and labor availabilities, the costs of mobility, and so on. (1968, p. 8)

This definition is incomplete, however, because it conspicuously lacks any mention of inflation. A more complete definition emerges from the remainder of Friedman's presidential address, in which he extensively examined the relationship between the unemployment rate and inflation. He argued that, whereas the natural rate of unemployment is determined by the real factors described in the passage quoted above, deviations from the natural rate are monetary phenomena: 'I use the term "natural" for the same reason Wicksell did – to try to separate real forces from monetary forces' (Friedman 1968, p. 9).

The Unemployment–Inflation Trade-Off

Friedman's 'natural rate hypothesis' maintained that '. . . there is a 'natural rate of unemployment' which is consistent with the real forces and with accurate perceptions; unemployment can be kept below that level only by an accelerating inflation; or above it only by accelerating deflation' (Friedman 1976, p. 458). This view of the relationship between the unemployment rate and inflation grew out of the experiences of the previous decades. In 1958, Phillips had observed a negative empirical relationship between the unemployment rate and the growth rate of wages (Phillips 1958). Understanding that high wage growth would ultimately translate into inflation, policymakers believed that there was a stable trade-off between unemployment and inflation that they could exploit. In other words, monetary and fiscal policy could be used to drive down unemployment at the cost of a certain degree of inflation. Experience showed, however, that the relationship was not stable. As individuals started to anticipate the inflation that resulted from attempts to exploit the trade-off, stimulative policy ceased to lower unemployment. Consequently, the Phillips curve appeared to have shifted outward, with higher inflation accompanying higher unemployment.

Friedman provided an explanation for this apparent shift. Over the long run, there is an unemployment rate determined by real factors that cannot be affected by monetary policy: the natural rate. In the short run, unanticipated inflation can temporarily push the unemployment rate below its natural rate. If workers do not perceive the higher inflation, then they will respond to higher nominal wages by increasing labour supply; similarly, employers who do not immediately perceive the higher inflation will respond to a higher price for their product by demanding more labour. This temporarily lowers unemployment, but the unemployment rate returns to its natural level when workers and employers begin to perceive the inflation. As emphasized in the literature on rational expectations (for example, Lucas 1973) that followed Friedman, inflation has no impact on real variables like the unemployment rate once individuals have already built the level of inflation into their expectations. In other words, as expectations about inflation change, the Phillips curve shifts.

Although the absence of any long-run trade-off between inflation and unemployment has gained wide acceptance, the possibility of a short-run trade-off has kept the natural rate of unemployment at the centre of policymaking. In particular, policy rules such as the Taylor rule (see Taylor 1999) maintain that central banks can stabilize the inflation rate by assessing where the economy stands relative to economic benchmarks such as the natural rate of unemployment, 'potential output', or the 'natural rate of interest'. When unemployment is high relative to the natural rate, and when output is below potential output, the policy rules call for stimulative monetary policy.

However, several important questions arise when one contemplates the usefulness of the natural rate of unemployment as a policy benchmark. First, although the natural rate clearly cannot be observed directly, can it be estimated with enough accuracy to be useful for policy? Or do movements in the natural rate itself make it too difficult to distinguish the natural rate and deviations from the natural rate in a sufficiently timely manner to be useful for policymakers? Second, rather than focusing so much on deviations from the natural rate, should policymakers also focus on policies that would alter the natural rate, either at low frequencies or perhaps even at business cycle frequencies? What would those policies be?

Identifying the Natural Rate

Although the natural rate is often simplistically described as the long-run average unemployment rate, economists widely recognize that this rate varies over time. Friedman (1968, p. 9) was clear on this point:

> To avoid misunderstanding, let me emphasize that by using the term 'natural' rate of unemployment, I do not mean to suggest that it is immutable and unchangeable. On the contrary, many of the market characteristics that determine its level are man-made and policy-made. . .. Improvements in employment exchanges, in availability of information about job vacancies and labor supply, and so on, would tend to lower the natural rate of unemployment.

Friedman (1968, p. 10) further argued that the mutability of the natural rate of unemployment significantly reduces its policy usefulness:

> What if the monetary authority chose the 'natural' rate – either of interest or unemployment – as its target? One problem is that it cannot know what the 'natural' rate is. Unfortunately, we have as yet devised no method to estimate accurately and readily the natural rate of either interest or unemployment. And the 'natural' rate will itself change from time to time.

Since Friedman's work, however, economists have achieved additional understanding of some of the factors that contribute to low-frequency fluctuations in the natural rate of unemployment. It is now generally understood that demographic changes can have a significant impact on the natural rate of unemployment (see Shimer 1998). For instance, young workers experience substantially more job turnover than more experienced workers, with the spells between jobs often spent in unemployment. Accordingly, when younger workers make up a larger fraction of the workforce (as they did in the 1970s when the baby boom generation entered the workforce in significant numbers), unemployment will be higher on average. Nevertheless, it is not clear whether this greater understanding of the factors that affect the natural rate can be translated into an estimate of the natural rate that is accurate enough to be useful for policy. Often changes in the natural rate can only be detected with a significant lag, after which time a policy response may actually increase volatility by causing the economy to overshoot its target.

Further complicating the question of the natural rate's usefulness as a policy benchmark is the question of whether even higher-frequency (that is, business cycle) fluctuations in the unemployment rate could in fact represent movements in the natural rate. For example, modern search theory views unemployment fluctuations at business cycle frequencies as movements in the natural rate, in the sense that they result from real rather than monetary forces. Evidence from data on job flows shows that jobs are constantly being reallocated across firms, industries, geographical regions, and so on (see Davis et al. 1996). Moreover, periods of above-average unemployment rates tend to coincide with an increased level of this reallocative activity. In this sense, unemployment rate fluctuations at business cycle frequencies can be viewed as the outcome of real phenomena of the type described in Friedman's famous quote – that is, as cyclical movements in the natural rate.

This emphasis on the real determinants of movements in the unemployment rate is part of the broader view that a significant portion of economic fluctuations reflects real factors as opposed to monetary phenomena. The vast real business cycle literature has explored this proposition since the seminal paper by Kydland and Prescott (1982). Hall (2005b) argues that real fluctuations, and the difficulty of distinguishing them from monetary phenomena, render useless the various benchmark concepts such as the natural rate of unemployment, potential output, and the equilibrium real interest rate.

Optimality of the Natural Rate and Policies to Alter It

If real sources of unemployment fluctuations are in fact as important as monetary sources, then the proper response by monetary policymakers to the fluctuations is much less clear. However, even if unemployment fluctuations are primarily driven by real factors, it would be incorrect to conclude that either the level or fluctuations of the natural rate are optimal. Accordingly, there may be a role for policy to improve welfare by affecting the natural rate (either at low frequencies or perhaps even at high frequencies). This suggests that research on the optimality of the natural rate, and on policies that can affect it, is as important as research aimed at detecting and proposing policies to counteract deviations from it.

The idea that the natural rate can be either too high or too low has been a primary focus of modern search and matching models of the labour market. In those models, the process whereby workers and firms meet may be subject to various externalities. When a worker chooses to search for a job, it has a positive externality on the probability that employers will find a suitable worker and a negative externality on the probability that other workers will find a job. Employers' search decisions cause similar externalities.

Hosios (1990) analyses the conditions under which, in a broad class of search and matching models, the various externalities result in an unemployment rate that is either too high or too low. He finds that in general there is no economic force that draws the unemployment rate towards its optimal level. One suspects that the wage might play that role. When employers decide whether to open job vacancies (the number of which ultimately determines the unemployment rate), they anticipate the wages that they will have to pay and the profits that they will earn when they form an employment relationship. However, the level of those wages and the resulting profits are determined after the fact by bargaining between workers and firms who have been matched, and who are not contemplating the impact that their bargain has on firms posting new vacancies. If the wages that result from bargaining are too low (high), firms anticipate this and create many (few) vacancies, and the unemployment rate is inefficiently low (high).

As a complement to this more theoretical examination of the optimal level of the natural rate, there is a more applied literature that tries to understand cross-country differences (particularly between continental Europe and the United States) in the average unemployment rate and how those differences relate to various policies. For example, Hopenhayn and Rogerson (1992) examine the impact of firing costs on unemployment and on productivity. They find that, in addition to increasing average unemployment, firing costs reduce productivity by impeding the reallocation of workers towards more productive employers. Ljungqvist and Sargent (1998) argue that the interaction between generous unemployment insurance in many western European countries and an increased turbulence in labour markets can explain the secular rise in European unemployment rates relative to the US rate over the last several decades.

In addition to this work on the determinants of average unemployment rates in the long run, recent work has also focused on trying to better understand the sources of non-monetary movements in the unemployment rate over the business cycle, and whether they are efficient. What real factors contribute to spikes in unemployment, and why is the subsequent recovery so slow? Pries (2004) argues that the slow recovery occurs because workers who lose their job in the initial spike may pass through several short-lived jobs, and several intervening unemployment spells, before ultimately settling into more stable employment. In this environment, policies that try to accelerate a recovery may be counterproductive if they encourage worker–firm pairs to hang on to low-quality matches.

Shimer (2005), on the other hand, argues that the slow recovery of the unemployment rate during economic downturns results from a significant reduction in posted vacancies and, consequently, a decline in workers' job-finding rates. More research is needed to understand the causes of the decline in posted vacancies. The canonical Mortensen–Pissarides (1994) matching model, in which wages are flexibly renegotiated as part of a Nash bargaining solution, struggles to produce a sizeable decline in vacancies during recessions. In the model, wages fall considerably during economic downturns, and the lower wages mean that firms still find it quite profitable to post vacancies. This model's failure to deliver the observed cyclicality in vacancies leads Hall (2005a) to suggest that in fact wages are much less flexible than assumed in Mortensen–Pissarides (1994). If so, then should the fluctuations be seen as monetary in nature, and is stimulative monetary policy the correct policy response? Or are tax incentives for investment, which may spur the creation of new jobs, a better policy response? As with countercyclical monetary policy, tax incentives may take effect with a lag and exacerbate fluctuations.

Milton Friedman's assertion in 1968 that there is a natural rate of unemployment that is determined by real economic forces and is impervious to monetary policy has become relatively uncontroversial. Nevertheless, important unresolved questions about the natural rate remain. What is the optimal natural rate? To what extent do unemployment rate fluctuations reflect movements in the natural rate as opposed to deviations from it? What policies, if any, are appropriate for counteracting movements in the natural rate or deviations from it?

See Also

- ▶ Friedman, Milton (1912–2006)
- ▶ Phillips Curve
- ▶ Real Business Cycles
- ▶ Search Models of Unemployment
- ▶ Taylor Rules

Bibliography

Davis, S., J. Haltiwanger, and S. Schuh. 1996. *Job creation and destruction*. Cambridge, MA: MIT Press.

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

Friedman, M. 1976. Nobel lecture: Inflation and unemployment. *Journal of Political Economy* 85: 451–472.

Hall, R. 2005a. Employment fluctuations with equilibrium wage stickiness. *American Economic Review* 95: 50–65.

Hall, R. 2005b. Separating the business cycle from other economic fluctuations. In *The Greenspan era: Lessons for the future.* Proceedings of the Federal Reserve Bank of Kansas City Symposium, August.

Hopenhayn, H., and R. Rogerson. 1992. Job turnover and policy evaluation: A general equilibrium analysis. *Journal of Political Economy* 101: 915–938.

Hosios, A. 1990. On the efficiency of matching and related models of search and unemployment. *Review of Economic Studies* 57: 279–298.

Kydland, F., and E. Prescott. 1982. Time to build and aggregate fluctuations. *Econometrica* 50: 1345–1371.

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

Lucas, R. 1973. Some international evidence on output-inflation tradeoffs. *American Economic Review* 63: 326–334.

Mortensen, D., and C. Pissarides. 1994. Job creation and job destruction in the theory of unemployment. *Review of Economic Studies* 61: 397–415.

Phillips, A. 1958. The relationship between unemployment and the rate of change of money wage rates in the United Kingdom, 1861–1957. *Economica* 58: 283–299.

Pries, M. 2004. Persistence of employment fluctuations: A model of recurring job loss. *Review of Economic Studies* 71: 193–215.

Shimer, R. 1998. In *Why is the U.S. unemployment rate so much lower?* NBER macroeconomics annual, ed. B. Bernanke and J. Rotemberg, vol. 13. Cambridge, MA: MIT Press.

Shimer, R. 2005. The cyclical behavior of equilibrium unemployment and vacancies. *American Economic Review* 95: 25–49.

Taylor, J. 1999. *Monetary policy rules*, NBER conference report series. Chicago/London: University of Chicago Press.

Natural Resources

Anthony C. Fisher

The adequacy of the resource base to support sustained growth of an agricultural, and later an industrial, society might be said to be one of the founding concepts of economics. Malthus's great treatise (1798) is concerned with population growth outstripping the (agricultural) resource base. Ricardo (1817) introduced a different, and probably more useful notion of scarcity, of higher quality, lower cost resources such as agricultural land, but also extractive resources like minerals. Both were pessimistic about prospects for long-term growth in the face of finite supplies of (good) land and related resources. The Ricardian scarcity concept was later applied by Jevons (1865) in a study of the British economy's dependence on coal. As Jevons noted, it is not simply, or so much the physical limits that matter, as the increasing costs of mining and processing lower-grade materials. From both classical and neoclassical sources, then, comes the idea that limited supplies and rising production costs of natural resources will exert a drag on growth, perhaps even preclude achievement of a steady state at a tolerable level.

Here I shall trace the evolution of thinking on this issue, and describe some additional concerns raised by contemporary economists. Chief among these is the question of how natural resources are allocated efficiently over time. Clearly the two concerns are related; if we are in danger of running out, we want to do the best we can with what we have. But most contemporary work has focused on one or the other, as I shall here.

The Great Scarcity Debate

Are resources limits to growth? For most of this century, and until quite recently – say the early 1970s – the prevailing view seems to have been, no, they are not, despite the earlier theories and predictions. In perhaps the most influential work on the subject, Barnett and Morse (1963) constructed indexes of the real costs of extractive output, and showed that these had tended to fall over the industrial history of the US to 1957. Later work, notably by Johnson, Bell and Bennett (1980), has extended these results to about 1970. The explanation is usually (and in my view correctly) given as technical change. Although this was foreseen even by Malthus, the broad and *sustained* nature of change was not. Other (related) factors considered responsible for the decline in costs include the discovery of new deposits and the substitution of more abundant materials for less abundant, as for example of aluminum for copper.

Recently a revisionist school of thought has arisen to challenge the prevailing view. Stimulated no doubt by the 'energy crisis' associated with the oil price shock of 1973–4, and possibly also by the nearly simultaneous appearance of *The Limits to Growth* (Meadows et al. 1972) and similar studies purporting to show that the US and global economies were doomed to collapse as they bumped up against resource limits in the near future, some economists have begun to question the Barnett–Morse results and consider whether, even if valid, they are accurate guides to the future. Looking at (mineral) resource *prices*, which as we shall see embody a kind of scarcity rent in addition to the cost of extraction, Smith (1979) finds that the rate of decline is itself declining. That is, if we plot price as a function of time, the relationship is most strongly negative for the early industrial years in the US. As the end point is extended, the relationship becomes weaker, and is scarcely perceptible over the full sweep of years (1870–1972). Put differently, there is no single linear trend. A kind of confirmation is provided by Slade (1982), who argues for a U-shaped or quadratic price path over time, and finds evidence of this in separate plots for major metals and fuels. All of this need not be inconsistent with the Barnett– Morse results. It appears that price first falls, as discoveries and technical change reduce costs. But, after a while, discoveries are harder to come by, and costs cannot be reduced indefinitely. The scarcity rent element then takes hold, and begins to drive price movements.

In my view, the revisionists have succeeded in raising doubts about the prevailing view, at least about its implications for the future. But does it matter? Suppose we are running out of (some) resources, can we not substitute others? Much econometric evidence suggests we can. Long run substitution elasticities have been studied extensively for energy materials, at least, and the results are encouraging (for a discussion of this and other results see Pindyck 1978). This does not deny that the transition – to abundant, sustainable energy sources, say – will be painful, at least for some. But given time to adjust and an avoidance of government policies that hinder adjustment (such as oil price controls), prospects seem good if not for continued growth then at least for maintenance of a steady state at something like today's levels in the industrialized countries. There is, however, a qualification. The production and consumption of extractive resources tend to involve relatively heavy use of *environmental* resources. Most air pollution, for example, is associated with energy conversion in one form or another. It is not yet clear to what extent this connection can be broken without at the same time adversely affecting conventional measures of economic welfare.

The Theory of Optimal Depletion

Most recent (post-1973) work in natural resources economics has been concerned with the question of how an exhaustible extractive resource is optimally allocated over time, and of how good a job the market does. The theory also sheds some light on the scarcity debate. Here I shall briefly work through a very basic model, indicate the relevance of the results to the scarcity issue, and sketch a couple of key extensions: to renewable resources, and to the environment.

Let us assume that the problem is to maximize the net present value of social benefit, defined as the sum of consumer and producer surpluses, from a resource deposit. In symbols, this is

$$\max_{\{y_t\}} \int_0^T \left[\int_0^{yt} p(z)\mathrm{d}z - c(y_t)\right] e^{-\mathrm{rt}}\mathrm{d}t \qquad (1)$$

where y_t is the amount of the resource extracted at time t; T is the end of the planning period; $p\,(\cdot)$ is demand for the resource; z is a variable of integration; $c(\cdot)$ is the cost of extraction; and r is the rate of discount. The constraint is given by the finite stock of the resource; in symbols,

$$\int_0^t y_\tau \mathrm{d}\tau = x_0 - x_t$$

Or

$$.x_t \ = -y_t \qquad (2)$$

where x_0 is the initial stock; x_t is the stock at time t; and τ is a variable of integration. Necessary conditions for a maximum are

$$p(y_t) - c'(y_t) - \lambda_t = 0' \qquad (3)$$

where λ_t is an auxiliary variable attached to the constraint equation, and is interpreted as the shadow price of a unit of the resource in the stock, and

$$.\lambda/\lambda = r. \qquad (4)$$

The first condition tells us that, for efficient allocation of an extractive resource, price is *not* equated to marginal cost. Instead, it is equated to marginal (extraction) cost plus the shadow price of the resource in the ground. The wedge between price and cost is often called the resource royalty, or scarcity rent. This is why cost alone can be a poor indicator of future scarcity; it does not capture, as price does, the rent accruing to the finite stock.

The second condition, due originally to Hotelling (1931), is perhaps the most widely known result in natural resource economics. It tells us that, over time, the royalty grows at a rate equal to the rate of interest. Efficiency requires that there be no gain in shifting a unit of extraction from one point in time to another. Proceeds of the sale of a unit extracted today can be invested to yield a rate of return, r. Alternatively, if left in the ground, the unit grows in value at rate r.

It is intuitively plausible, and readily verified by setting up a similar optimization problem for a competitive firm, that the same conditions characterize competitive depletion. This of course assumes no market failure of any kind; one that can be important in a problem where time plays a crucial role is a difference between private and social rates of discount. If, as some have argued, the private rate is above the social rate, then from equation (4) royalty and price will be rising ‘too fast’. Given a downward-sloping demand, this implies that too much of the resource is extracted too soon.

Two Extensions: Renewable Resources and the Environment

The basic model can be extended to deal with renewable resources in a simple and instructive way. The only change is in the constraint equation, which becomes

$$.x_t = g(x_t) - y_t \quad \mathrm{M} \qquad (2')$$

Where $g\,(\cdot)$ is the natural growth, or renewal, as a function of stock size. The second optimality condition becomes

$$.\lambda/\lambda = r - g'(x_t) \qquad (4')$$

The required rate of growth in the royalty is reduced, for $g'(x) > 0$ A unit in the stock yields not just a capital gain, as with an exhaustible resource, but a dividend, in the shape of extra growth. In a steady state $.\lambda/\lambda = 0$, so $g'(x) = r$, and the marginal unit in the stock grows at a rate equal to the rate of interest.

To incorporate environmental considerations into the basic model, let the objective function, equation (1), include a term for value attached to the stock in the ground, $v(x_t)$. This represents the gain from not disturbing the environment (from which the resource is extracted). Then equation (4) becomes

$$.\lambda/\lambda = r. \qquad (4'')$$

The rate of growth in the royalty is reduced, implying that it pays to leave more of the resource in the ground. As in discussion of the scarcity issue, we are only scratching the surface with respect to environmental considerations – the hard choice dictated by lack of space.

Concluding Remarks

Natural resources have played an important role in the evolution of economic thought. Going back at least to Malthus and Ricardo, we might even say that considerations of the impact of resources on economic welfare were central to the founding of the discipline. Yet for much of the 20th century economists have neglected resources, as findings have tended to suggest that they are not growing more scarce, that they are not the limits to growth feared by the classical economists. The pendulum swings, and the classical concern has re-emerged, though in a less dramatic way, and in part tied to environmental impacts of resource use. In the meantime, theory has been enriched by considerations special to extractive neutral resources; price need not be equated to marginal cost, and the behaviour of the wedge in turn has a bearing on the scarcity debate.

See Also

- ▶ Bioeconomics
- ▶ Common Property Rights
- ▶ Depletion
- ▶ Energy Economics
- ▶ Exhaustible Resources
- ▶ Fisheries
- ▶ Renewable Resources
- ▶ Water Resources

Bibliography

Barnett, H.J., and C. Morse. 1963. *Scarcity and Growth; The economics of natural resource scarcity*. Baltimore: Johns Hopkins University Press.

Hotelling, H. 1931. The economics of exhaustible resources. *Journal of Political Economy* 39: 137–175.

Jevons, W.S. 1865. *The coal question*. London: Macmillan.

Johnson, M., F. Bell, and J. Bennett. 1980. Natural resource scarcity: Empirical evidence and public policy. *Journal of Environmental Economics and Management* 7(3): 256–271.

Malthus, T.R. 1798. *An essay on the principle of population*. Reprint of 6th ed, 1826. London: Ward, Lock and Co, 1890.

Meadows, D.H., et al. 1972. *The limits to growth*. New York: Universe Books.

Pindyck, R.S. 1978. *The structure of world energy demand*. Cambridge, MA: MIT Press.

Ricardo, D. 1817. *Principles of political economy and taxation*. Reprinted, London: Everyman, 1926.

Slade, M.E. 1982. Trends in natural-resource commodity prices: An analysis of the time domain. *Journal of Environmental Economics and Management* 9(2): 122–137.

Smith, V.K. 1979. Natural resource scarcity: A statistical analysis. *Review of Economics and Statistics* 61(3): 423–427.

Natural Selection and Evolution

Sidney G. Winter

Important theoretical concepts tend to resist satisfactory definition (cf. Stigler 1957). Such concepts are in the service of the expansive ambitions of the theories in which they occur, and must accordingly respond flexibly to the changing requirements for maintaining order in a changing intellectual empire. The term ‘evolution’ – obviously important in biology, but also in the physical and social sciences – provides a good illustration of this principle. A prominent biologist and author of a highly expansive treatise on biological evolution had the following to offer in his glossary:

> Evolution. Any gradual change. Organic evolution, often referred to as evolution for short, is any genetic change in organisms from generation to generation, or more strictly, a change in gene frequencies within populations from generation to generation (Wilson 1975).

Note the abrupt and radical reduction in the breadth of the conceptual field from the first phrase of this definition to the last. The beginning connects the term to common discourse; the reference to gene frequencies at the end clearly brands the term as belonging to biology, but does not do much to explicate it. The layman is left wondering whether this is meant to cover what happened to the dinosaurs, and perhaps puzzled also as to whether ‘gradual change’ adequately captures the common features of organic evolution, cultural evolution and stellar evolution.

To the extent that biology ‘owns’ the concepts of natural selection and evolution, the meanings of these terms tend to be regarded as biology-specific. It then seems to follow that the application of evolutionary thinking in other realms falls under the rubric ‘biological analogies’, whence it is believed to follow, further, that the appropriateness of an evolutionary approach somehow depends on the closeness of the parallels that can be drawn between the situation in view and situations considered in biology.

The quest for close parallels is substantially impeded by the fact that a prominent feature of the biological scene, sexual reproduction, is, one might say, peculiar. Although asexual or haploid reproduction plays a significant role in biological reality, and this is suitably reflected in portions of biological theory, critics of ‘biological analogies’ tend to stress the question ‘what is the analogue of genetic inheritance?’ with sexual reproduction in mind. A persuasive case can be made that the inability to complete an analogy in this respect is not necessarily a bar to its utility. It is certainly true, nevertheless, that a great deal of biological theory cannot readily be adapted for use in non-biological arenas because the implications of sexual reproduction are so central to the analysis.

This essay puts forward a radical approach to these issues: it challenges biology’s basic ownership claim to the concept of evolution by natural selection. An account of the basic framework of evolutionary analysis is set forth, and while this account attaches meanings to ‘evolution’ and ‘selection’ that are obviously strongly influenced by evolutionary biology, it adapts more readily to discussion of various types of cultural evolution than to biological evolution (at least to the extent that the latter involves sexual reproduction). Examples of the application of the evolutionary viewpoint to economics are then provided in discussions of two areas, the evolution of productive knowledge and the character of Economic Man.

The Framework of Evolutionary Analysis

Fundamentally, and in the most abstract terms, an evolutionary process is a process of information storage with selective retention. Consider, for illustrative purposes, the books in an undergraduate library. Such a library typically has many copies of some books. Given the hazards of loss, pilferage and wear and tear, as contrasted with the comparative constancy of much of the subject matter, the library will not infrequently order new copies of books it has long possessed.

Although each individual volume is informationally complex and in some respects unique, there are nevertheless ‘types’ of books, for example, volumes with the same author and title. Formally, ‘same author and title as’ is an equivalence relation on the set of books, and a relation of particular interest to librarians, students, professors and others. There are, however, a great many other equivalence relations: ‘same publisher as’, ‘same Library of Congress classification as’, ‘same colour as’, and so forth. In fact, given the complexity of the individuals (volumes) that make up the library, the possibilities for defining equivalence relations – which in effect describe alternative approaches to describing the library – are virtually endless.

Now consider the change in such a library over the course of a year – say, at successive annual inventory times when the academic year is over, no books are circulating and all those that are going to be returned have been returned. In terms of a hypothetical exhaustive description of the library, which for example would note every change in yellow highlighting and marginal question marks, the amount of change is enormous in the sense that it would take a great many bytes of information to describe it. A more practical approach to describing the change is to take one or more interesting equivalence relations and count members of equivalence classes at the two dates. For example, for each title-and-author the number of elements in that equivalence class and in the library at t could be counted and the result compared with the number in that same equivalence class and in the library at $t + 1$. While a librarian might be chiefly interested in accounting for the difference in the two numbers, an evolutionary theorist is more likely to divide the latter number by the former and call the result the (observed) 'fitness' of that title-and-author. (Of course, this can only be done provided the denominator is not zero.)

Proceeding along this line, it is possible to discuss how the library evolves (at the title-and-author level) by 'natural selection'. This term refers to the action of the complex collection of processes that are involved in the introduction to and disappearance from the library of individual volumes. The word 'natural' connotes the expectation that these processes cannot be entirely explained by reference to the intent of some individual actor who is effectively in charge of the whole situation – perhaps the head librarian. (Were this expectation not held, the evolutionary approach to understanding the library might well be abandoned in favour of an attempt to fathom the intentions of the controlling actor.)

As described thus far, the evolutionary approach to understanding the library may provide a useful framework, but it is not a theory. In particular, the notion of 'fitness' provides a purely tautological 'explanation' of how the library changes over time. (It is also only a partial explanation, first because of the problem of new acquisitions (zero denominators), but more fundamentally because it treats of a small structure of equivalence relations and does not aspire to complete description.) There is no difficulty in converting this framework into a genuine theory; for example; just assume that 'title and author fitnesses' are constant over time. This theory has abundant empirical content; unfortunately, it is false. A weaker version, substituting 'approximately constant' will fare very little better. The difficulty lies not in the construction, within such an evolutionary framework, of genuine theories with empirical content, but in producing successful ones. More specifically, some non-tautological propositions about theoretical fitness must be derived and turn out to be true of observed fitness. Whether the quest for such propositions proves successful depends on the equivalence relations chosen for study.

In the library example, the choice of title-and-author as the focal equivalence relation for the theory is a masterstroke of creative insight (or would be if it were not obvious). With title and author as taxonomic criteria, a great deal of detailed information about individual volumes is succinctly captured. Also, the fact that there are printers and publishers (and copyright laws) has strong implications for the precision of the 'inheritance' mechanism in this evolutionary system, and the selection mechanism has persistent features reflecting the existence and persistence of academic departments, professors, large enrolment courses, reading lists and library budget levels.

Detailed knowledge of the actual systems governing inheritance and selection would certainly be helpful to the evolutionary scientist seeking to understand the library, but it is not essential. Once 'on to' the idea that 'same title and author' is an important relation in the larger context that affects the evolution of the library, the investigator can make progress without necessarily knowing the answers to a lot of questions about why this idea is fruitful.

So far as the formal, tautological structure of the evolutionary approach is concerned, the investigator could just as well be working with the equivalence classes induced by the relation ‘same word appears as the first word on page fifteen’. The investigator can still count volumes and measure fitness, and it will still be true (*ex post*) that the fittest types come to dominate the library – or more precisely, that approximately equal fitness is a requirement for long-term coexistence in the library environment. It would be surprising, however, if interesting empirical regularities emerged from such an inquiry.

If the foregoing discussion of the evolution of the undergraduate library were an attempt at developing a biological analogy, it would be time to pull back the veil from the correspondences that have not been made explicit thus far. The equivalence classes of ‘same title and author as’ correspond to species. Different editions or printings of a given book correspond to genotypes because there are systematic differences among them, yet the differences are small compared to the differences between classes. Underlining, yellow highlighting, torn pages and the like are examples of phenotypic variation, which reflect the incidents and accidents encountered by an individual volume over its life cycle. The Library of Congress provides a readymade taxonomic structure to facilitate discussion of evolution above the ‘species’ level. Journals are apparently a different life form altogether, since the usual close association of title and author does not prevail.

One could just as well, however, take evolutionary bibliography as the prototypical evolutionary science and think of biology in terms of bibliographic analogies (setting aside, of course, the facts of history and the wide difference in degree of development of the two subjects). In this perspective, the key idea on which the power of the evolutionary approach is seen to rest is that of an equivalence class within which the elements (individuals) are close copies of each other in observable respects. The meaning of ‘close’ involves a contrast between small intra-class variation and large inter-class variation in the system of equivalence classes. Related fundamental ideas are the idea of counting or otherwise measuring the aggregate of elements in such an equivalence class at different points in time, plus the notion that, over time, new individuals appear in a previously existing class – implying that somewhere and somehow, the capacity to produce new individual copies exists.

Biological species that reproduce sexually represent a complex variant of this basic evolutionary paradigm. The part of the process that involves the production of the most exact copies, the replication of chromosomes in the course of gameteogenesis, involves information that is a complete genetic description neither of the parent nor of the offspring. The concept of genetically identical individuals – individuals that are alike the way different copies of the same printing of a book are alike – is prominent in theoretical models, but because of the genetic complexity of individuals and the character of sexual reproduction the phenomenon is rare in the part of nature where sexual reproduction prevails. One consequence is that the concept of a ‘species’, which is so central to evolutionary biology, displays imperfectly resolved tensions between taxonomic criteria and reproductive (inter-breeding) criteria. This difficulty is a peculiarity associated with the phenomenon of sexual reproduction. Perhaps it is in part a reflection of the fact that the major substantive problem of the origin of species is not conclusively solved, and it would be counterproductive to leave no flexibility in the definition of species while pursuing that important goal.

In any case, the contention here is that the empirical application of the framework of evolutionary analysis requires in general the development of a taxonomic system (or more formally, a system of equivalence relations on the set of individuals considered) to which generalized concepts of inheritance, fitness and selection can be applied.

Evolution of Productive Knowledge

Many prominent economists have endorsed some version of the idea that evolutionary principles, or biological science, provide intellectual models that economists would do well to emulate. Marshall's famous dictum that 'The Mecca of the economist lies in economic biology rather than in economic dynamics' (Marshall 1920, p. xiv) is an obvious and important case in point. Thomas (1983) analyses with admirable thoroughness the origin, meaning and implications of this statement in the development of Marshall's thought, emphasizing the central importance of the idea of *irreversible* evolutionary change in economic life. Somewhat less well known, perhaps, is Schumpeter's statement that

> The essential point to grasp is that in dealing with capitalism we are dealing with an evolutionary process . . ..Capitalism, then, is by nature a form or method of economic change and not only never is but never can be stationary (Schumpeter 1950, p. 82).

In Schumpeter's case, too, irreversible change is probably dominant among the connotations of 'evolution', a term which he employed quite frequently.

Neither Marshall nor Schumpeter presented what the above discussion argues to be the key to the development of a predictive evolutionary science – a suggestion about how to interpret economic reality in terms of a system of equivalence relations that effectively breathes empirical content into generalized notions of inheritance and selection. Such a suggestion was advanced, albeit sketchily, by Thorstein Veblen in his paper, 'Why Economics is not an Evolutionary Science' (1898, pp. 70–71, emphasis supplied):

> For the purpose of economic science the process of cumulative change that is to be accounted for is the sequence of *change in the methods of doing things* – the methods of dealing with the material means of life.

Although perhaps not as a result of direct influence from Veblen, a similar proposal (emphasizing imitation of 'rules of behaviour') figures in the classic essay on evolutionary economics by Alchian (1950). The idea is featured more prominently in Winter (1971), and more prominently still, under the rubric of 'routines', by Nelson and Winter (1982). It is the evolutionary economist's answer to an important element in the critique of 'biological analogies' offered by Penrose (1952). (For further discussion, *see* ► Competition and Selection.)

Evolutionary economics thus attaches central importance to a question that is not merely unanswered, but unasked in the context of orthodox economic theory: what are the social processes by which productive knowledge is *stored?* Certainly the concepts of production sets and functions do not seriously evoke this question, and even the bulk of the theoretical literature concerned with technical change disregards the issue as it probes the causes and consequences of things becoming 'known' that were formerly 'unknown'. From an evolutionary viewpoint, abstracting from the storage process in this fashion inevitably has a crippling effect on the effort to understand the appearance of new methods of doing things and the selective pressures to which innovations and innovators are subjected. In particular, the fact may be overlooked that the role of business firms as sources of innovation is intimately related to their social role as repositories of productive knowledge.

These themes cannot be explored in detail here. By way of illustration, however, consider one example of a method of doing things – the method of producing written text that resembles print, called 'typewriting'. There is an equivalence relation 'same (alphabet) keyboard as' on the set of machines used for this purpose, and an equivalence class called 'standard (QWERTY) keyboard'. There is a related human skill called 'touch typing', and an equivalence class of skilled typists 'trained on standard keyboard'. The early evolutionary history of these familiar phenomena has been nicely analysed and described by Arthur (1984) and David (1985). It stands as a warning against simplistic ascriptions of optimality to the outcomes of evolutionary processes. As David explains, the familiar arrangement of keys on the standard keyboard originated as an adaptive response to a particular technical problem – the problem of key jamming produced by typists typing on a machine vastly different from the modern typewriter (be it mechanical, electric, electronic, or a facet of the capabilities of a computer). In particular, the text being produced was invisible to the typist, and jamming of the keys was both hard to detect and serious in its consequences. After many decades of evolution, during which the typewriter itself has been radically transformed, the QWERTY keyboard survives and still performs its intended function of slowing typists down.

David argues convincingly that a central feature of the social process that replicates QWERTY over the generations, to the exclusion of alternatives that permit faster typing, is the complementarity between typewriters and skilled typists. Absent machines with an alternative keyboard, nobody learns an alternative touch typing skill. Absent a good supply of appropriately trained typists, a shift to alternative machines does not pay.

There are some interesting facets of this situation that Arthur and David do not touch upon. One reason that the supply of typists plays the role it does is that touch typing is a tacitly known skill. Although concerned with symbol production, it is not transferable from individual to individual by symbolic communication. One cannot give a lecture to a roomful of typists and thereby convert their skills from one keyboard to another. Typists do not know (in a conscious or articulable way) how they do what they do. As a matter of fact, the level of performance displayed by a highly skilled typist remains mysterious even upon scientific analysis, seemingly surpassing bounds set by known facts of human neurophysiology (Salthouse 1984). The tacit character of typing skill implies high switching costs; the high performance levels achievable even under the QWERTY handicap presumably reduce the incentives to switch (assuming the demand for typing services is price inelastic).

The social process that maintains the QWERTY typewriting method on a large scale is a complex and multi-faceted phenomenon, involving a host of factors traditionally regarded as economic, plus others, such as tacit knowledge, that have more recently entered the disciplinary lexicon. The story of this somewhat obsessive social memory is the story of an innovation; on the hand, it is also a story of how success was precluded for a number of other innovative efforts. In both of its aspects, it has counterparts today. For them, as for QWERTY, understanding how and why methods of doing things *do not* change is fundamental to understanding how and why they *do* change.

Economic Man: The Evolutionary Critique

Economists are wont to regard themselves as hard-headed realists in their assessments of the world in general and of human nature in particular. The trained eye of the economist penetrates facades of pompous pretence, cunning deceit and impassioned demagoguery, discerning the rational pursuit of self-interest in martyr, merchant and murderer alike. Many such penetrating analyses contain, no doubt, an important element of truth. Arguably, the making of them is an important role played by economists and others in a free society. For the purposes of economic science, however, the model of the rational self-interested individual has serious limitations. When it is not a transparent caricature (the textbook consumer who cares only about consumption of goods and services), it is often an obscure tautology (with no definite limits set on what may affect 'utility' and hence choice).

From an evolutionary viewpoint, the key question is which, if any, of the various theoretically described subspecies of *homo economicus* might have been well adapted to the real environments that have shaped humanity. A realistic and *scientific* appraisal of human nature (and the degree and nature of the self interest manifested therein) is an appraisal supportable by reference to the biological and cultural determinants of contemporary human behaviour and the evolutionary forces that have shaped those determinants. If, in a particular instance, the implications of such an appraisal turn out to be different from those of 'hard headed' economic analysis, then economics ought to change – presuming, of course, that the objective in view is the advance of economic science.

Outside of the realm of human motivation, economists routinely (but often implicitly) make use of theoretical assumptions that are plainly not ‘hard headed’ but the reverse. The leading case in point is the assumption that society somehow provides perfect and costless enforcement of contracts. A second case is disregard of social networks (defined by various criteria) as determinants of transacting patterns. One does not have to be imbued with an evolutionary viewpoint, but only moderately experienced in the world, to acknowledge that economic analysis based on such assumptions may yield a seriously distorted image of reality. Where an evolutionary viewpoint comes in handy is in discussing how and why the economy functions as well as it does in spite of the limitations of third party contract enforcement, and the role that non-economic social relations may play in making this possible.

To some extent, the errors introduced by excesses of hard and soft headedness tend to cancel out. Markets perform sometimes well and sometimes poorly, and economics has managed to discover a good deal about this matter in spite of the fact that it has left entirely out of account two major categories of reasons. The burdens of carrying along the two sets of errors have, nevertheless, been heavy. It is important to leave them behind.

Progress is being made in doing so. As economics breaks out of the shell formed by its first approximation assumptions, its relationships to other social sciences and to biology become both more obvious and more fruitful. The interwined themes of the role of self interest in behaviour and the bases of social cooperation are fundamental not just in economics but in all of social science, and in much of biology as well. Jack Hirshleifer, who has repeatedly and insightfully emphasized the universality of these themes, recently proclaimed that ‘*there is only one social science’* (1985, p. 53). For a ‘generalized economics’ to serve as that one social science, economics ‘will have to deal with man as he really is – self-interested or not, fully rational or not’ (ibid., p. 59).

Although it is probably premature to announce a contest to provide the best name for unified social science – a contest that would no doubt evoke numerous alternatives to ‘generalized economics’ – it does seem that many of the elements are at hand for a move toward unification. Major contributions from a variety of directions have vastly improved understanding of how cooperative behaviour in general and exchange behaviour in particular can arise in spite of weak or nonexistent institutional support. Some of these involve explicit use of the evolutionary framework (e.g. Axelrod 1984); some do not (e.g. Williamson 1985). All are at least potentially adaptable to a general multi-level evolutionary scheme in which patterns reproduced by a variety of mechanisms are subjected to selective pressure. Major difficulties, and major controversies, attend the problem of characterizing the linkages between the levels. On this front too there is recent progress, particularly the work of Boyd and Richerson (1985), who study the interactions of biological and cultural evolution with the aid of a collection of ‘dual inheritance’ models. Such interactions have, of course, implications for the understanding of human biology as well as for the study of culture.

In sum, natural selection and evolution should not be viewed as concepts developed for the specific purposes of biology and possibly appropriable for the specific purposes of economics, but rather as elements of the framework of a new conceptual structure that biology, economics and the other social sciences can comfortably share.

See Also

- ▶ Bioeconomics
- ▶ Competition and Selection
- ▶ Game Theory
- ▶ Hunting and Gathering Economies

Bibliography

Alchian, A. 1950. Uncertainty, evolution and economic theory. *Journal of Political Economy* 58: 211–221.

Arthur, W.B. 1984. Competing technologies and economic prediction. *Options* (I.I.A.S.A., Laxenburg, Austria), 10–13.

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

Boyd, R., and P. Richerson. 1985. *Culture and the evolutionary process*. Chicago: University of Chicago Press.

David, P. 1985. CLIO and the economics of QWERTY. *American Economic Review* 75(2): 332–337.

Hirshleifer, J. 1985. The expanding domain of economics. *American Economic Review* 75(6): 53–68.

Marshall, A. 1920. *Principles of economics*, 8th ed. London: Macmillan, 1953.

Nelson, R., and S. Winter. 1982. *An evolutionary theory of economic change*. Cambridge, MA: Harvard University Press.

Penrose, E. 1952. Biological analogies in the theory of the firm. *American Economic Review* 42: 804–819.

Schumpeter, J. 1950. *Capitalism, socialism and democracy*, 3rd ed. New York: Harper.

Stigler, G. 1957. Perfect competition, historically contemplated. In *Essays in the history of economics*, ed. G. Stigler. Chicago: University of Chicago Press, 1965.

Salthouse, T. 1984. The skill of typing. *Scientific American* 250(2): 128–135.

Thomas, B. 1983. *Alfred Marshall on economic biology*. Paper presented to the History of Economics Society, May.

Veblen, T. 1898. Why economics is not an evolutionary science. In *The place of science in modern civilization*, ed. T. Veblen. New York: Russell & Russell, 1961.

Williamson, O. 1985. *The economic institutions of capitalism*. New York: Free Press.

Wilson, E. 1975. *Sociobiology: A new synthesis*. Cambridge, MA: Harvard University Press.

Winter, S. 1971. Satisficing, selection and the innovating remnant. *Quarterly Journal of Economics* 85(2): 237–261.

Natural Wage

Krishna Bharadwaj

The notion that there exists a fixed subsistence level of wages appears to have emerged in Europe in the 17th and 18th centuries, both as an empirical observation on the extant conditions of the labouring poor and as a plank for mercantilist labour policy. An analytical advance was gained by the Physiocrats when they considered the implications of a 'given wage' in terms of the circular process of reproduction of the social economy. Attempts followed thereafter to define the norm of subsistence and mechanisms by which a variation from the norm sets up tendencies to restore it. Adam Smith, more than any of his predecessors, perceived clearly the logic of the evolving capitalist system and provided definitions, categories and the basic frame of analysis in terms of which the future questions in political economy were to be cast and developed. Recognizing profits as a category separate from rents and wages, the emergence of 'free' labour and the competitive tendencies towards the uniformity of the rate of profit and of wages, he made an analytical distinction between persistent (or 'permanent') and transitory (or, accidental) forces in operation – the former tending the economy to a 'natural' state while the latter characterized by 'market' forces, generating fluctuations around the 'natural' or central position. Thus a significant and later well-established distinction was made between 'natural price' and 'market price'.

> When the price of any commodity is neither more nor less than is sufficient to pay the rent of the land, the wages of the labour, and the profits of the stock employed in raising, preparing and bringing it to market, according to their natural rates, the commodity is then sold for what may be called its natural price. (Smith 1776, p. 55)

At such a price, the quantity brought to market is just sufficient to supply the effectual demand. In case of any deficiency or excess of supply over effectual demand, the market price deviates from the 'natural'.

A natural rate of wages was analogously conceptualized by Smith, around which there could be deviations due to particular transitory factors. The distinction between natural wage and market wage was to be formally and rigorously spelled out by Ricardo, following Torrens (see below). Adam Smith wove his theory of what determines the level of wages from an interesting variety and complex of factors, synthesizing the preceding discussions on wages by Petty, Child, Necker, Cantillon, the Physiocrats and Turgot. His theory of wages proceeded on two related strands: having clearly identified the three classes with their respective revenues, profits, rents and wages, the first strand explored the struggle for distribution among the classes, with 'rents' and 'profits' perceived as deductions from the produce of labour.

> What are the common wages of labour, depends everywhere upon the contract usually made between the two parties, whose interests are by no means the same. The workmen desire to get as much, the masters to give as little as possible. (Smith 1776, p. 66)

In the uneven contest, the masters enjoy powerful advantages in the ease with which they can combine ('Masters are always and everywhere in a sort of tacit, but constant and uniform combination, not to raise the wages of labour above their actual rate': pp. 66–7). In the protection of their interests through the tacit of explicit support from the state and its statutes (e.g. outlawing strikes by workers) and in economic security, a privilege of the propertied classes, contrasted with the abject dependence of the workers who can not 'hold out' as long as their employers. ('In the long run, the workman may be as necessary to his master as his master to him, but the necessity is not so immediate' p. 66.) This struggle implied that no definite fixed level could be ascribed to wages, but Smith held that there was a lower limit, determined by the necessary means of subsistence 'below which it seems impossible to reduce, for any considerable time, the ordinary wages of even the lowest species of labour' (p. 67). This, however, was not a physiologically determined subsistence; for 'in order to bring up the family, the labour of the husband and wife together must, even in the lowest species of common labour, be able to earn something more than what is precisely necessary for their own maintenance' (p. 68). The lowest rate was in fact considered as the one 'consistent with common humanity' (Smith 1776, p. 68). The second strand in Smith was the economic factors that influenced and, were influenced by, the social struggle. In a rapidly progressing economy where demand for labour is increasing, competition among masters 'breaks through their natural combination not to raise wages'. In a stationary economy, even if incomes were at a high level, competition among workers and masters would soon reduce wages to the lowest rates 'consistent with common humanity'. In a decaying economy, 'want, famine, mortality' would provide the corrective through 'the number of inhabitants in the country getting reduced to what could be easily maintained by the revenue of the state'. To this was also added the response of population to wages when they are higher or lower than the average; although he observed that 'in civilized society, it is only among the inferior ranks of people that the scantiness of subsistence can set limits to the further multiplication of the human species' (p. 79). 'The liberal reward for labour', which is seen as a result of rapid accumulation, can enable them 'to provide for the children and consequently bring up a greater number'. A wage decline has the contrary effect. Thus, while the pace of accumulation would influence the demand for labour and wages, the supply of workers could also adjust. However, it was supply which was seen basically *adjusting* to demand: 'It is in this manner that the demand for men, like that for any other commodity, necessarily regulates the production of men' (p. 80). This led Smith to advance the concept of natural wage as

> The wages paid to journeymen and servants of every kind must be such as may enable them, one with another, to continue the race of journeymen and servants, according to the increasing, diminishing, or stationary demand of the society may happen to require. (p. 80)

While in this statement, it would appear as if Smith had considered a purely supply-and-demand determined wage, the following position of Smith indicates that he was concerned with the systematic shifts in the natural rates which then continue to act as the new central norms: 'The demand for labour, according as it happens to be either increasing, stationary, or declining, or to require an increasing, stationary, or declining population', regulates the subsistence of the labourer and determines in what degree it shall be either liberal, moderate or scanty. In fact that such a norm was presupposed is evident from Smith's acceptance of the proposition that money wage moves with the price of provisions and a tax on necessities is shifted onto rents and profits. Smith also considered explicitly, as did Ricardo, following him, that there could be a spectrum of natural wage rates for different skills, gradations and intensity of labour, 'the proportion between different rates both of wages and profit in the different employments of labour and stock seems not to be much affected . . . by the riches or poverty, the advancing, stationary, or declining state of the society' (p. 143).

It was left to Ricardo to set out clearly the distinction between 'natural' and 'market' wage and discuss the relation between the two. Ricardo's view of wages was greatly influenced by Torrens's *Essay on the Corn Trade* (1815) where, regarding labour as a commodity, Torrens stated

> It therefore has, as well as anything else, its market price and natural price. The market price of labour is regulated by the proportion which, at any time, and at any place, may exist between the demand and the supply; its natural price is governed by other laws and consists in such a quantity of the necessaries and comforts of life, as from the nature of the climate and the habits of the country are necessary to support the labourer, and to enable him to rear such a family as may preserve in the market an undiminished supply of labour.

Thus there could be variations in the natural price of labour due to differences in habit, custom and also 'different stages of national improvement' but may be regarded as 'very nearly stationary' in any given time and place; whereas, the market price of labour 'fluctuates perpetually according to the proportion between demand and supply'. Again the difficulty or ease of maintaining family, deaths or prudential checks on marriage, tended to push the market wage towards the natural rate via the adjustment of the supply of labour called forth by the deviation.

Ricardo, while closely following Torrens, defined the natural price of labour as 'that price which is necessary to enable the labourers, one with another, to subsist and perpetuate their race without increase or diminution' (*Principles*, p. 93); or they are the wages that maintain the population *stationary*. The natural wage is also stable in real terms ('quantity of food, necessaries and conveniences become essential to him from habit') so that a rise in price of necessaries raises the natural wage and the great difficulty encountered in producing food, the major component of wage, induces a tendency for the natural price of wages to rise. (Here, it must be remembered that Ricardo often uses the term natural price of wage to mean 'value of wages', or, labour embodied in the production of the necessary wage.)

Ricardo formulates more clearly the tendency of market wage to conform to natural wage via the adjustment of the supply of labour. The demand for labour is itself generated by the process of accumulation which however appears as given independently. Ricardo concedes that, notwithstanding, the tendency of wages to conform to their natural rate, a continuous and constant increase of capital may keep the market rate above the natural rate for an indefinite period. Ricardo thus analyses the effects of accumulation on two counts – quantity of capital (food and necessaries) may increase while at the same time, the difficulty of their production may increase too, increasing thus the *value* of capital along with its quantity. In this case, the natural price of wages would rise along with the price of necessaries. If, on the other hand, capital does not meet increases in value, but only in quantity, the natural price of labour remains stationary or may even fall (because of the possible cheapening of other non-food products in the wage). However with the increasing capital (in quantity), the market price of labour would rise in both cases because of the increased demand for labour and would set to work the adjustment process of the supply of labour. How far and how fast the tendency of the market price to restore the natural price of labour would work, depends both on the influence of accumulation on the natural price itself and the rapidity of the supply adjustment.

Ricardo did not subscribe to any 'iron law of wages' and made that explicit:

> It is not to be understood that the natural price of labour, estimated even in food and necessaries, is absolutely fixed and constant. It varies at different times in the same country, and very materially differs in different countries. (p. 96)

Malthus objected to Ricardo's concept of natural price of labour which he himself defined as 'that price which, in the natural circumstances of the society, is necessary to occasion an average supply of labour sufficient to meet the average demand' (*Principles*, p. 29). As Cannan (1903, p. 257) remarks, '. . . by this rather cloudy phrase he seems to mean nothing more or less than the actual wages which are paid in a year not marked by any exceptional circumstances'. He thus rejected entirely not only the idea of a rigid level of wages faced by physiological necessity but also of a ‘given’ level rendered stable by ‘habit’.

Among the Ricardian followers, the centrality of the notion of natural wage appears to have been subordinated to the theory that wages are determined by the proportions of capital (wage fund) to labour and was even eliminated altogether. James Mill in his *Elements* concentrated his theory entirely on *changes* in wage, depending upon the varying proportion of capital to labour, without any mention of the natural wage. The wages fund doctrine emerged in John Stuart Mill’s *Principles*.

Among the marginalists, Marshall, who sought to establish continuity with classical writers, emphasized, in his descriptive accounts, the element of custom, habit and conventions, but introduced, apart from necessaries of subsistence, ‘earnings for efficiency’ linked with productivity. He saw these elements as ‘the many peculiarities in the action of demand and supply with regard to labour which are of a vital character’. For, ‘they affect not only the form but also their substance’, and ‘limit’ to some extent the action. However, ‘the correct position to take’, he advised, was ‘not to measure the influence of supply and demand by their first and obvious effects’. The influence of custom was only the ‘cumulative effect’ of their past operation (*Principles*, p. 559).

See Also

- ▶ Corn Model
- ▶ Iron Law of Wages
- ▶ Natural Price
- ▶ Wage Fund Doctrine
- ▶ Wages in Classical Economics

Bibliography

Cannan, E. 1903. *A history of the theories of production and distribution*, 2nd ed. London: P.S. King & Son.

Malthus, T.R. 1836. *Principles of political economy considered with a view to their practical applications*, 2nd ed. London: William Pickering.

Marshall, A. 1920. *Principles of economics*, 8th ed. London: Macmillan, 1947.

Mill, J. 1821. *Elements of political economy*, 3rd ed. London: Baldwin, Cradock, and Joy, 1826.

Mill, J.S. 1848. In *Principles of political economy, with some of their applications to social philosophy*, ed. Ashley Sir William. London: Longmans, Green & Co, 1909.

Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*. London: W. Strahan and T. Cadell.

Torrens, R. 1815. *An essay on the external corn trade*. London: Hatchard.

Navier, Louis Marie Henri (1785–1836)

R. F. Hébert

Keywords

Consumption externalities; Cost–benefit analysis; Demand theory; Dupuit, A.-J.; Jointness of consumption; Navier, L.; Pigou, A.; Public goods; Public works; Samuelson, P, on public goods; Subjective utility; Utility measurement

JEL Classifications

B31

A French engineer and economist, Louis Marie Henri Navier was a pioneer in the construction of suspension bridges, and is also known as the creator of that branch of mechanics known as structural analysis. In his economic inquiries, he sought a practical measure of public utility that provided the springboard for Dupuit’s pioneer contributions to demand theory. Orphaned at the age of nine, Navier was adopted by his great-uncle, the celebrated architect–engineer, Émiland-Marie Gauthey (1732–1806), who likely inspired his adopted son to follow in his illustrious footsteps. Navier died prematurely at the age of 51, thus cutting short a distinguished career of public service.

Navier was one of the earliest formulators of a cost–benefit rule to guide the construction of public works. His rule advocates expenditures on public works if the total benefit derived – in the form of before–after cost savings – exceeds the total recurring costs of the new construction. In choosing recurring costs over total costs as the element to be covered by tolls, Navier was showing a greater appreciation of consumption externalities than Pigou (1947, p. 3n.), who wrote more than a century later. In fact, Navier's rule is a somewhat less sophisticated version of Stephen Marglin's (1967, pp. 22–4) 'myopic rule' of public investment.

Navier's rule was the proximate cause of Dupuit's innovative attempt to establish demand based on subjective utility. Dupuit (1844) objected to Navier's attempt to measure utility on two grounds: (*a*) in competitive markets the proper measure of utility of the quantity of goods and services consumed is not the reduction of transport costs but rather the reduction of production costs; (*b*) increases in the quantity taken at lower prices do not all have the same utility, but rather take on smaller values as more is consumed. Thus, Dupuit's rule overcame the limitations of Navier's rule, and, in addition, launched the neoclassical theory of demand. Kölm (1968) argues that, in the context of public finance, Dupuit's rule moves us closer to Samuelson's (1954, pp. 387–9) decision rule regarding public goods. However, a valid comparison of Dupuit's performance with Samuelson's must recognize that Samuelson employed a highly restrictive definition of a public good and the assumption of true consumption jointness – aspects missing from Dupuit's analysis or from Navier's.

See Also

- ▶ Consumption Externalities
- ▶ Cost–Benefit Analysis
- ▶ Dupuit, Arsene-Jules-Emile Juvenal (1804–1866)
- ▶ Pigou, Arthur Cecil (1877–1959)
- ▶ Public Goods
- ▶ Public Works

Selected Works

1832. De l'exécution des travaux publics, et particulièrement des concessions. *Annales des Ponts et Chaussées: Mémoires et Documents,* 1 ser. 3, 1–31.

1835. Note sur la comparison des avantages respectifs de diverse lignes de chemins de fer, et sur l'emploi des machines locomotives. *Annales des Ponts et Chaussées: Mémoires et Documents,* 1 ser. 9, 129–179.

Bibliography

Coronio, G. 1997. *250 ans de L'École des Ponts en cent portraits*, 86–87. Paris: Presses de l'école des Ponts et Chaussées.

Dupuit, J. 1844. On the measurement of the utility of public works. *Annales des ponts et chaussées,* 2d ser. 8, 332–375. Trans. R. Barback, *International Economic Papers* 2 (1952), 83–110.

Ekelund, R. Jr., and R. Hébert. 1978. French engineers, welfare economics, and public finance in the nineteenth century. *History of Political Economy* 10: 636–668.

Ekelund, R. Jr., and R. Hébert. 1999. *Secret origins of modern microeconomics: Dupuit and the engineers*. Chicago: University of Chicago Press.

Etner, F. 1987. *Histoire du calcul économique en France*. Paris: Economica.

Hébert, R. 1994. Fondements et développements de l'économie publique. *Revue Du Dix-Huitième Siècle* 26: 37–49.

Kölm, S.-C. 1968. Léon Walras' correspondence and related papers: The birth of mathematical economics. *American Economic Review* 58: 1330–1341.

Marglin, S. 1967. *Public investment criteria: Studies in the economic development of India*. Cambridge, MA: MIT Press.

Pigou, A. 1947. *A study in public finance*. 3rd ed. London: Macmillan.

Samuelson, P. 1954. The pure theory of public expenditures. *Review of Economics and Statistics* 36: 387–389.

Necessaries

G. Vaggi

The classical economists, Smith and Ricardo in particular, used the term 'necessaries' to indicate 'the commodities which are indispensably necessary for the support of life', and also 'whatever the custom of the country renders it indecent for creditable people, even of the lowest order, to be without' (Smith 1776, vol. 2, pp. 869–70). Thus, necessaries include not only the goods which are strictly required for the survival of workers and their families, but also all the commodities which by habit and custom are regarded as 'necessary to the lowest rank of people' (ibid.). Thus the term includes a purely physical element and a sociological one. Smith distinguishes necessaries from luxuries, which are all the goods which are not strictly required to guarantee the workers a decent standard of living.

The prices of necessaries are extremely important in the determination of the money wages of the workers, because these commodities make up the consumption basket which defines the historically determined level of subsistence. Changes in the prices of necessaries modify money wages and in this way they influence the prices of all manufactured products (ibid.; Ricardo 1821, p. 93). The distinction between necessaries and luxuries is important with respect to fiscal policy; a tax on the sales of necessaries increases their price and money wages, and has negative effects on the markets for all other commodities. In fact, it is typical of necessaries to enter into the production of every commodity, because they are the consumption goods of the workers. An excise tax on a luxury good does not influence the prices of all other commodities (Smith 1776, vol. 2, pp. 873, 888; Ricardo 1821, p. 241). Since real wages cannot be compressed indefinitely, a tax which raises the prices of necessaries will ultimately fall on the revenue of the landlords and of rich people, who have no interest in taxing these commodities (Ricardo 1821, p. 235).

The necessaries of life are not only primary goods, but also include manufactured products (ibid., p. 243). However, both Smith and Ricardo accepted the Physiocratic view that agricultural products make up most of the value of real wages (Quesnay 1767, p. 258). For this reason Ricardo believed that the price of necessaries would rise with the progress of society and the increase of population – because of the existence of diminishing returns in agriculture more labour is required in the production of necessaries of life (see Ricardo 1821, p. 101). Thus the wages of productive workers become more expensive, and the rate of profit falls because wages make up the largest part of the country's circulating capital.

The distinction between necessaries and luxuries was a traditional feature of classical economics (see J. Mill 1808, pp. 126–9; J.S. Mill 1848, p. 193). This concept was criticized by Marshall because of the difficulty of establishing whether a commodity belonged to necessaries or luxuries (see Marshall 1920, pp. 56–7). Nevertheless he still used the notion of necessaries, in particular in his analysis of the elasticity of demand. Since necessaries are essential elements of consumption, their demand schedule is highly inelastic (ibid., pp. 89–91).

See Also

- ▶ Engel's Law
- ▶ Wage Goods

Bibliography

Marshall, A. 1920. *Principles of economics*, 8th ed. London: Macmillan, 1964.

Mill, J. 1808. *Commerce defended*. Edinburgh: Oliver & Boyd, 1966.

Mill, J.S. 1848. *Principles of political economy*. Harmondsworth: Penguin, 1970.

Quesnay, F. 1767. The General Maxims for the economic government of an agricultural kingdom. In *The economics of physiocracy*, ed. R.L. Meek. London: Allen & Unwin, 1962.

Ricardo, D. 1821. *On the principles of political economy and taxation*. 3rd ed. In *The works and correspondence of David Ricardo*, ed. P. Sraffa. Cambridge: Cambridge University Press, 1951.

Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*. Oxford: Oxford University Press, 1976.

Necker, Jacques (1732–1804)

A. Courtois

Necker was born and died at Geneva. His character was an unusual mixture of qualities rarely united in one individual. A very able and honest banker, he established a house of the highest standing at Paris – Thélusson, Necker & Co. –

and rapidly accumulated a large fortune; satisfied with the wealth he had acquired, he retired from business at the age of forty to devote himself to politics and literature. He believed himself possessed of sufficient capacity to lead the political world, and that at a moment when it was in the utmost disorder. Dexterous in the use of expedients, and but slightly burdened with theory, he flattered himself that he would eclipse Turgot, whose inferior he was, especially in grasp of principle. His first work, the *Eloge de Colbert*, received a prize from the French Academy in 1773, he then wrote *De la législation et du commerce des grains* (1775), which, dogmatic in style and opposed to the views of Turgot, had considerable success, and even contributed to the fall of that minister (19 May 1776). On Turgot's successor, de Clugny, dying, 30 October 1776, Taboureau des Reaux was appointed to succeed him, and compelled to accept Necker as his coadjutor. This led to his resignation 1 July 1777, when his duties were handed over to Necker under the title of Directeur-général des finances. Though acting as Contrôleur-général, he was not granted that title, as this would have admitted him to the council of state, and he was a protestant. In this, his first essay in finance, Necker showed marked ability, diminishing the expenses, simplifying the machinery of the administration, and, through his connection with the great Bank, obtaining exceptionally favourable terms for the treasury. The tide of public opinion began now to set in the direction of the convocation of the Etats Généraux. In 1781 Necker's famous *Compte Rendu au Roi* appeared, addressed rather to the public than to the head of the state. His popularity increased; the success of his report, the first of its class, though incomplete, was great. The condition of the finances of the country was improved, but an unexpected result occurred. Cabals were roused against him, perhaps fomented by Necker's extraordinary vanity and his folly in mixing praises of his wife, whose *salon* was celebrated, with his official reports. The court became hostile, and in 1781 he was compelled to resign. But the weaknesses of the best-known of his successors, Calonne, caused the public to think with regret of the fallen minister, and the publication of *De l'administration des finances de la France* (1784), contributed to strengthen his popularity. This work, like those which Necker had written previously, is marked by an absence of general principle; it was declamatory and exaggerated in style, but valuable to those who would study how the finances of France were managed in the last days of the old régime.

Necker was detested by the court as a protestant and a bourgeois, nevertheless Louis XVI found himself compelled to recall him to power, 20 August 1788, this time also with the title of Directeur-général des finances. The financial position was serious. The payment of the interest of the public debt was suspended, the treasury empty; Necker's return to power inspired confidence, and, as if by magic, money reappeared. He had, however, to employ his private resources to sustain the public credit. Though the court was still hostile, the multitude applauded him. When he spoke of retirement the court was compelled to ask him to remain in office, but by one of those sudden turns of fortune so frequent at this period, the king intimated to him his dismissal, 11 July 1789, and ordered him to leave France secretly. Necker obeyed and returned to Geneva. The effect of his departure on public opinion was terrific. In the midst of these disturbances the Bastille was taken, and on 29 July, Necker was recalled by the court with the title of Premier ministre des finances, and was admitted to the council. His return was an unparalleld triumph. In every town that he passed through between Switzerland and Paris the horses were taken out of his carriage and he was drawn by the admiring people. This mad enthusiasm could not last. Some slight errors in judgement alienated public opinion, and on 8 September 1790 he was again compelled to leave office and France, this time for ever. The populace was indifferent, if not hostile. In a small town in Champagne, he, who had never deigned to accept the salary attached to his high office, was arrested as a malefactor. How little he had deserved this may be understood from the fact that he had left behind him at the treasury, to assist the public credit, £96,000, his own property, which was only returned to his daughter the well-known Madame de Staël-Holstein in the early years of the Restoration. An order had to be obtained from the national assembly to enable Necker to regain his liberty and to return to Switzerland.

Of Necker's later works we need only mention: *Sur l'administration de M. Necker par lui-même*, in one volume, 1791. His work on *La législation et le commerce*, is inserted in the economic collection of Guillaumin.

[Adam Smith called Necker 'a mere man of detail'. Sir J. Mackintosh is the authority for this (Rae 1895, p. 206).

Selected Works

1773. *Eloge de Colbert*. Paris.

1775. *De la législation et du commerce des grains*. Paris. 1781. *Compte rendu au Roi*. Paris.

1784. *De l'administration des finances de la France*, 3 vols. Paris.

1791. *Sur l'administration de M. Necker par lui-même*. Paris.

References

Carré, A. 1903. *Necker et la question des grains à la fin du 18e siècle*. Paris.

Coquelin, C., and Guillaumin, M. 1852. *Dictionnaire de l'économie politique*, 2 vols. Paris.

Rae, J. 1895. *Life of Adam Smith*. London: Macmillan.

Nef, John Ulric (Born 1899)

Colin G. Clark

Born in Chicago on 13 July 1899, Nef was educated at Harvard (SB, 1920) and the Robert Brookings Graduate School in Washington, DC (PhD, 1927). Almost his entire academic career was spent at the University of Chicago.

The coal trade in London was the subject of one of Nef's first researches (1932). At the time it was believed that in the early 19th century (apart from tariff protection) business was highly competitive, and that cartels were to come only late in the century. Nef's book on combination in that trade showed that, contrary to expectation, a high degree of cartelization could and did occur even then.

London was a great consumer of fuel, with negligible supplies of fuel wood, and, before the railway age, dependent on coal transported from the north of England and sailing ships. The costs of wagon transport being so high, commercially available coal could only be brought from mines close to navigable estuaries, particularly the Tyne. Even so, a certain amount of wagon transport was necessary. Some coal owners began laying wooden rails from the mines to the waterfront, thereby greatly increasing the load which each horse could draw. From these primitive railways was obtained the standard gauge of 4 feet 8 1/2 inches, which was to spread around the world. Cartelization and price fixing were strongly enforced among both North Country suppliers and shippers delivering in London. These cartels had some connection with the ancient medieval guilds of privileged traders, which elsewhere had died out.

There was once a widely held idea that an 'Industrial Revolution' occurred quite suddenly in Britain in the closing decades of the 18th century. Rostow's work has done much to reinforce this misconception. Nef showed (e.g. 1943) that while there had been an acceleration of progress in the last decades of the 18th century, the industrial development of England should really be said to have begun as early as the 16th century. A comparable study for France showed that development started later, and that it was slowed down by the extraordinary quantity and detail of bureaucratic regulations; also, he added, by the slow growth of the size of the market – French population growth had already slowed down by the latter decades of the 18th century.

On one occasion I asked Nef how it was that the Dutch, clearly Europe's most productive economy in the 17th century (Sir William Petty had established Dutch superiority over France and England in productivity per head), had failed to get into industrial development until much later. Nef gave the interesting reply that the Dutch were more concerned with quality than with quantity. The Dutch obtained their high incomes mainly from trade and shipping, and there was not the same economic compulsion to embark on manufacture. The 17th century had been the Dutch great age, in war, commerce, colonization, art. The Dutch themselves regard the 18th century as a period of decadence. There was also a marked slowing down in population growth.

Nef was closely associated with Robert Hutchins, the dramatic head of Chicago University, appointed at a very early age, in 1930. Hutchins encouraged Nef to establish the ‘Committee on Social Thought’ as a department in the University. The object was to provide for interchange of ideas between different departments in the University, which had become, he considered, too narrowly specialized. However, like his other reforms, this initiative of Hutchins was not a success, and was quickly abandoned after his retirement.

Having married into the influential Castle family in Hawaii (who had originally gone there as missionaries, and then developed large interests in sugar and shipping), Nef had many valuable contracts and opportunities for meeting people from many countries. He was able to bring to Chicago, city as well as university, a considerable intinction of European philosophy, culture and art. He was one of the few foreigners who had the honour of being elected to the Collège de France.

In 1950 Nef published a book on the ominous subject of whether the world could have made the same progress, economic and social, without the stimulus of war. His great historical knowledge certainly provided plenty of material for this case. It was only war, in the physical sense, not merely international tension, which brought about the principal developments in the European metal trades; and the same might be said of the side-effects of the American Civil War. Keynes said that war was ‘a great sifter, bringing the right men to the top’. In present times, it would be hard to deny that our extraordinary progress in all branches of electronics would have been at the same pace without the continuous stimulus of military demand.

Selected Works

1932. *The rise of the British coal industry*. London: G. Routledge & Sons.

1940. *Industry and government in France and England, 1540–1640. Memoirs of the American Philosophical Society*, vol. 25. Philadelphia: American Philosophical Society.

1941a. Industrial Europe at the time of the reformation (ca. 1515–ca. 1540) Pts I–II. *Journal of Political Economy* 49: 1–40; 183–224.

1941b. Silver production in central Europe, 1450–1618. *Journal of Political Economy* 49: 575–591.

1942a. War and economic progress 1540–1640. *Economic History Review* 12(1–2): 13–38.

1942b. *The United States and civilization*. Chicago: University of Chicago Press. 2nd ed, revised and enlarged 1967.

1943. The industrial revolution reconsidered. *Journal of Economic History* 3: 1–31.

1944. Wars and rise of industrial civilization 1640–1740. *Canadian Journal of Economics and Political Science* 10: 36–78.

1950. *War and human progress; an essay on the rise of industrial civilization*. Cambridge, MA: Harvard University Press; London: Routledge & Kegan Paul. 2nd ed, 1963.

1958. *Cultural foundations of industrial civilization*. London: Cambridge University Press.

1964. *The conquest of the material world*. Chicago: University of Chicago Press.

1973. *Search for meaning; the autobiography of a nonconformist*. Washington, DC: Public Affairs Press.

Negative Income Tax

Harold W. Watts

The negative income tax is a concept which inspired an interesting crop of income transfer proposals aimed at reforming the welfare system of the 1960s. A substantial amount of analytic effort and empirical research was generated around the basic notion of negative taxes. Income transfer policy has been and continues to be influenced by this innovation.

A negative income tax, or NIT for short, assesses the size of entitlement for a beneficiary unit on the basis of its income flow. This is entirely analogous to an income tax that assesses liability on an income base. In this sense both positive and negative income taxes are potentially as various as the range of schedules that can be devised to define the relation between income and the tax liability or entitlement for particular types or sizes of units. Typically, the NIT provides subsidy benefits that increase linearly or at least continuously with the (negative) deviation of income from some zero-benefit or 'break-even' income level, denoted by B. A schedule of this kind extended to the zero level of income defines a maximum benefit payable to units with no income, and that amount is often called the 'guarantee', or G. In the case of a simple linear NIT the 'tax rate' or benefit reduction rate, r, must be equal to the ratio of the guarantee to the break-even level of income or, in symbols:

$$r \equiv G/B.$$

Using this identity the relation between the benefit payment, P, and income, Y, can be written as:

$$P = r(B - Y)$$

or

$$P = G - rY,$$

where

$$Y \leq B$$

and

$$P = 0,$$

where

$$Y > B.$$

The guarantee determines the minimum income for an eligible filing unit. The filing unit is usually the income sharing household or family, and guarantees are varied according to its size. Some versions specify the guarantee according to the age of each individual and simply add them up to get the family guarantee.

As developed in the 1960s the NIT owed something to the intellectual heritage of the Social Dividend promoted by Lady Rhys-Williams in England after World War II (Rhys-Williams 1943). The Speenhamland system of 'outdoor relief' that was abandoned in England with the passage of the Poor Law of 1834 is an earlier predecessor. The NIT idea developed with much more recognition of incentive effects than Speenhamland, and much less sweeping comprehensiveness than the Social Dividend (Green 1967).

Milton Friedman (1962) is generally credited with coining the term 'negative income tax' when he introduced a simple scheme of that kind in *Capitalism and Freedom*. His plan used existing exemptions and standard deductions of the current positive tax law to determine breakeven income thresholds. Friedman applied a 50 per cent 'tax' rate to the difference between gross income and the threshold to determine the size of the benefit entitlement. These rules produce a guarantee equal to half the total of exemptions and standard deductions.

Lampman (1965) Tobin (1965) and Tobin et al. (1967) contributed importantly to the development and currency of schemes using benefit formulae and structures with strong analogues in the tax system as a means of paying transfers to the poor. Proposals by these economists and others sometimes included full integration with the positive tax system. Unified tax and transfer mechanisms can achieve both administrative economies and reduced distortion of incentives by imposing constant marginal rates over the range of negative and positive tax liabilities.

The NIT concept was developed mainly by economists who were at once convinced of the need for expanded transfers as a component of the effort to eliminate poverty and concerned about the complex and dysfunctional incentives that were apparent in existing transfer programmes. In that context the NIT has some very attractive features, at least potentially. Its explicit tax rate provides a focus for considering and adjusting the effect on work incentives of reducing after-tax net wage rates. The guarantee, where the negative tax is the only source of subsidy for all or most poor persons, can be directly evaluated in terms of the adequacy of the budgets they afford or compared with poverty thresholds themselves. An NIT paying its benefits in cash and replacing a group of programmes offering a mixture of cash and in-kind benefits that fails to cover all those equally needful offers clear gains in both efficiency and equity as these are understood by economists.

The NIT also makes it possible to subsidize the income of households that are poor despite the full-time efforts of at least one breadwinner (usually due to the combined effect of low earning capacity and large family size). Such units had been excluded categorically from existing Federally supported welfare programmes because of fears that honest workers would develop dependent habits or that employers would conspire to cut wages knowing that the welfare programme would make up the difference. The implicit 100 per cent tax rate applied to earnings in pre-1969 welfare programmes made this a quite realistic concern.

With an NIT, both ‘working poor’ families and families without breadwinners can be given equivalent levels of support in a framework that enhances work incentives for those previously on welfare and retains substantial incentives for the working poor to continue working and seeking better wages. The exclusion of the working poor can also induce breadwinners to abandon their families (or appear to) so that they will be eligible for welfare benefits. A negative tax available to households with and without breadwinners effectively removes that temptation.

In addition to the positive work incentives that are provided by the wage rate net of the fractional tax rate, $(1 - r)W$, it is possible to make entitlement conditional on some sort of work test. Such provisions are typical in welfare or other transfer programmes for persons who are able-bodied and not otherwise fully occupied with schooling or care of dependent family members. This is one feature that does not have a direct parallel in positive taxation, although it is interesting to consider making personal exemptions or standard deductions conditional on productive activity.

There are, of course, weaknesses in the NIT concept which are often only the obverse of its strengths depending on the point of view. The notion of treating all groups of poor the same and with a fixed and rigid benefit formula seems very retrogressive to many in the ‘helping professions’ who are trained to be sensitive to different needs and to fashion individualized therapies. Similarly, the replacement of in-kind benefits, whether goods or services, by a cash benefit may yield an efficiency dividend for the recipient, but the donor or typical taxpayer may not be content to allow recipients to allocate aid freely. Moreover, groups that supply the in-kind benefits (food producers, housing contractors, public employees who implement the programmes, etc.) quite predictably identify a national interest in in-kind benefits.

Two somewhat technical features, both important in determining the impact of an NIT must be mentioned at this point. One is the income concept and the other is the accounting period over which income flows are to be measured. Most NIT plans count money earnings from all sources, but there are differences with regard to imputing income to owned housing and to other assets that may not earn current money income. Some NIT proposals arbitrarily count part of net assets above some level as available for current expenditure regardless of their liquidity or earning rate.

The accounting period traditional for ordinary income taxes is a year, with some allowance for carry forward or carry back of income or losses that partially extends the accounting period. For welfare programmes the traditional accounting period in the 1960s was a month, and benefits were usually based on caseworker projections rather than on *ex post* income measures. For an NIT to be a plausible substitute for welfare benefits, it is necessary for the payments to be responsive to short-run variations in income. At the same time it seems important to give equal treatment to units with the same average income over a year or more regardless of how stable their income stream. Practical compromises have been found that use carry-over rules to rectify short-term overpayments. Integrated linear tax and transfer systems can handle this problem as a simple extension of current withholding policies for those with regular employment.

Of course, no programme can be as simple and complete as the NIT appears to be at first encounter. By the time a legislative committee has carefully considered and specified the tax unit, income concept, accounting period and administrative mechanism, it is possible that the plan will be feasible, but it will certainly be more difficult to implement than Friedman's vision implied. Overstimulated expectations also damage the appeal of the NIT when people realize that other programmes will still be needed for persons on the margin of competence to get along outside institutions or that other measures may be necessary to reinforce or enforce financial responsibility of parents for children. No one programme, and certainly not a basic cash transfer programme, can relieve all social ills, and some of the initial NIT enthusiasts neglected to mention the problems that remain once a minimum of spending power is assured.

Although no negative income tax has been fully adopted so far, the concept has had important impacts on both income transfer policies and on how these policies are analysed. Besides the theoretical and conventional empirical evaluations of the NIT ideas, controlled field experimentation was used to establish better estimates of the effects of alternative negative tax rates and guarantees on labour supply behaviour. These experiments demonstrated that the NIT can be implemented and administered at relatively low cost. The elasticity of labour supply of primary wage-earners generally turned out to be modest for variations within the range studied. Larger elasticities were observed for secondary earners, but in no case did reduced earnings substantially dilute the income enhancement which is, of course, the main reason for making transfers to the poor (Watts and Rees 1977; Robins et al. 1980).

The Family Assistance Plan first proposed by the Nixon Administration in 1969 and the Program for Better Jobs and Income proposed during Carter's term both incorporated major reforms of the welfare system that reflected NIT ideas. Both initiatives failed to win congressional approval. Smaller reforms have been more successful. There are a number of 'partial' or 'mini-' negative taxes in existence. In 1969 the AFDC programme adopted a formula that allowed recipients to keep the first $30 earned in a month and a third of any additional earnings. Although the implied 67 per cent tax rate may seem high, it was a major change from the 100 per cent rate previously in effect. (President Reagan eliminated this feature, and the response has been a small but distinct reduction in paid work for welfare recipients.) The Food Stamp programme now operates like a low-benefit negative tax that pays benefits in pseudo-money that can be spent only for food, but usually without binding constraint. Even the working poor are eligible for Food Stamp benefits! Supplemental Security Income operates as a Federal NIT with a 50 per cent tax rate for persons that are aged, blind or disabled (the former 'adult' categories of public assistance). Although labour supply is not a major public issue for these groups, it is not at all uncommon for SSI recipients to have other income, including labour earnings. This innovation allows them to enjoy at least half of the fruits of their efforts.

In these various ways, the NIT as an innovative approach to income transfers has enjoyed at least a modest level of success. Many economists now find welfare reform and related analysis to be intellectually challenging. Because of the theoretical and empirical efforts inspired by the negative income tax, policy analysts now have better analytical tools and evidence with which to work, and it seems likely that future income transfer policies will continue to be heavily influenced by this branch of economic thought.

See Also

- ▶ Built-in Stabilizers
- ▶ Direct Taxes
- ▶ Public Finance
- ▶ Taxation of Income
- ▶ Transfer Payments

Bibliography

Friedman, M. 1962. *Capitalism and freedom*. Chicago: University of Chicago Press.

Green, C. 1967. *Negative taxes and the poverty problem*. Washington, DC: Brookings.

Lampman, R. 1965. Approaches to the reduction of poverty. *American Economic Review* 55: 521–529.

Rhys-Williams, Lady J. 1943. *Something to look forward to*. London: MacDonald.

Robins, P., R. Spiegelman, S. Weiner, and J. Bell. 1980. *A Guaranteed annual income: Evidence from a social experiment*. New York: Academic.

Tobin, J. 1965. Improving the economic status of the Negro. *Daedalus*, 889–895.

Tobin, J., J. Pechman, and P. Mieszkowski. 1967. Is a negative income tax practical? *Yale Law Journal* 77: 1–27.

Watts, H., and A. Rees (eds.). 1977. *The New Jersey income maintenance experiment*, vols. 2 and 3. New York: Academic.

Negative Quantities

F. Y. Edgeworth

Negative quantities occur in economics, as in other sciences, when a variable, passing through zero, becomes less than nothing, so that the addition thereof causes not augmentation but diminution. Most economic quantities are susceptible of this *change of sign.* Thus wealth, affected with the *minus* sign, becomes debt. The utility attending the consumption of wealth being taken as positive, the disutility of labour incurred by the production of wealth must be regarded as negative. Consumption is negative production. Jevons proposes to employ discommodity to signify any substance or action which is the opposite of *commodity,* that is to say, *anything which we desire to get rid of,* like ashes or sewage *(Theory,* 2nd edn, p. 63). Such an article may be said to have negative value. Among articles which have a negative value agents of production may occur. The loss attending the use of old-fashioned machinery and plant may be considered as a negative 'quasi-rent' (Marshall). It is conceivable that, capital becoming superabundant, borrowers would pay a 'negative interest', that is, receive a payment for safeguarding and keeping up the capital borrowed (Prof. Foxwell, 'The Social Aspect of Banking', *Journal of the Institute of Bankers,* vol. vii. p. 71, 1886). The practical limit to this class of payment would be soon attained. The payment which a waiter makes in order to be allowed to serve in a fashionable restaurant where there is a prospect of gratuities might be described as negative wages.

The geometrical representation of a negative quantity, by reversing the direction of a line, is common in mathematical economics. Thus Jevons *(Theory,* 2nd edn, p. 187) represents the disutility of labour by ordinates measured downwards, the utility of consumption being represented by ordinates measured upwards. Of course the pleasure which may attend initial stages of labour is to be measured in an opposite direction from fatigue. A beautiful example of this construction is given by Gossen.

[The philosophy of the subject is stated ably and authoritatively by Cournot in his *Revue Sommaire,* in a passage directed against Mr. H.-D. Macleod's peculiar use of negative quantities in economics.]

Neighbours and Neighbourhoods

Ingrid Gould Ellen

Keywords

Census data; Chicago School of Sociology; Employment; Externalities; Hedonic regression analysis; House prices; Human capital; Neighbourhood; Networks; Residential segregation; Social capital; Tipping; Zoning

JEL Classifications

D85

The concept of neighbourhood has long been a topic of popular discourse and a subject of academic interest. Despite this attention, there is little agreement on what the term 'neighbourhood' means. The *American Heritage Dictionary* (Pickett 2000) simply defines a neighbourhood as ‘a district or an area with distinctive characteristics’.

‘A district or an area’ is not very specific, and social scientists (outside of economics) have struggled for decades to define more precisely the geographic boundaries of neighbourhoods (Keller 1968). Beyond the fact that neighbourhoods are sub-jurisdictional units, characterized by some degree of social cohesion, there is no accepted standard. The report prepared by the National Commission on Neighborhoods (1979, p. 7) stated that ‘each neighborhood is what the inhabitants think it is’. Yet the evidence suggests that such subjective perceptions vary greatly (Keller 1968).

For economists, who generally focus on externalities when considering neighbourhoods, an individual’s neighbourhood should theoretically extend as far as the individuals or facilities that affect her satisfaction with the community (Segal 1979; Galster 1986). In practice, economists and other social scientists studying neighbourhoods in the United States typically use census tracts to proxy for neighbourhoods. Including between 2,500 and 8,000 people on average, census tracts are close in size to what most envision as a neighbourhood and have the practical advantage of supplying demographic and economic data from the decennial census. In Australia and Europe, census data are typically available at sub-jurisdictional levels, defined by electoral wards or postcodes, and in some cases, smaller enumeration or collection districts (Overman 2002; Bolster et al. 2004; Drever 2004). Increasingly, researchers in the United States and Europe are able to link individual census data and other national household surveys to geographic identifiers, and they are experimenting with smaller and more flexible neighbourhood definitions (Bolster et al. 2004; Ioannides 2004; Bayer et al. 2005).

As for the term ‘distinctive characteristics’, economists identify several types of goods or services delivered by neighbourhoods. First, neighbourhoods offer distinct physical amenities, ranging from the style and condition of local housing to the number and quality of local parks. Second, neighbourhoods embody a particular set of ‘neighbours’, who have a distribution of income, human capital, and racial characteristics. Third, neighbourhoods often approximate local public service delivery areas such as attendance zones for public elementary schools, which often vary significantly in performance, even within the same jurisdictions. Fourth, neighbourhoods provide accessibility to shopping and employment opportunities. Finally, economists increasingly view neighbourhoods as possessing a stock of social capital, or norms and networks that facilitate interaction and can help residents work together to address problems like crime (Glaeser 2000).

Social scientists have been preoccupied with the evolution and nature of neighbourhoods for decades. Modern academic discourse on neighbourhoods has its roots in the Chicago School of the 1920s. These University of Chicago sociologists hypothesized that cities naturally grow outward in a series of concentric rings. Through this growth, a neighbourhood life cycle emerges, from richer residents to poorer, as more affluent residents opt for newer, less dense and quieter areas (Park et al. 1925).

Economists came later to the study of neighbourhoods, also initially drawn by an interest in the transition of neighbourhoods from high to low income and from predominantly white to predominantly minority residents. Muth (1972) and Sweeney (1974) propose variations of the filtering model, which, similar to the Chicago School theory, posits that neighbourhoods decline because, as their housing ages and deteriorates, higher-income residents exit, opting for newer neighbourhoods with newer housing. Other economists focused instead on the role of racial or class preferences in driving neighbourhood change (Bailey 1959). In his simple, elegant model, Schelling (1971) shows that, if households care about the composition of their neighbours, then small changes in demographic make-up can lead to the rapid tipping of a neighbourhood from one group to another.

Another strand of economic literature examines the relationship between various neighbourhood attributes and housing prices, typically using hedonic regression analysis (Kain and Quigley 1970; Bartik and Smith 1987). Mills and Hamilton (1994) argue that economists have historically failed to identify the external effects of housing quality and neighbourhood conditions. But more recent research finds strong evidence that housing prices are lower in areas with higher crime, lower- quality schools, dilapidated housing and vacant lots, and fewer homeowners (Grieson and White 1989; Black 1999; Coulson et al. 2003; Schwartz et al. 2003, 2005). As for the impacts of racial composition, more recent papers find that a neighbourhood's housing prices are negatively correlated with the percentage of black residents (Yinger 1976; Kiel and Zabel 1996; Myers 2004).

Finally, following Wilson (1987), economists have more recently turned to the study of how neighbourhoods and social interactions in them influence resident behaviour and outcomes.

See Also

- ▶ Ghettoes
- ▶ Residential Segregation
- ▶ Spatial Mismatch Hypothesis
- ▶ Urban Housing Demand

Bibliography

Bailey, M. 1959. Note on the economics of residential zoning and urban renewal. *Land Economics* 35: 288–292.

Bartik, T., and K. Smith. 1987. Urban amenities and public policy. In *Handbook of regional and urban economics*, vol. 2, ed. E. Mills. Amsterdam: North-Holland.

Bayer, P., S. Ross, and G. Topa. 2005. *Place of work and place of residence: Informal hiring networks and labor market outcomes*, Working Paper 11019. Cambridge, MA: NBER.

Black, S. 1999. Do better schools matter? Parental valuation of elementary education. *Quarterly Journal of Economics* 114: 579–599.

Bolster, A., S. Burgess, R. Johnston, K. Jones, C. Propper, and R. Sarker. 2004. *Neighbourhoods, households, and income dynamics: A semi-parametric investigation of neighbourhood effects*, Research Discussion Paper 4611. London: Centre for Economic Policy.

Coulson, N., S.-J. Hwang, and S. Imai. 2003. The value of owner-occupation in neighborhoods. *Journal of Housing Research* 13(2): 153–174.

Drever, A. 2004. Separate spaced, separate outcomes? Neighbourhood impacts on minorities in Germany. *Urban Studies* 41: 1423–1439.

Galster, G. 1986. What is a neighborhood? *International Journal of Urban and Regional Research* 10: 243–263.

Glaeser, E. 2000. The future of urban research: Non-market interactions. *Brookings-Wharton Papers on Urban Affairs* 2000: 101–138.

Grieson, R., and J. White. 1989. The existence and capitalization of neighborhood externalities: A reassessment. *Journal of Urban Economics* 25: 68–76.

Ioannides, Y. 2004. Neighborhood income distributions. *Journal of Urban Economics* 56: 435–457.

Kain, J. 1968. Housing segregation, negro unemployment, and metropolitan decentralization. *Quarterly Journal of Economics* 82: 175–197.

Kain, J., and J. Quigley. 1970. Measuring the value of housing quality. *Journal of the American Statistical Association* 5: 532–548.

Katz, L., J. Kling, and J. Liebman. 2001. Moving to opportunity in Boston: Early results of a randomized mobility experiment. *Quarterly Journal of Economics* 116: 607–654.

Keller, S. 1968. *The urban neighborhood: A sociological perspective*. New York: Random House.

Kiel, K., and J. Zabel. 1996. House price differentials in U.S. cities: Household and neighborhood racial effects. *Journal of Housing Economics* 5: 143–165.

Mills, E., and B. Hamilton. 1994. *Urban economics*, 5th ed. New York: HarperCollins College Publishers.

Muth, R. 1972. A vintage model of the housing stock. *Regional Science Association Papers Proceedings* 30(2): 141–156.

Myers, C. 2004. Discrimination and neighborhood effects: Understanding racial differences in U.S. house prices. *Journal of Urban Economics* 56: 279–302.

National Commission on Neighborhoods. 1979. *People, building neighborhoods*. Final Report to the President and the Congress of the United States. Washington, DC: Government Printing Office.

Overman, H. 2002. Neighbourhood effects in large and small neighbourhoods. *Urban Studies* 39: 117–130.

Park, R., E. Burgess, and R. McKenzie (eds.). 1925. *The city*. Chicago: University of Chicago Press.

Pickett, J. (ed.). 2000. *The American Heritage® dictionary of the English language*, 4th ed. Boston: Houghton Mifflin Company.

Schelling, T. 1971. Dynamic models of segregation. *Journal of Mathematical Sociology* 1: 143–186.

Schwartz, A., S. Susin, and I. Voicu. 2003. Has falling crime driven New York City's real estate boom? *Journal of Housing Research* 14: 101–136.

Schwartz, A., I. Ellen, I. Voicu, and M. Schill. 2005. *The external effects of place-based, subsidized housing*, Working Paper. New York: Furman Center for Real Estate and Urban Policy, New York University.

Segal, D. 1979. Introduction. In *The economics of neighborhood*, ed. D. Segal. New York: Academic Press.

Sweeney, J. 1974. A commodity hierarchy model of the rental housing market. *Journal of Urban Economics* 1: 288–323.

Wilson, W. 1987. *The truly disadvantaged: The inner-city, the underclass and public policy*. Chicago: University of Chicago Press.

Yinger, J. 1976. Racial prejudice and racial residential segregation in an urban model. *Journal of Urban Economics* 3: 383–396.

Neisser, Hans Philipp (1895–1975)

Edward J. Nell

Born in Germany on 3 September 1895, Hans Neisser came to the United States after his dismissal from his post as Deputy Director of Research in the Institute of World Economics at Kiel University in 1933, a post which had followed a distinguished career as an economic adviser to the Weimer government. While at Kiel he wrote *Der Tauschwert Des Geldes*, an important contribution to monetary theory, which also led him to formulate his critique of the Walrasian system. In the United States he first became Professor of Monetary Theory at the University of Pensylvania, and then worked in the Office of Price Administration during World War II, finally joining the Graduate Faculty of the New School for Social Research in 1943, where he remained as Professor until his retirement in 1965, and as an active Emeritus until his death in 1975.

Neisser's interests were extraordinarily broad, and he made important contributions first in monetary theory and macroeconomics, where, already working on his own critique of Say's Law, he was one of the few who immediately understood the message of Keynes's *General Theory.* His practical work in the OPA led him to rethink oligopoly in the light of game theory. Although he had little formal training in mathematics, he developed mathematical models instinctively, developing original ideas not only in the above areas, but also in international trade, and in growth theory. Moreover, he was one of the early pioneers of econometrics, in which he collaborated with his most distinguished pupil, Franco Modigliani. In addition he wrote extensively on philosophy and the sociology of knowledge, and their relationship to economic method. His approach was always analytic and critical, but his almost legendary openmindedness enabled him to appreciate the contributions as well as the flaws in systems as diverse as the neo Classical, the Keynesian and the Marxist.

His main publications include *Der Tauschwert des Geldes*, (1927), mentioned approvingly by Keynes in the *Treatise; Some International Aspects of the Business Cycle* (1936); *National Income and International Trade*, (with Franco Modigliani, 1953); *On the Sociology of Knowledge* (1965). In addition he published technical articles in almost every major economic journal, together with a series of papers on methodological and socio/philosophical issues in *Social Research*, of which he was a contributing editor for many years.

Selected Works

1927. *Der Tauschwert des Geldes*. Kiel.

1936. *Some international aspects of the business cycle*. Philadelphia: University of Pennsylvania Press.

1953. (With F. Modigliani.) *National income and international trade*. Urbana: University of Illinois Press.

1965. *On the sociology of knowledge*. New York: James H. Heineman.

Nemchinov, Vasily Sergeevich (1894–1964)

M. C. Kaser

Keywords

Linear programming; Mathematical economics; Nemchinov, V. S.

JEL Classifications

B31

Born the son of a State Bank messenger in Grabovo, Russia, on 2 January 1894; died in

Moscow on 5 November 1964. Nemchinov graduated from the Moscow Commercial Institute between the February and October Revolutions of 1917, but joined the Communist Party only in 1940 on appointment as Director of the K.A. Timiryazev Agricultural Institute, the Statistics Faculty of which he had headed since 1928. He showed courage in prohibiting from his Institute the pseudo-genetics ('Michurinism') of T.D. Lysenko, but when at Stalin's instigation mainstream genetics were condemned in 1948 he was forced from the directorship. The Academy of Sciences (to which he had been elected in 1946) then made him chairman of its Council for the Study of Productive Resources, a post retained (with a chair at the party's Academy of Social Sciences) until his fatal illness. In 1958 he established the first group in the USSR to study mathematical economics (from 1963 the Central Economic Mathematical Institute) and was posthumously awarded a Lenin Prize for elaborating linear programming and economic modelling for the USSR.

The research embodied in Nemchinov (1926, 1928) was distorted to justify Stalin's coercion of the peasantry: his data on rural social stratification gave cover to 'liquidation of the kulaks as a class' (though Nemchinov had avoided the term 'kulak'); his measurement of absolute gross harvest (Nemchinov 1932) was used to extort deliveries from collective farms. As soon as Stalin died, Nemchinov campaigned for the publication of official statistics and for more sophisticated techniques to utilize them – cybernetics had been damned as a pseudo-science serving capitalist interests. His organization of experimental national and regional input–output tables led him to question the meaningfulness of administered pricing, and his last book (1962) sought, as his widow put it (Nemchinova 1985, pp. 202–21), 'a broad-based system of social valuations ... as a single, internally consistent set of values'.

Selected Works

1926. O statisticheskom izuchenii klassovogo rassloenniya derevni [On the statistical study of rural class stratification]. *Bulleten' Ural'skogo oblastnogo statisticheskogo upravleniya* [Bulletin of the Urals Regional Statistical Administration] 1. Reprinted in *Selected works*, vol. 1.

1928. Opyt kalssifikatsii krest'yanskikh khozyaistv [Experience from the classification of peasant households]. *Vestnikstatistiki* [Statistical bulletin] 1. Reprinted in *Selected works*, vol. 1.

1932. Vyborochnye izmereniya urozhainosti [Sampling measurement of yields]. *Narodnoe khozyaistvo SSSR* [National economy of the USSR] 5–6. Reprinted in *Selected works*, vol. 1.

1962. *Ekonomiko-matematicheskie metody i modeli* [Methods and models of mathematical economics]. Moscow: Sotsegiz. 2nd (posthumous) ed, 1965. Reprinted in *Selected works*, vol. 3.

1967–9. *Izbrannye proizvedeniya* [Selected works]. 6 vols. Moscow: Izdatel'stvoNauka.

Bibliography

Nemchinova, M.B. 1985. The scientific work of Vasily Sergeevich Nemchinov (on the 90th anniversary of his birth). *Matekon. Translations of Russian and East European Mathematical Economics* 21(2) (1984–5): 3–25; translation of an article in *Ekonomika i matematicheskie metody* [Economics and mathematical methods] 20(1) (1984).

'Neoclassical'

Tony Aspromourgos

Keywords

Cambridge School; Classical economics; Dobb, M. H.; Hedonistic psychology; Hobson, J. A.; Marginal productivity theory of distribution; Marginalist theory; Marshall, A.; Methodological individualism; Mitchell, W. C.; Neoclassical; Neoclassical economics; Neoclassical synthesis; Roll, E.; Subjective theory of value; Utilitarianism; Veblen, T

JEL Classifications

B13

The term 'neoclassical' was first used by Veblen (1900, pp. 242, 260–2, 265–8), in order to characterize Marshall and Marshallian economics. Veblen did not appeal to any similarity in theoretical structure between the economics of Marshall and classical economics in order to defend this novel designation. Rather, he perceived Marshall's Cambridge School to have a continuity with classical economics on the alleged basis of a common utilitarian approach and the common assumption of a hedonistic psychology. Derivative from Veblen's use, this meaning of the term subsequently gained some currency, particularly in the 1920s and 1930s; for example, in the writings of Wesley Mitchell, J.A. Hobson, Maurice Dobb and Eric Roll. It is evident that the emergence of this notion of Marshallian economics as a 'neoclassical' project also involved, at least in part, an acquiescence to Marshall's portrayal of his own economics as a continuation of the classical tradition, though Marshall's sense of the continuity is not really that perceived by Veblen. Keynes (1936, pp. 177–8) also employed the term, though in an idiosyncratic matter, derivative from his equally idiosyncratic notion of classical economics.

The use of the term with the meaning which became the accepted convention after the Second World War, extending it to embrace marginalist theory in general, can be traced to Hicks (1932, p. 84) and Stigler (1941, pp. 8, 13, 297). From what source they derived the term is not certain. It is highly unlikely that either of them coined it independently. Perhaps the likeliest source of Hicks's use is Dobb's article, published as it was in the London School of Economics' 'house journal', *Economica*. Following Hamilton (1923), Dobb (1924, p. 68) writes that 'neoclassical' is not an entirely inappropriate term to describe Marshallian economics, 'for what the Cambridge School has done is to divest Classical Political Economy of its more obvious crudities, to sever its connection with the philosophy of natural law, and to restate it in terms of the differential calculus. The line of descent is fairly direct from Smith, Malthus, and Ricardo'. Hicks's article, or Veblen, is the most likely source of Stigler's use. He refers to both of them. Hicks and Stigler were certainly more correct than Veblen in perceiving the unifying core of the marginalist theories to be, on the one hand, methodological individualism and on the other, the marginal productivity theory of distribution developed in connection with the subjective theory of value. However, neither of them offered any significant defence for their (then) implicit view that the writings of the classical economists also can be characterized in terms of this theoretical approach. Subsequently this characterization and the nomenclature for marginalism associated with it – has given way to a recognition of the sharp theoretical disjuncture between classical and marginalist economics. Stigler's use, albeit hesitant, was probably as influential as his book. The term first gained wide currency in the debates on capital and growth in the 1950s and 1960s. It was no doubt also popularized by the extensive use made of it in Samuelson's textbook. From the third edition, Samuelson (1955, p. vi) presents the book as setting forth a 'grand neoclassical synthesis'. (For a fuller account, see Aspromourgos 1986.)

The question may be raised whether the depiction of 'neoclassical economics' in the mid-20th century, understood as a characterization of the mainstream of the discipline, continues to represent an accurate picture of dominant beliefs within economics. Colander (2000), for example, has questioned this. But, even though the term was never sensible, the majority of the profession remains committed to the fundamental convictions which were at issue in those earlier capital and growth debates – in particular, the notion that competition brings about a tendency to full employment of resources (especially labour) and the marginal productivity theory of functional income distribution.

See Also

- ▶ Robinson Crusoe
- ▶ 'Supply and Demand'

Bibliography

Aspromourgos, T. 1986. On the origins of the term 'neoclassical'. *Cambridge Journal of Economics* 10: 265–270.

Colander, D. 2000. The death of neoclassical economics. *Journal of the History of Economic Thought* 22: 127–143.

Dobb, M. 1924. The entrepreneur myth. *Economica* 4: 66–81.

Hamilton, W.H. 1923. Vestigial economics. *New Republic*, 4 April.

Hicks, J.R. 1932. Marginal productivity and the principle of variation. *Economica* 12: 79–88.

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

Samuelson, P.A. 1955. *Economics: An introductory analysis*. 3rd ed. New York: McGraw-Hill.

Stigler, G.J. 1941. *Production and distribution theories*. New York: Macmillan.

Veblen, T.B. 1900. The preconceptions of economic science III. *Quarterly Journal of Economics* 14: 240–269.

Neoclassical Growth Theory

F. H. Hahn

Abstract

Neoclassical growth theory is mostly that of the equilibrium of a competitive economy through time. It stresses capital accumulation, population growth and technical progress. It distinguishes momentary equilibrium (when the capital stock, the working population and technical know-how are fixed) from long-run equilibrium (when none of these elements is given). Long-run equilibrium is not a sequence of momentary equilibria, since it embodies the rational expectations of agents. The theory has little to say about the 'animal spirits' that may determine an economy's potential growth rate, but provides a good base camp for sallies into the study of particular economies.

Keywords

Accumulation of capital; Animal spirits; Arrow, K. J.; Capital–labour ratio; Classical saving function; Cobb–Douglas functions; Convergence; Duality; Elasticity of substitution; Expectations; Factor–price frontier; Hahn, F. H.; Harrod, R. F.; Investment behaviour; Kaldor, N.; Keynes, J. M.; Knife-edge problem; Liquidity trap; Long-run equilibrium; Lucas, R.; Meade, J. E.; Modigliani, F.; Momentary equilibrium; Natural and warranted rates of growth; Neoclassical economics; Neoclassical growth theory; New macroeconomics; Overlapping generations; Population growth; Proportional savings assumption; Rational expectations equilibrium; Robinson, J. V.; Samuelson, P. A.; Savings; Solow, R.; Steady-state equilibrium; Technical progress; Technical progress function; Unemployment; von Neumann, J.; Warranted path; Wicksell effect

JEL Classifications

O4

Neoclassical growth theory is not a theory of history. In a sense it is not even a theory of growth. Its aim is to supply an element in an eventual understanding of certain important elements in growth and to provide a way of organizing one's thoughts on these matters. For instance, the question of whether technical progress is bound to be associated with unemployment cannot be decisively answered by the theory but it goes a long way in pinpointing those considerations on which an answer depends.

Most of the theory is that of the equilibrium of a competitive economy through time. In particular, attention is paid to the accumulation of capital goods, growth in population and technical progress. Two kinds of equilibria are distinguished. One is the short period or *momentary* equilibrium of the economy when the stock of capital goods, the working population and technical know how can be taken as fixed. The other is the *long-run* equilibrium when none of these three elements are taken as given. It is important to understand that while long-run equilibrium implies momentary equilibrium for all dates it is not the case that a sequence of momentary equilibria constitutes a long-run equilibrium. For the latter has the property that the actions of agents taken at a given date in the light of their expectations of events at subsequent dates are not regretted when these dates arrive. In other words, it is what we would now call a *rational expectations equilibrium.* Harrod (1939) called a path of an economy with this property the *warranted path***.**

In principle a warranted path (say of output or output per man) could be quite irregular. Indeed it could be cyclical (Lucas 1975). But except in very simple models such generality is intractable and most of the attention has been devoted to long-run equilibria which are *steady-state* or *quasi-stationary.* (If a variable $x(t)$ obeys the dynamic equation $x(t) = \mathrm{e}^{gt}x(0)$ then $\hat{x}(t) = x(t)\mathrm{e}{-}^{gt} = x(0)$ is a constant, that is x is stationary.) This is one of the reasons why the theory is not really a theory of growth. It is also unwise to identify the steady state – say, the steady state rate or growth in output per head – with historical trends in the variable. That would require a good deal more argument than the theories provide. A steady state equilibrium is simply an extension of stationary equilibrium (an equilibrium in which the stock of capital goods, the population and technical knowledge are all constant). But it allows this now to include accumulation and technical change.

It is of interest to ask whether a steady state equilibrium is possible and if it is, whether a sequence of short period equilibria guides the economy to it. There is also another qst: do all warranted paths eventually become steady states? (See Hahn 1987) However the literature on these matters is sometimes confused and confusing. Short period equilibrium plainly depends on agents' expectations and so if they are not postulated to be always correct there are many possible evolutions of such equilibria. In fact except for Harrod's (1939) pioneering discussion of *actual* growth paths and one or two others, little attention has been paid to the expectational problem. Instead the path of the economy has been studied on the hypothesis that what is saved is also invested without explicit attention to what this implies for expectations concerning prices and interest rates. When that is made explicit it turns out that only warranted paths have been examined and not a sequence of short period equilibria. This procedure has been also adopted by the 'new macroeconomics' (e.g. Lucas 1975).

Connected with this is the treatment of investment and savings. The latter are usually taken to be either proportional to income or to come only from profits. Savings are not explained by the optimizing choices of households. This, however, is against the spirit of neoclassical economics. In order to improve on conventional savings theory one either takes a world which one can study 'as if' agents were infinitely long lived or one considers an economy of *overlapping generations* first studied by Samuelson (1958). Neither of these moves is discussed in what follows. But I re-emphasize that until savings behaviour has been explained the theories are not fully neoclassical.

Investment behaviour is a more difficult matter. Since the bulk of the theory is one of the warranted path, the marginal return to any investor is always equal to the marginal cost of investment. Thus investment is never regretted and is simply explained by it not being profitable to undertake more or less investment than is thus warranted. But difficulties arise if the warranted path and particularly the steady state is not unique, and also if investment is in some sense the carrier of technical progress. 'Animal spirits', as Keynes called entrepreneurial investment propensities, may be determinants of the rate of growth which the economy is capable of. Equally important is the circumstance that investment behaviour will be of prime importance in the evolution of a sequence of short run equilibria. Neoclassical theory has little to offer on these matters and is open to criticism on these grounds.

This brings me back to the beginning. As will be seen from what follows neoclassical theory states quite precisely what kind of economy in what kind of state is being considered. This economy and this state may be considered to be of low descriptive power. That, however, needs empirical argument and neither proponents nor opponents have produced any clinching ones. But an equally interesting question is whether the theory provides a good base camp for sallies into the study of particular economies. For instance, does it allow us to find just that feature of such an economy which is at variance with the postulates of the theory and thence to a modification of the latter, step by step? To this question at the moment the answer must be yes.

There is one last matter. The theories here discussed have provided the arena for much controversy concerning the *logical* coherence of neoclassical theory in general (Robinson 1965; Harcourt 1969). This controversy is not here discussed. For what it is worth it is this writer's view that neoclassical theory has survived this controversy unscathed. But the emphasis here is on 'logical'. There is little to be said for those economists who have taken the question of the descriptive merit of the theory as having been decisively settled in its favour.

The Simple Model

The Single Good Economy: No Technical Progress

Consider an economy in which a single good is produced by means of itself and labour. The good can also be consumed. The stock of it devoted to production is denoted by K and called capital. The stock does not depreciate either through use or the passage of time. Further notation is as follows: Y is output, L is the amount of labour used in production, L^0 is the labour force, $y = Y/L$, $k = K/L$, $e = L/L^0$.

Assumption 1 The production possibilities of the economy can be represented by a C^2 production function.

$$Y = F(K, L)$$

with the following properties:

(a) For all $h > 0 : hY = F(hK, hL)$. (Constant Returns to Scale)

(b) $f'(k) > 0, f''(k) < 0$ for $k \in [0, \infty]$. Also $f'(0) = \infty, f'(\infty) = 0$

(The 'Inada Conditions'; see Inada 1963).

From these assumptions it follows that we may represent the production possibilities by

$$y = f(k).$$

Assumption 2 The working population L^0 grows at a constant geometric rate λ[i.e. $L^0(t) = L^0(0)e^{\lambda t}$].

Assumption 3 A constant fraction s of output is not consumed.

It will thus be a condition of equilibrium that output which is not consumed is invested:

$$sf[k(t)] = sy(t) = \frac{\dot{K}(t)}{L(t)} = \dot{k}(t) + k(t)\frac{\dot{L}(t)}{L(t)}. \quad (1)$$

Definition 1 The economy is said to be in steady state equilibrium if $k(t)$ and $e(t)$ are constants, profits are maximized and (1) holds.

If $e(t)$ is constant then

$$\frac{\dot{L}(t)}{L(t)} = \frac{\dot{L}^0(t)}{L^0(t)} = \lambda.$$

Using this and the condition $\dot{k}(t) = 0$ in (1) yields

$$\lambda = \frac{sf(k)}{k} \quad (2)$$

as a condition for steady state equilibrium. Harrod (1939) called $sf(k)/k$ the *warranted rate of growth* and we shall abbreviate by writing

$$\frac{sf(k)}{k} \equiv w(k).$$

Clearly $w(k)$ gives us the rate of growth of output required to keep investment and savings equal to each other in steady state. On the other hand, λ is the rate of growth of employment which is needed to keep the proportion employed (possibly = unity) constant. Harrod called it the *natural rate of growth* of output for it tells us the rate at which output grows at a constant e.

Now by Assumption 1(b) one has $w(0) > \lambda$ and $w(\infty) < \lambda$ so there exists k^* satisfying (2). Since $(w'k) < 0$ everywhere, k^* is the only value of the capital labour ratio satisfying 2. But then for profit maximization, the real wage w^* and the real interest rate, ρ in steady state equilibrium are:

$$w^* = f(k^*) - k^* f'(k^*) \text{and } \rho^* = f'(k^*). \quad (3)$$

So the steady state equilibrium exists and is uniquely characterized by (3) and

$$\lambda = w(k^*) \quad (4)$$

Now return to (1) and consider the path $k(t)$ out of steady state but with $e(t)$ constant at e. In our new notation we find

$$\frac{\dot{k}}{k} = [w(k) - \lambda] \quad (5)$$

by dividing (1) by k and rearranging. Now let

$$V(k) = \frac{1}{2}[w(k) - \lambda]^2$$

so that *V(k)* is a measure of the deviation of the warranted from the natural rate of growth. One has:

$$V(k) \geq 0 \text{ all } k \text{ and } V(k^*) = 0. \quad (6)$$

Also using (5):

$$\begin{aligned}\dot{V}(k) &= [w(k) - \lambda]w'(k)\dot{k} \\ &= [w(k) - \lambda]^2 k w'(k) < 0 \quad \text{all } k > 0 \text{ and } k^* \neq k.\end{aligned} \quad (7)$$

These two results together with the Inada conditions suffice for the conclusion:

$$\text{For all } k(0) \geq 0, \quad \lim_{t \to \infty} k(t) = k^*.$$

We sum up:

Proposition P.1 An economy satisfying Assumption 1–3 has the following properties:

(a) There exists a unique steady state equilibrium
(b) The path of the economy along which savings are always equal to investment and the proportion of the workforce employed is constant (e is constant) approaches the steady state equilibrium as $t \to \infty$.

Discussion of the Model

There are many lacunae in the theory just presented and we shall be able to fill in some of these below. But first I discuss what can be learned from it.

Harrod (1939) writing in a Keynesian spirit held the view that a steady state equilibrium might not exist. He was particularly interested in the possibility that the warranted growth rate was always above the natural rate. In that case output would have to grow faster than is physically possible in order for investment to take up the savings generated and that is not possible. There would be a permanent tendency to depression. For many commentators this view of Harrod's rested implicitly on an assumed production function of the form:

$$Y = \min[aK, bL] \quad (8)$$

that is on fixed coefficients of production (see e.g. Solow 1956). However, a careful reading of Harrod suggests that he rather based his argument on the Keynesian liquidity trap. That is he thought that monetary forces set a positive lower bound on the rate of interest which thus on neoclassical theory set an upper bound on k and so, given s, a lower bound on $w(k)$.

This argument, however, is suspect. It is the real and not the nominal interest rate which governs (together with the real wage) the choice of k. Liquidity preference may set a lower bound on the nominal interest rate (the cost of holding money) but not on the real rate. Thus suppose r is the nominal interest rate. Then

$$\rho = r - \frac{\dot{p}}{p}$$

where p is the price of the good. Then if r is at its minimum level $\underline{r}$ we have from (3)

$$\left(\frac{\dot{p}}{p}\right)^* = \underline{r} - f'(k^*) \quad (9)$$

as a condition of steady state equilibrium. By assumption $f'(k^*) < \underline{r}$ so for such an equilibrium one requires a constant inflation rate:

$$\left(\frac{\dot{p}}{p}\right)^{*} > 0.$$

So provided we can graft a monetary sector onto the simple model it would seem that the liquidity trap is not an obstacle to the existence of steady state equilibrium.

But this argument reveals a central weakness in the reasoning which supports Proposition 1(b). For suppose at a historically given k one has $(wk) > \lambda$. If we *impose* the condition that savings are equal to investment, then indeed there would be pressure on resources and one could tell a story to explain the generation of the required inflation rate of (8). But we have no good reason for imposing that condition. By doing so we are not really asking: what actually happens?, that is, what is the actual growth rate?, but rather we are implicitly postulating that the inflation rate is always such that excess savings for k constant are taken up by capital deepening $\left(\dot{k} > 0\right)$. But why should this be so? If, for instance, the economy grew at λ then there would be excess supply of the good and normal arguments would lead us to suspect falling prices. But these would raise the real rate of interest and raise $w(k)$ above λ even further. The steady state equilibrium even if it exists is an unstable 'knife-edge' (Harrod 1939).

(b) Solow's celebrated paper (1956) established Proposition 1. But Solow was mistaken in his belief that it disposed of Harrod's knife-edge. The latter does not deal with paths on which the condition: savings = investment at a constant e has been imposed. That is did not postulate that the actual path was an equilibrium path. In this he was right since there is no good explanation of the Solow condition.

(c) An alternative procedure leading to Proposition 1(a) even if 8 is the form of the production function is to drop Assumption 3 (Hahn 1951; Kaldor 1955; Robinson 1965). This is done by supposing that the saving ratio out of profits is higher than that out of wages. Now if there are fixed coefficients of production (8) the equilibrium conditions (3) have no meaning since marginal products are not defined. This leaves it open to determine the real wage and interest rate by the requirement that they should generate that distribution of income between wages and profits which makes the warranted growth rate equal to the natural rate. From (8) one finds

$$\frac{Y}{K} = a,\ \frac{Y}{L} = b \ \text{ and } \ k = \frac{b}{a} \equiv \beta \ \text{ say}.$$

Let s_0 be the saving propensity out of wages and s_1 the saving propensity out of profits, with $s_0 < s_1$. Then the aggregate saving propensity, s, of the economy is given by

$$\frac{s_1 \rho}{a} + s_0 \frac{w}{b} = s.$$

Imposing the condition $sa = \lambda$ (the warranted rate = natural rate) yields

$$s_1 \rho + s_0 \frac{w}{\beta} = \lambda. \tag{10}$$

But also

$$\frac{\rho}{a} + \frac{w}{\beta} = 1 \tag{11}$$

so that we have two equations to determine what w^* and ρ^* must be in steady state equilibrium. A special case arises when $s_0 = 0$ (no saving out of wages) and $s_1 = 1$ (no consumption out of profits). Then

$$\rho^* = \lambda \tag{12}$$

is the condition of equilibrium. The reader should avoid interpreting (12) as saying that λ 'determines' the rate of profit. Equation 12 tells us what ρ must be if there is to be steady state equilibrium.

Once again a version of Proposition 1-(a) survives. Also stability fares slightly better than in (a). For if the actual growth rate is less than the warranted rate (because w and ρ have the 'wrong' values), and the latter is greater than λ then investment will be less than savings and competition between firms may lead to lower prices, higher real wages and so a fall in s. This will lower the warranted rate and bring it closer to λ as well as reducing the investment-savings gap. This *may* be so but what has just been said is not a proof. Indeed, as for instance Meade (1966) has shown, falling profitability may reduce the willingness to invest and so lead the system away from steady state equilibrium.

(d) Of course, (8) is not a plausible production function. Suppose we combine the savings assumption of (c) with a neoclassical production function satisfying Assumption 1. Then certainly (14) must hold in equilibrium. But (13) will now read

$$(s_1 + s_0)f'(k) + s_0\frac{f(k)}{k} = \lambda \tag{13}$$

from which we can find k^*. (Since

$$s_1\rho\frac{K}{Y} + s_0\frac{wL}{Y} = s.$$

So

$$s_1\rho + s_0\frac{w}{k} = s\frac{Y}{K} = \lambda.$$

Then substitute from (14) for ρ and w. So while the saving hypothesis will be reflected in the steady state value of k it will leave the equality between marginal productivity and factor rewards as an equilibrium condition. Indeed without this, the steady state values of w and ρ would be unknown. This is so even under the 'classical' savings assumption that $s_0 = 0$. The equation derived from (13) is then

$$s_1f'(k) = \lambda$$

and it tells us what k must be in order to generate a profit rate which, given the savings hypothesis, generates just the right amount of savings required for a growth in the capital stock at the rate λ. Thus the savings hypothesis has no direct bearing on the neoclassical equilibrium condition that the rate of profit must equal the marginal product of k.

(e) If workers save and invest their savings at the current rate of return on capital then the foregoing arithmetic needs to be changed. This was first noticed by Pasinetti (1962) whose paper gave rise to a number of others (Meade and Hahn 1965; Modigliani and Samuelson 1966).

Let $\sigma = s_1 - s_0 > 0$ Let μ be the fraction of k owned by capitalists – that is by agents who have no income from work. Then savings per employed worker are given by

$$s_0f(k) + \sigma f'(k)\mu k.$$

So in steady state equilibrium one requires

$$\frac{s_0f(k)}{k} + \sigma f'(k)\mu = \lambda. \tag{14}$$

From which

$$\frac{\mu f'(k)k}{f(k)} = \frac{1}{\sigma}\left[\frac{\lambda k}{f(k)} - s_0\right]. \tag{15}$$

The left-hand side measures the capitalists' share in income which cannot be negative. But there is nothing which guarantees a solution to (15) with $\lambda k \geq s_0f(k)$. Pasinetti (1962) simply made the latter (with strict inequality) a condition of the model. But God may have made the world otherwise.

In fact there are two possibilities. Suppose (15) has an admissible solution. One notes that in steady state one must have

$$1 - \mu = \frac{s_0[f(k) - \mu kf'(k)]}{\lambda k}. \tag{16}$$

That is the ratio of workers' capital to total capital must equal the ratio of their savings to total savings which in steady state equilibrium is equal to λk. Solving (16) for μ yields.

$$\frac{\lambda k - s_0f(k)}{k[\lambda - s_0f'(k)]} = \mu. \tag{17}$$

Solving (14) for μ yields

$$\left[\frac{\lambda k - s_0 f(k)}{k}\right]\frac{1}{\sigma f'(k)} = \mu. \tag{18}$$

Equating (17) to (18) then yields

$$s_1 f'(k) = \lambda. \tag{19}$$

So even though workers save, the long run equilibrium rate of profit bears the same relation to λ as it does under the classical savings hypothesis. Note that $\lambda k > s_0 f(k)$ is here required as before. In particular write (18) as

$$\max\left[0, \frac{\lambda k - s_0 f(k)}{k}\right]\frac{1}{\sigma f'(k)} = \mu. \tag{20}$$

Then this always has an admissible solution. If that gives $\mu = 0$ then from (14)

$$\frac{s_0 f(k)}{k} = \lambda \tag{21}$$

Harrod solution. It should now be emphasized that $\mu = 0$ does *not* mean that capitalists own no capital. All it means is that their share in total capital is zero.

Modigliani and Samuelson (1966) have shown how a warranted growth path may converge to k^* given by (12) or to k^{**} given by (21) depending on the technology and savings propensities.

(f) It will have been noticed that the whole of the above discussion has been conducted for L/L^0 constant and not $L/L^0 = 1$; that is the steady state is consistent with permanent unemployment. This should cause no surprise since the assumption of constant returns to scale and of constant savings propensities makes all equilibrium conditions independent of scale. if there is unemployment in a steady state equilibrium it can be argued with equal lack of real sense that either the capital stock is too low or that the real wage is too high. The present model is not suited to a discussion of whether falling interest rates and or money wages as long as there is unemployment would lead the economy to a steady state with full employment.

The Single Good Economy with Technical Progress

Growth theory without technical progress seems pretty useless. Yet no really satisfactory account exists of the determinants of technical progress, at least no such account based solely on considerations of economic theory exists. (Schumpeter (1934) is probably still the most interesting attempt but it excludes the possibility of steady state equilibrium.) What follows is therefore rather ad hoc and mechanical.

Technical progress shifts the production function through time and so in its most general form when technical progress is *disembodied*, one writes

$$Y(t) = F[K(t), L(t), t] \tag{22}$$

and retains the assumption of constant returns to scale for each t. Progress is disembodied if it can be taken full advantage of by the stock of the good (capital) accumulated in the past and by the same kind of labour. Even with this strong assumption we need more structure to build a model and accordingly postulate that all technical progress is *factor-augmenting,* that is (22) can be written as

$$Y(t) = F[\alpha(t)K(t), \beta(t)L(t)]$$
$$\text{with} \quad \alpha(t) \geq 0, \beta(t) \geq 0 \quad \text{all} \quad t.$$

Let

$$\hat{K}(t) = \alpha(t)K(t), \quad \hat{L}(t) = \beta(t)L(t)$$

and

$$\hat{k}(t) = \frac{\hat{K}(t)}{\hat{L}(t)}, \hat{y}(t) = \frac{Y(t)}{L^{\wedge}(t)}.$$

Then the equilibrium real interest rate is given by $\alpha f'[\hat{k}(t)]$ when $\hat{y}(t) = f[\hat{k}(t)]$.

In steady state equilibrium the real interest rate is constant. Let the operator E applied to a function $g(x)$ denote its elasticity

$$\left[Eg(x) = \frac{g'(x)}{g(x)}\right]x.$$

Then for the real interest rate to be constant one requires:

$$\frac{\dot{\alpha}}{\alpha} + \{Ef'[\hat{k}(t)]\}\left[\frac{\dot{\alpha}}{\alpha} - \frac{\dot{\beta}}{\beta} + \frac{\dot{k}}{k}\right] = 0. \tag{23}$$

Suppose first that $\alpha(0) = \beta(0) = 1$ and that $\dot{\alpha}(t) = 0$ all t, $\dot{\beta}(t) = b\beta(t)$ all t. Technical progress is purely labour augmenting (at a constant rate) or *Harrod- Neutral.* Clearly $\beta(t) = e^{bt}$. Hence (23) will be satisfied if

$$\frac{\dot{k}(t)}{k(t)} - b = 0 \text{ or } \frac{\dot{K}}{K} = b + \lambda. \tag{24}$$

Let $n = b + \lambda$ and call it the *natural rate of growth.* If savings are proportional to income, equilibrium requires

$$\frac{\alpha(0)sf[\hat{k}(t)]}{k^\wedge(t)} = n \tag{25}$$

which can be uniquely solved for $\hat{k}^*$ when the production function is concave and satisfies the Inada conditions. By (24), $\dot{k}(t) = 0$ and so we conclude that (i) the capital output ratio and the real interest rate are both constant and (ii) the real wage and the capital labour ratio (k) are rising at the rate b. But the wage per efficiency unit of labour and capital per efficiency unit of labour are both constant. Hence we are essentially in the same situation as that discussed for the absence of technical progress.

Next suppose that $\dot{\alpha}(t) = a\alpha(t)$ and $a = b$. Technical progress is said to be *Hicks-neutral.* Then (23) becomes

$$a + \{Ef'[k(t)]\}\frac{\dot{k}}{k} = 0. \tag{26}$$

Suppose that the production function is characterized by an elasticity of substitution equal to minus one. Then since with Hicks-neutrality one can write:

$Y = e^{bt}F[K(t), L(t)]$one has that KF_K/F is constant when K is changed but F is constant (if one is moving along an isoquant). This implies

$$Ef'[\hat{k}(t)] = -1$$

and so once again (using (25) one obtains (24). A constant rate of profit and a constant share of profits then implies a constant capital output ratio. In other words, Harrod-neutrality is equivalent to Hicks-neutrality with a unit elasticity of substitution (Robinson 1938). Uzawa (1961) has shown that only a Cobb–Douglas production function will give this equivalence.

If $a \neq b$ technical progress is 'biased' in favour of the higher of a and b. However, there is no fundamental reason why technical progress should be of the factor-augmenting type nor, if it is, why it should proceed at a steady rate. Hence technical progress makes the idea of steady state equilibrium somewhat unconvincing.

However, there have been attempts to formulate a theory which focuses on endogenous economic forces that may cause technical progress to be of a certain kind (Kennedy 1964; Samuelson 1965). These attempts are not notably successful or convincing and will only be sketched.

Given a factor-augmenting production function which exhibits constant returns to scale, one can write the minimum unit cost function as

$$c = c[q(t)/\alpha(t), w(t)/\beta(t)]$$

where $q(t)$ is the rental of capital of $w(t)$ the wage. Let s_K and s_L respectively be the shares in unit cost of capital and labour. Then from elementary Duality Theory (e.g. Varian 1978), if $\dot{w}(t) = \dot{q}(t) = 0$:

$$\frac{\dot{c}}{c} = -[s_K a(t) + s_L b(t)] \tag{27}$$

where $b(t) = \dot{\beta}(t)/\beta(t), a(t) = \dot{\alpha}(t)/\alpha(t)$. The idea now is as follows. Firms can choose to 'produce' $a(t)$ and $b(t)$ according to a 'production possibility' function.

$$T[a(t), b(t)] = g[b(t)] - a(t) \geq 0 \tag{28}$$

and the pairs (a, b) satisfying (28) form a convex compact set with a differentiable boundary. Also $g'(b) < 0$. If the firm's objective is to minimize

$\dot{c}/c$ subject to (28) it will choose $b(t)$ so as to satisfy

$$-g'[b(t)] = \frac{s_L}{s_K}. \tag{29}$$

As Samuelson (1965) has noted, (29) is *not* some novel theory of income distribution unrelated to the Neo-classical one. The latter was needed in the definition of c and the derivation of (27).

Now s_L/s_K will depend on the relative prices of efficiency units. Since $g(\cdot)$ is monotone (28) can be inverted:

$$b(t) = (g')^{-1}(s_L/s_k)$$

and so we write

$$b(t) = h\left[\frac{w(t)}{q(t)}\frac{\alpha(t)}{\beta(t)}\right]. \tag{30}$$

The Eqs. 28 and 30 are two differential equations in *a(t),* $\beta(t)$ and relative factor prices. It is easy to show that

$$h'(1-\sigma) \geq 0$$

where σ is the elasticity of substitution.

If one can take w/q constant then one proceeds as follows.

$$\begin{aligned} b(t) - a(t) &= \frac{\mathrm{d}\log[\beta(t)/\alpha(t)]}{\mathrm{d}t} = b(t) - g[b(t)] \\ &= v[b(t)] \text{ say.} \end{aligned}$$

Substituting from (30) one obtains the differential equation

$$\frac{\mathrm{d}\log[\beta(t)/\alpha(t)]}{\mathrm{d}t} = v\left\{h\left[\frac{w}{q}\frac{\alpha(t)}{\beta(t)}\right]\right\}. \tag{31}$$

This equation gives the evolution of relative factor augmentation. If for some $[\alpha/\beta]^*$ one has a critical point of v and (31) is convergent then there will be a constant relative rate of labour augmentation so $b(t) - a(t) \to 0$. (This does not necessarily imply that $b(t)$ and $a(t)$ become constant.) In that situation innovations are derived to be Hicks-Neutral. Even if the rate of innovation is then constant we know that this will not be consistent with steady state unless the elasticity of substitution is unity. But Samuelson (1965) has shown that the stipulated convergence of (31) requires an elasticity of substitution which is less than one in absolute value.

All of this is on the assumption w/q = constant. In fact we know from our earlier discussion that w/q will depend on $\widehat{k}(t)$ so we can replace the r.h.s. of (31) by:

$$v^*\left[k(t)\frac{\alpha(t)}{\beta(t)}\right].$$

We then need a differential equation for the evolution of *k(t)* which we can obtain from the appropriate warranted growth path.

Samuelson (1965) has studied the case: $\dot{k}(t) = 0$. The literature can be consulted for further detail. At this level of aggregation the story is hardly persuasive nor can much be said in favour of the objective function which has been stipulated. On the other hand, all of this is a considerable advance on meaningless claims like: 'high wages induce labour-saving innovation' first exposed by Fellner (1961). After all, the marginal return per unit cost of the factor is the same for all factors in equilibrium. None the less one must conclude that the theory of induced innovations and their relations to growth have a long way to go yet.

The One Sector Model with Embodied Technical Progress

In this section two related ideas are considered. The first is that capital and labour are substitutable *ex ante* ('putty') before investment has been congealed in concrete machines but it is not substitutable *ex post* ('clay') once the investment has been made. The second is that technical progress does not benefit old machines; it is embodied in the latest machines. These two ideas are related but can be combined in various ways. Thus one can have embodied technical progress with (traditional) putty–putty (Solow 1970) or with clay–clay (Solow et al. 1967). One can also have disembodied technical progress as in the previous section with putty–clay. The main lessons are perhaps best learned by combining embodied technical progress with putty–clay. The classic reference here is Bliss (1968).

Some of the technicalities of the analysis now called for are somewhat involved and what follows is more in the nature of a summary of the economic implications.

An investment undertaken at date θ gives rise to machines of vintage θ. If at that date the investment is $I(\theta)$ and employment is $L(\theta, \theta)$, output per man is $y(\theta, \theta)$ and given by

$$y(\theta,\theta) = \mathrm{e}^{a\theta} f(k(\theta)) \ \text{ where } \ k(\theta) \\ = I(\theta)/L(\theta,\theta)\mathrm{e}^{a\theta}.$$

Let $f(\cdot)$ satisfy Assumption1.1. The output per man on vintage θ at date $t \geq \theta$ is written as $y(t, \theta)$. It is assumed that as long as output is produced on vintage θ that

$$y(t,\theta) = y(\theta,\theta) \tag{32}$$

This departs somewhat from the 'clay' assumption. It will be noticed that Harrod- neutral technical progress has been assumed. It can be shown (Bliss 1968) that this is necessary for a steady state equilibrium to exist.

Any firm in this technological environment will make its investment and employment decisions in the light of long term expectations. For once machines have been installed they no longer share in technical progress yet the latter will raise real wages and reduce quasi-rents on old machines. These will be scrapped when quasi-rents have fallen to zero so that the economic life of the machines is endogenous to the economic process. The economic life is relevant to the investment decision and hence expectations of the course of real wages are relevant. In the theory it is assumed that all expectations are always correct. None of these considerations apply to the case of disembodied technical progress with putty-putty.

If $w(t)$ is the real wage at t then if $y(t, \theta) - w(t) > 0$ it will pay the firm to set $L(t, \theta) = L(\theta, \theta)$ because of (32). It will set $L(t, \theta) = 0$ when$y(t, \theta) - w(t) = 0$. These conditions determine the economic life of a machine. It is easy to show that if T is the economic life of a machine that it must be constant in steady state equilibrium. The value of T is determined by the condition$w(t) = y(t - T, t - T)$, that is, the wage equals its average product on the last vintage in use. When that is the case the firm is indifferent whether it employs labour on that vintage or not. If it does employ some then if the economy had a little more or less labour it would be employment on the last vintage in use which is varied and so $w(t)$ would measure labour's marginal social product. If no labour is employed of the last vintage then a small reduction in labour would mean reducing employment on the next oldest vintage. If there is a continuum of vintages then the economy would still lose just $y\,(t - T, t - T)$.

Now let $n = a + \lambda$ as in (1). We are looking for a steady state equilibrium as before in which output and investment grow at the rate n because gross savings are proportional to income. As before also the ratio of capital to labour measured in efficiency units of the latest vintage (i.e. $k(\theta)$) should be constant. So if $Y(t)$ is aggregate output at t and $Y(\theta, \theta)$ total output with capital of vintage θ we have

$$\begin{aligned} Y(t) &= \int_{t-T}^{t} y(\theta,\theta)L(\theta,\theta)\mathrm{d}\theta \\ &= \int_{t-T}^{t} Y(\theta,\theta)\mathrm{d}\theta \\ &= \frac{\mathrm{e}^{nt}Y(t-T,t-T)(1-e^{-nT})}{n}. \end{aligned} \tag{33}$$

If $I(\theta)$ is investment at θ then $I(t) = \mathrm{e}^{nt}I(t - T)$ and that must equal $sY(t)$. So using (33) and writing $v = Y(t - T)/I\,(t - T)$ we obtain

$$sv = \frac{n}{1 - \mathrm{e}^{-nT}}. \tag{34}$$

The left-hand side of (34) is again Harrod's warranted growth rate. But the rate at which the economy is capable of expanding indefinitely now depends on T, the economic life of equipment and that is an economic variable and *not* a parameter like n. One must, of course, show that (34) has a solution. If as in Solow et al. (1967) the technology is clay–clay then v is given as fixed. Profit maximization together with the condition that the present value of quasi-rents equals the cost of the investment which gives rise to them at the scrapping, fix the equilibrium value of T. It is then possible that Harrod's view that (34) has no solution is valid. This is a fortiori true if the solution of (34) requires $s > 1$.

One can show that the real interest rate (= profit rate) must be constant in steady state equilibrium (see Bliss 1968). However, the relation between the latter and the equilibrium value of T is not straightforward and depends on the elasticity of substitution. That is because in steady state the scrapping condition is $t = 1/a$ log (inverse of share of wages in vintage $(t - T)$) and the share will depend on the elasticity of substitution. One can also show that if a steady state exists that the warranted growth path of the economy will approach the steady state. This is even the case with clay–clay.

All in all the simple neoclassical model survives 'the bolting down' of concrete machines and embodied technical progress rather well. That does not mean that the resulting model is satisfactorily 'realistic'. What it does mean is that the theory is a good deal more robust than critics once thought it to be. This is also illustrated by the following episode in the related theory of technical progress.

Kaldor took the view that it was not possible to distinguish between finding another 'page in the book of blueprints' (Robinson 1965), i.e. movements along the production function and finding a new page, i.e. innovations. He proposed that all that could be observed was a relation between the rate of growth in labour productivity and investment per man. This relation he called the 'technical progress function' and justified by the view that every act of investment led to learning. He and Mirrlees (1962) constructed a model on this basis. However, except for the assumption that firms required investment 'to pay for itself' in a predetermined period, the results of the model were not notably different from the ones already discussed. (A linear technical progress function can be integrated into a Cobb–Douglas production function. A non-linear one of the right shape has the advantage of making steady state equilibrium investment be at the rate at which the capital output ratio is constant, i.e. Harrod-neutrality is a consequence and not a hypothesis of the model.)

Arrow (1962) kept the production function (he uses clay–clay) but made technical improvement depend on the total investment undertaken over the past. This was again justified by learning. The steady state again is one of Harrod-neutral progress which is explained endogenously. There are now obvious external benefits from investment but otherwise the 'learning by doing' steady state equilibrium is of the kind we have already discussed.

Two Sector Growth Models

One considers an economy with a consumption good and an investment good sector. This was first proposed by Uzawa (1961) and then gave rise to a very large literature (e.g. Solow 1962; Inada 1963; Takayama 1963). We shall discuss only the case where both sectors have 'well behaved' constant returns to scale production functions, capital does not depreciate and there is no technical progress. For the latter see Diamond (1965).

Steady State

It is well known (e.g. Samuelson 1957; Mirrlees 1969) that given these assumptions, the equilibrium relative prices of the two goods are determined once ρ (the real interest rate) is determined. So with a classical saving hypothesis we know that steady state requires:

$$\rho = \lambda$$

and so q the price of the investment good in terms of the consumption good can be written as $q(\lambda)$. If w is the wage in terms of consumption good, y_c is output per man employed in the consumption good sector and $\mu = L_c/L$ is the proportion of the labour force employed in that sector, the classical savings assumption yields the equilibrium condition

$$w = y_c\mu \;\; or \;\; \mu = w/y_c. \qquad (35)$$

(Demand for consumption good equals supply.) But w/y_c is a unique function of ρ. For by profit maximization the marginal product of capital in the consumption sector must equal $\rho q = \lambda q(\lambda)$. So λ determines a unique capital/labour ratio and so a unique share of wages in the consumption sector. Hence we can write $\mu = \mu(\lambda)$. If k is the overall capital labour ratio, k_c and k_I the capital/labour ratios in the consumption and investment sectors respectively then $k = \mu k_c + (1 - \mu)k_I$ It is plain that k is uniquely determined by λ.

Matters are somewhat more complicated with a proportional saving function and we shall not derive all the results in full. Let v be the capital output ratio *in value terms*. In steady state, as usual, we require$s = v\lambda$. The question now is whether putting $v = s/\lambda$ uniquely determines k, k_c k_I and hence the rate of profit and real wage. The answer is: no.

Let ψ be the wage rental ratio. A rise in that ratio will lower q if the consumption goods sector is more labour intensive than the investment goods sector. Hence k_c and k_I will be raised and v will be lowered. But the value of investment output is a constant fraction s of the value of output and q is lower so that output of investment good must rise relatively to that of consumption good and so μ, must be lower $(1 - \mu$ is higher). Hence k will be higher (since $k_I > k_c$) and this will tend to increase v. It follows that v can have the same value at different k's and ψ's. This is really the story of what Professor Robinson (1965) called the Wicksell effect. To get uniqueness one needs the not very persuasive assumption: $k_c > k_I$ always, or some assumption on the elasticities of substitution (Takayama 1963).

Stability

The question may be asked whether a sequence of short period equilibria of the economy starting with an arbitrary $k(0)$ at time $t = 0$ lead the economy to steady state equilibrium.

At any moment of time k is given from the past. A short period equilibrium is a division of the capital stock and of the labour between the two sectors such that at the resulting prices all markets clear and profits are maximised. The resulting investment good output will augment the capital stock. At the next moment there will also be more labour so we know the new value of k. So given $k(0)$ it looks as if we could deduce $k(t)$ for all $t > 0$ and so study the convergence to steady state.

But this is only true if momentary equilibrium is unique. If it is not then there will be a variety of paths the system can follow and we do not know which it will be. More seriously in this case we may have, say, there equilibria for some k and only one for another k'. In that case at the point at which we 'lose' equilibria there is a 'catastrophe' (in the technical sense). For this see Inada (1963).

Now consider the proportional savings assumption. It says that consumption and investment are proportional to *aggregate income*, that is, the distribution of income has no effect on the demand for either good. But this is just the case for which non-intersecting community indifference maps exist (see Gorman 1953) and in that case momentary equilibrium must be unique: it is given by the tangency of the transformation curve between investment and consumption good and the indifference curve. So in this case momentary equilibrium is unique.

But this is not true for the classical saving function where it is clear that demand does depend on the distribution of income so that in general no community indifference maps exist and there may be multiple momentary equilibria. Once again more detailed assumptions concerning elasticities of substitution or $k_c > k_I$ can rescue the situation. They really amount to the postulate of a certain kind of gross- substitutability (Hahn 1965).

Once uniqueness of momentary equilibrium is assured it is not hard to show that the sequence of momentary equilibria approach the steady state (see Hahn and Matthews 1964, for an intuitive account). For instance, for a classical saving postulate, $k(0)$ must be inversely related to $\psi(k(0))$, the wage rental ratio. So if $k*$ is the steady state capital labour ratio, $\rho(k(0)) < \rho(k^*)$ whenever $k(0) > k*$. But $\rho[k(0)] = K/K$ while $\rho(k^*) = \lambda$ hence

$$\frac{\dot{k}}{k} = \rho[k(0)] - \rho(k^*) < 0$$

and $k(0)$ in declining at $t = 0$. In fact the reader can check that $[k(t) - k^*]^2$ is always declining with t as long as $k(t) \neq k^*$which suffices here to establish convergence to the steady state value k*.

On the other hand, it should be noted that this argument is very much at risk when there is a variety of capital goods (see Hagemann 1987).

Technical Progress

With two sectors the nature of technological change in the economy as a whole will clearly depend on what kind of progress occurs in each of the sectors and on the composition of output. For instance, if by Harrod neutrality we mean that the capital/ output ratio in value terms is constant when the rate of profit is constant we need to know how the capital/output ratio in each of the sectors is changing as well as what is happening to the relative outputs of the two sectors.

The case of disembodied technical progress is fully analysed in Diamond (1965) while there seems to be no literature on two-sector embodied technical progress.

As an example consider steady state with a proportional savings function. The value share of investment in output must remain constant. Technical progress in the investment sector will have to be Harrod-neutral because the rate of profit equality with the marginal product of capital is there independent of relative prices (input and output are the same). So in steady state the marginal product of capital should remain constant. If the capital labour ratio in both sectors remains constant then technical progress in the consumption goods sector must also be Harrod-neutral. Differences in the rate of technical progress in the two sectors will be reflected in a changing price of consumption good in terms of investment good. However, there could be steady state equilibrium with the labour allocation between the two sectors changing. In that case in general technical progress in the consumption good sector will not be Harrod-neutral.

It is not profitable to go into greater detail.

Many Sectors

As long as one is only concerned with steady state equilibrium there is no difficulty for neoclassical theory when there are many sectors. Although it was somewhat special the foundations for the study of this case were laid by von Neumann (1945). (He assumed labour to be in infinitely elastic supply (in fact producible) at a given vector input of consumption goods. He also considered a ‘spectrum’ of techniques.) More recent formulations are best studied in Morishima (1964). For a survey see Hahn and Matthews (1964).

The essentials of this case can be illustrated for a classical savings function with only intermediate goods used in production (i.e. no long lived inputs) and no joint production.

Suppose there are N produced goods and one non-produced good (e.g. labour). Production takes time. Let q be the price vector of the N produced goods in terms of the non-produced good. Let all inputs be paid for when purchased and let $c(q)$ be the minimum unit cost function in terms of labour. That is $c(q)$ is the unit cost of production when inputs have been chosen to minimise costs. We can write it in this way because constant returns prevail everywhere. If that were not so there would be no hope of finding a steady state equilibrium.

In such an equilibrium if all goods are produced and relative prices are constant it must be that

$$q = (1 + \rho)c(q). \qquad (36)$$

If the economy is productive and indecomposable and every good needs labour in its production then one can solve (35) uniquely for $q(p) \gg 0$ provided ρ lies in some bounded interval. The function $q(\rho)$ is the *factor-price frontier*.

It is easy to prove that

$$\frac{\partial q_j}{\partial_\rho} > 0. \qquad (37)$$

Provided that the ratio in which wage earners consume goods depends only on q and not on their level of income one can now complete the story. The solution $q(\rho)$ is plainly independent of the scale or composition of output. So one can always make demand equal to supply in each sector provided there is enough labour in the economy. Suppose that labour is inelastically supplied. Then the scale of output can be anything. But if the ratio of employed to unemployed is to remain constant then output must grow at the rate λ hence so must investment and we get $\rho = \lambda$ as a further equilibrium condition. Relative prices will then be given by $q(\lambda)$. In equilibrium the present value of an input's marginal product will equal its price. Moreover ρ can be shown to measure the increase, at constant prices, in consumption made possible tomorrow if there is a little less consumption today and resources saved thereby are allocated efficiently.

An alternative scenario is to suppose that labour can always be had at a constant real wage w^* where the real wage is written as some function of q, say, $w(q)$. Then $w^* = w(q)$ together with (36) determine both q^* and ρ^* for steady state equilibrium. Given that there are classical savings the economy will grow at the rate ρ^* which will in fact be the highest (balanced) rate of growth the economy is capable of.

Perhaps a more general insight into these models can be gained as follows. Let Y and X be two n-vectors where the latter is the input of goods at one date and Y the output resulting at the subsequent date. Let L be the labour input. Then

$$T(Y, X, L) \geq 0 \tag{38}$$

is the economy's transformation locus which is homogeneous of degree one in its argument. Now a perfectly competitive economy is production efficient. So if all goods are produced in the steady state $(Y^*/L^*, X^*/L^*)$ there must be prices q^* and profit rate p^* such that

$$q^*Y^* - (1+\rho^*)[q^*X^* + L^*] = 0 \tag{39}$$

is a supporting hyperplane of the set of (Y, X, L) satisfying (38) at (λ^*, X^*, L^*) Net output is $q^*(Y^* - X^*)$. If there are proportional savings at the rate s then one requires

$$sq^*(Y^* - X^*) = \lambda(q^*X^*) \tag{40}$$

if employment is to grow at the rate λ and Y/L and X/L are constant. But that is just the Harrod equation.

Now

$$q^*Y^* - (1+\rho^*)[q^*X^* + L^*] \geq q^*Y - (1+\rho^*)[q^*X + L] \tag{41}$$

for all (Y, X, L) satisfying (38. Hence (39) is the maximum value of the r.h.s. of (41) subject to (38). Hence if T is differentiable:

$$q_i^* = \frac{T_{X_i}}{T_L} = -(1+\rho^*)\frac{T_{Y_i}}{T_L} \tag{42}$$

as can be verified by carrying out the maximization. Write (38) as

$$T(Y, kX, L) \geq 0 \tag{43}$$

take $k = 1$ and differentiate with respect to k at (Y^*, X^*, L^*) to get

$$\left[\sum T_{Y_i}\frac{\mathrm{d}Y_i}{\mathrm{d}k} + \sum T_{X_i}X_i\right]\mathrm{d}k = 0. \tag{44}$$

Substitute from (42) into (44) writing

$$\Delta y_i = \frac{\mathrm{d}Y_i}{\mathrm{d}k}\mathrm{d}k, \Delta x_i = X_i\mathrm{d}k,$$

to obtain

$$\sum q_i^*\Delta y_i = (1+\rho^*)\sum q_i^*\Delta x_i$$

or

$$\frac{\sum q_i^*\Delta y_i - \sum q_i^*\Delta x_i}{\sum q_i^*\Delta x_i} = \rho^* \tag{45}$$

Hence the equilibrium rate of profit measures the increase in the value of net output at equilibrium prices as a fraction of the increase in the value of inputs at equilibrium prices. Or the rate of substitution between present and future consumption bundles of constant composition, evaluated at q^*. Of course, there is no sense to the claim that (45) 'determines' ρ^*.

The literature on growth theory is vast and this essay can usefully be supplemented by other accounts such as Meade (1962), Hahn and Matthews (1964), and Solow (1970).

See Also

- ▶ Classical Growth Model
- ▶ Neoclassical Growth Theory (New Perspectives)
- ▶ Ramsey Model
- ▶ Two-Sector Models
- ▶ von Neumann, John (1903–1957)

Bibliography

Arrow, K.J. 1962. The economic implications of learning by doing. *Review of Economic Studies* 28 (3): 155–173.

Bliss, C.J. 1968. On putty-clay. *Review of Economic Studies* 35 (2): 105–132.

Diamond, P. 1965. Disembodied technical change in a two-sector model. *Review of Economic Studies* 32 (2): 161–168.

Fellner, W. 1961. Two propositions in the theory of induced innovations. *Economic Journal* 71: 305–308.

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

Hagemann, H. 1987. Capital goods. In *The New Palgrave: A Dictionary of Economics*, ed. J. Eatwell, M. Milgate, and P. Newman, vol. 1. London: Macmillan.

Hahn, F.H. 1951. The share of wages in national income. *Oxford Economic Papers* 3 (2): 149–157.

Hahn, F.H. 1965. On two sector growth models. *Review of Economic Studies* 32 (4): 339–346.

Hahn, F.H. 1987. Hahn problem. In *The New Palgrave: A Dictionary of Economics*, ed. J. Eatwell, M. Milgate, and P. Newman, vol. 2. London: Macmillan.

Hahn, F.H., and R.C.O. Matthews. 1964. The theory of economic growth: A survey. *Economic Journal* 74: 779–902. Reprinted in *Surveys of Economic Theory*, vol. 2. London: Macmillan 1965.

Harcourt, G.C. 1969. Some Cambridge controversies in the theory of capital. *Journal of Economic Literature* 7 (2): 369–405.

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

Inada, K. 1963. On a two-sector model of economic growth: Comments and a generalisation. *Review of Economic Studies* 30: 119–127.

Inada, K. 1964. On the stability of growth equilibrium in two-sector models. *Review of Economic Studies* 31 (2): 127–142.

Kaldor, N. 1955. Alternative theories of distribution. *Economic Journal* 23 (2): 83–100.

Kaldor, N., and J. Mirrlees. 1962. A new model of economic growth. *Review of Economic Studies* 29 (3): 174–192.

Kennedy, C. 1964. Induced bias in innovation and the theory of distribution. *Economic Journal* 74: 841–847.

Lucas, R. 1975. An equilibrium model of the trade cycle. *Journal of Political Economy* 83: 1113–1144.

Meade, J.E. 1962. *A neoclassical theory of economic growth*. London: Allen & Unwin.

Meade, J.E. 1966. The outcome of the Pasinetti process: A note. *Economic Journal* 76: 161–165.

Meade, J.E., and F.H. Hahn. 1965. The rate of profit in a growing economy. *Economic Journal* 75: 445–448.

Mirrlees, J.A. 1969. The dynamic non-substitution th. *Review of Economic Studies* 36 (1): 67–76.

Modigliani, F., and P.A. Samuelson. 1966. The Pasinetti Paradox in neo-classical and more general models. *Review of Economic Studies* 33: 269–301.

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

Pasinetti, L.L. 1962. Rate of profit and income distribution in relation to the rate of economic growth. *Review of Economic Studies* 29 (4): 267–279.

Robinson, J.V. 1938. The classification of inventions. *Review of Economic Studies* 5 (2): 139–142.

Robinson, J.V. 1965. *The accumulation of capital*. 2nd ed. London: Macmillan.

Samuelson, P.A. 1957. Wages and interest: A modern dissection of Marxian economic models. *American Economic Review* 47: 884–912.

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.

Samuelson, P.A. 1965. A theory of induced innovations along Kennedy-Weizsacker lines. *The Review of Economics and Statistics* 47: 343–356.

Schumpeter, J.A. 1934. *The theory of economic development*. Cambridge, MA: Harvard University Press.

Solow, R.M. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70 (1): 65–94.

Solow, R.M. 1962. Comment (on Uzawa 1961). *Review of Economic Studies* 29 (3): 255–257.

Solow, R.M. 1970. *Growth theory: An exposition*. Oxford: Clarendon Press.

Solow, R.M., J. Tobin, C.C. von Weizsacker, and M. Yaari. 1967. Neo-classical growth with fixed proportions. *Review of Economic Studies* 33 (2): 79–115.

Takayama, A. 1963. On a two-sector model of economic growth: A comparative statics analysis. *Review of Economic Studies* 36: 95–104.

Uzawa, H. 1961. On a two-sector model of economic growth. *Review of Economic Studies* 29 (1): 40–47.

Varian, H. 1978. *Micro-economic analysis*. New York: W.W. Norton.

von Neumann, J. 1945. A model of general economic equilibrium. *Review of Economic Studies* 13: 1–9.

Neoclassical Growth Theory (New Perspectives)

Rodolfo E. Manuelli

Abstract

The neoclassical growth model captures the basic trade-off between saving and investment. It has proven to be a useful tool to study development paths, and the interactions of technology shocks, money and fertility choices with growth.

Keywords

Competitive equilibrium; Convexity; Endogenous growth; Fertility; Human capital accumulation; Infinite horizons; Innovation; Marginal rate of transformation; Neoclassical growth theory; Optimal development paths; Optimal quantity of money; Optimal taxation; Population growth; Recursive equilibrium; Representative agent; Solow–Swan growth model; Technical change; Technology shocks; Transversality condition; Turnpike property

JEL Classifications

O4

This article complements neoclassical growth theory. It discusses some developments of the neoclassical growth theory that endogenize the saving rates.

Infinite Horizons

The Planning Problem

The standard neoclassical growth model assumes that the planning horizon is infinite. One justification is that forward-looking parents act 'as if' they were to live forever. To see this, assume that each individual lives for one period and has exactly one descendant. The utility of a member of generation 0 is given by

$$U_0 = u(c_0) + \beta U_1, \tag{1}$$

where u is an increasing, continuous and concave function of consumption at time t, c_t. Iterating on this expression yields

$$U_0 = \sum_{t=0}^{\infty} \beta^t u(c_t), \beta = \frac{1}{1+\rho}, \rho > 0, \tag{2}$$

which shows that altruism implies that the effective planning horizon for each individual is infinite.

In the simplest one-sector version of the model, the technology is summarized by

$$c_t + x_t \leq zf(k_t), t = 0, 1, \ldots \tag{3a}$$

$$k_{t+1} \leq (1 - \delta_k)k_t + x_t, t = 0, 1, \ldots \tag{3b}$$

$$k_0 > 0, given, \tag{3c}$$

where k_t is the stock of capital per person available at the beginning of period t, x_t is gross investment, z is a measure of productivity, and δ_k is the depreciation rate of capital. The function f is assumed to be increasing, continuous and strictly concave.

The planning problem corresponds to the maximization of the utility criterion (2), subject to the feasibility constraints (3). The analysis of this problem was initially carried out by Ramsey (1928), Cass (1965) and Koopmans (1965). A thorough analysis of the model can be found in Stokey and Lucas (1989).

The model has sharp predictions for the properties of an optimal development path. The relevant first-order conditions (in the interior case) require that the marginal rate of substitution between consumption at time t and $t + 1$ equal the marginal rate of transformation,

$$\begin{aligned} \frac{u(c_t)}{\beta u(c_{t+1})} &= 1 - \delta_k + zf'(k_{t+1}), t \\ &= 0, 1, \ldots, \end{aligned} \tag{4}$$

and a transversality condition which is naturally interpreted as requiring that the value, at time 0, of the stock of capital at time $T + 1$ converge to 0 as $T \rightarrow \infty$. Formally, the condition is

$$\lim_{T\to\infty} \beta^T u'(c_T)k_{T+1} = 0.$$

Some properties of the solution are as follows:

1. There exists a unique steady state; that is, there are constant sequences of consumption, investment and capital that satisfy (3) (except at time 0) and (4). From (4) it follows that, in the steady state, the marginal product of capital equals the sum of the discount rate, ρ, and the depreciation factor, δ_k,

$$\rho + \delta_k = zf'(k^*), \tag{5}$$

which determines capital per worker. The steady state level of consumption is given by

$$c^* = zf(k^*) - \delta_k k^*. \tag{6}$$

2. For any $k_0 > 0$, the solution to the problem converges to the steady state. Convergence is monotone.
3. In general, the savings rate – defined as $1 - c_t/zf(k_t)$ – is not constant, or even monotone. This distinguishes the optimal neoclassical growth model from the Solow–Swan version that assumes exogenous (and generally constant) saving rates.

The steady state is the model's prediction about the long-run levels of capital, consumption and investment. From the point of view of a theory of growth there are some interesting results:

1. The steady state level of output per worker is independent of the form of the utility function.
2. If a fixed level of government consumption, g, is introduced in the model, the steady state condition (5) remains unchanged. The new steady state level of consumption is $c^* = zf(k^*) - \delta_k k^* - g$. Thus the model predicts that, in the long run, permanent increases in government spending have no impact on output per worker, and they crowd out private consumption one for one, with no effect on investment.

The basic model has been extended in many dimensions. In the case of multiple sectors, existence of optimal paths has been established very generally. Burmeister (1980) provides conditions for the existence and uniqueness of steady states with many capital goods.

The properties of optimal paths depend on the specification of the economic environment. In the case of a discounted twice differentiable utility and dominance diagonal of a matrix of first-order conditions, it is possible to show that the turnpike property holds (see the excellent survey in McKenzie 1986). Formally, McKenzie shows that if $\{k_t\}$ is an optimal path starting from k_0, then, for every capital stock k_0' near k_0 the associated unique optimal path converges exponentially to $\{k_t\}$.

The monotonicity properties of optimal paths do not extend to the multicapital or multisector case. In general, optimal paths can display cycles (see Burmeister 1980) and even more complex behaviour.

To illustrate this let the feasible technology set be described as

$$c_t \le T(k_t, k_{t+1}),$$

and let the (indirect) utility function over capital stocks be

$$v(k_t, k_{t+1}) \equiv u(T(k_t, k_{t+1})).$$

With this notation, the planning problem reduces to

$$\max_{\{k_{t+1}\}} \sum_{t=0}^{\infty} \beta^t v(k_t, k_{t+1}).$$

Let's denote a candidate solution by a function g where

$$k_{t+1} = g(k_t).$$

Boldrin and Montrucchio (1986) showed that – under standard conditions – given any twice differentiable function g, there exists a pair (v, β) so that the associated planner's problem has g as its optimal policy function. Since g can exhibit arbitrary complex dynamics, the result shows that in order to endow the theory with predictive power it is necessary to 'force' the chosen specification to quantitatively match moments of the (actual) economy under study. Most recent research using the neoclassical growth model disciplines the choices of functional forms and parameters by requiring that they predict behaviour consistent with the empirical evidence.

Equilibrium Growth

Even though the analysis of the growth model was motivated by normative considerations, under the stated assumptions the planner's solution of the growth model coincides with the competitive equilibrium of the economy. The argument – using the traditional definition of a competitive equilibrium – follows from Debreu (1954). In macro applications – the field in which the model has proved to be most useful – it is more natural to define a competitive equilibrium using the notion of recursive equilibrium first introduced by Prescott and Mehra (1980).

In order to account for wages, let the production function be given by

$$y \le zF(k, n),$$

where F is concave and homogeneous of degree one, and it satisfies

$$f(k) \equiv F(k, 1).$$

Even though there are many alternative ways of defining an equilibrium, it is easiest to consider the case in which there are rental spot markets for capital and labour, and the households trade consumption, labour and capital services and one-period bonds. The problem solved by the representative household is

$$\max \sum_{t=0}^{\infty} \beta^t u(c_t)$$

subject to

$$\begin{aligned} b_{t+1} + c_t + x_t \le w_t n_t = q_t k_t + (1 + r_t) b_t \\ = 0, 1, \ldots k_{t+1} \le (1 - \delta_k) k_t + x_t, \\ t = 0, 1, \ldots 0 \le n_t \le 1, t = 0, 1, \ldots \end{aligned}$$

and the initial conditions, $[(1 + r_0)b_0, k_0]$, given. As stated, this problem has no solution since the budget set is unbounded. Different alternative assumptions on how to deal with debt at infinity have been used to guarantee that the problem is well defined. The most general specification is to rule out Ponzi games by imposing that the present value of debt be nonnegative. Formally, any solution must satisfy

$$\lim_{T \to \infty} \prod_{j=0}^{T} \frac{1}{1 + r_j} b_{T+1} \ge 0.$$

which is the analogue – in the market setting – of the transversality condition in the planning problem.

Firms solve a static problem

$$\max_{k_t n_t} zF(k_t, n_t) - q_t k_t - w_t n_t.$$

A competitive equilibrium is an allocation $[\{c_t\}, \{n_t\}, \{x_t\}, \{k_{t+1}\}]_{t=0}^{\infty}$, a price system $[\{q_t\}, \{w_t\}, \{r_{t+1}\}]_{t=0}^{\infty}$ and a sequence of bond holdings $\{b_{t+1}\}_{t=0}^{\infty}$ such that:

1. Given the price system, the allocation solves the maximization problems of households and firms.
2. Markets clear.

Given that Debreu (1954) shows that the solution to the planner's problem can be decentralized as a competitive equilibrium, the first-order conditions (on the assumption of interiority and differentiability) corresponding to the maximization of utility and profits imply that equilibrium prices (as a function of the planner's allocation) are given by

$$q_t = zf'(k_t), \tag{7a}$$

$$w_t = zf(k_t) - k_t zf'(k_t), \tag{7b}$$

$$r_{t+1} = q_{t+1} - \delta_k. \tag{7c}$$

It is possible to state the implications of the neoclassical growth model more intuitively using equilibrium prices. The consumer's optimal choice between consumption and saving requires that

$$\frac{u(c_t)}{\beta u(c_{t+1})} = 1 + r_{t+1},$$

that is, that the marginal rate of substitution between present and future consumption equal to (gross) interest rate. Optimality on the part of firms requires that the marginal product of labour be equal to the wage rate and that the marginal product of capital equal the cost of capital, $r_{t.} + \delta_k$.

The basic neoclassical growth model (and some of the extensions mentioned) has had a significant impact on how economists view the process of development and the role of markets supporting optimal development paths. It is clear that there is nothing special about dynamic problems that make it more (or less) likely for competitive markets to fail to deliver optimal allocations. In the basic model of this note, Theorems I and II of welfare economics apply.

Applications

Some of the most notable extensions are as follows.

Technology Shocks

Brock and Mirman (1972) studied a version of the neoclassical growth model in which the representative agent maximizes the expected value of the discounted flow of utility, and the technology is as in the deterministic growth model except that the technology level, z, is replaced by a stochastic process $\{z_t\}$. Brock and Mirman assumed that the process $\{z_t\}$ is i.i.d. They established the existence of a solution and they showed that, under standard concavity assumptions, the resulting stochastic process of the capital stock has a unique invariant measure, which is the stochastic analogue of the steady state in the deterministic version of the problem. They also showed that the optimal policy function which determines k_{t+1} as a function of k_t and z_t is monotone. The results were extended to the case of serially correlated shocks by Donaldson and Mehra (1983).

This research has provided the theoretical foundations for a large literature that analyses the impact of economic fluctuations on savings and growth. When the model is extended to include an elastic labour supply, this is a natural setting in which to study cyclical movements of employment. For an introduction to this literature see Cooley (1995).

Human Capital and Development

The neoclassical growth model, extended to allow for human capital accumulation, is a natural candidate to understand the role that technological differences play in accounting for differences in output per worker. In the standard specification – using a Cobb–Douglas specification for f – it follows that output per worker is given by

$$y = z^{1/(1-\alpha)}\overline{y}_0$$

where α corresponds to capital share, $\overline{y}_0$ and (and all the $\overline{y}_j$ in this section) is a constant. This version of the theory implies that the elasticity of output per worker with respect to z is $1/(1 - \alpha)$. Since accepted estimates of α cluster around 0.33 – which, approximately, correspond to the share of national income that accrues to capital – the elasticity is estimated to be approximately 1.5. If this model is to explain the differences in output per worker between the richest and poorest countries (which are of the order of 15–20 to 1), it must assume fairly large differences in productivity that exceed the best available estimates.

Klenow and Rodríguez-Claire (1997) (see also, Bils and Klenow 2000) consider a production function of the form

$$y = zk^{\alpha}\left(h^e\right)^{1-\alpha},$$

and they use the specification $h^e = e^{\psi s}$, where s corresponds to years of schooling to estimate the role of human capital. In this case, the equilibrium level of output per worker is given by

$$y = z^{1/(1-\alpha)}e^{\psi s}\overline{y}_1$$

Klenow and Rodríguez-Claire use data to determine s and ψ. To highlight the role of productivity differences, let $e^{\psi s} = z^{\nu}$. Output per worker is

$$y = z^{1/(1-\alpha)+\nu} e^{\psi s} \overline{y}_1.$$

Klenow and Rodríguez-Claire find that the implied ν is not large. They conclude that productivity differences account for much of the differences in output.

Manuelli and Seshadri (2007a) endogenize the human capital decision. They adopt Ben Porath's (1967) specification. In discrete time, their model assumes that human capital evolves according to

$$h_{t+1} = z_h (n_t h_t)^{\gamma 1} x_{ht}^{\gamma 2} + (1-\delta_h) h_t,$$

where $n_t h_t$ is the fraction of the available time allocated to producing human capital, and x_{ht} denotes market goods used in the production of human capital. In this setting, $h^e = (1 - n)h$. It is possible to show that, in the steady state, output per worker is given by

$$y = z^{\gamma_2/[(1-\alpha)(1-\gamma_1-\gamma_2)]} \overline{y}_2.$$

This version of the model implies that the elasticity of output with respect to the productivity parameter z is $\gamma_2/[(1 - \alpha)(1 - \gamma_1 - \gamma_2)]$. Manuelli and Seshadri use life age–earnings profile evidence to estimate that $\gamma_1 = 0.63$ and $\gamma_2 = 0.30$. This results in an elasticity of output per worker with respect to productivity of 6.5. This high elasticity implies that productivity differences have a large impact on (endogenously chosen) human capital. As a result, even small productivity differences are consistent with large variations in output per worker. The relative importance of human capital and productivity is an active area of research. More work is needed before the roles of technology and education in accounting for differences in output can be accurately estimated.

The Role of Taxation

The neoclassical growth model has been widely used to analyse the effect of specific tax policies and to derive properties of optimal tax systems.

Consider a version of the model in which labour is elastically supplied. Let the period utility function be given by $u(c, \ell)$, where ℓ is interpreted as leisure. In an economy in which consumption, capital income and labour income are taxed (at constant rates) it follows that the steady state is characterized by

$$\rho = \left(1 - \tau^k\right)\left(F_k(k,n) - \delta_k\right) \qquad (8a)$$

$$u_\ell(c, 1-n) = u_c(c, 1-n) F_n(k,n) \frac{1-\tau^n}{1+\tau^c} \qquad (8b)$$

$$F(k,n) = c + \delta_k k \qquad (8c)$$

$$\rho = \left(1 - \tau^b\right) r^b. \qquad (8d)$$

From a formal point of view the system of Eq. (8) contains four equations in four unknowns. Let $\Phi(c, \ell) = u_\ell(c, 1 - n)/u_c(c, 1 - n)$, and assume that $\Phi(c, \ell)$ is increasing in c and decreasing in ℓ. In this case, it is possible to show that:

1. An increase in the tax rate of capital income, τ^k, decreases the amount of capital, but has ambiguous effects on employment.
2. An increase in tax rate on labour income (consumption) decreases both k and n.

The effect of taxes on employment and growth is a subject that continues to receive substantial attention.

In the mid-1980s Chamley (1986) and Judd (1985) asked the following question: If a government has to finance a given (say, constant) stream of consumption, and if the only available taxes are distortionary taxes (for example, in the previous example, set $\tau^c = 0$ and add government spending to (8c)), how should those taxes be chosen? Chamley and Judd showed that the optimal tax system is such that, in the steady state, capital income taxes are zero while labour income taxes are positive.

This result is delicate in the sense that it does not hold if some of the assumptions are slightly modified. For example, if the function F is strictly concave, and pure profits cannot be taxed away, then the optimal long-run tax rate on capital income need not be zero. Similarly, if there are different types of labour (for example, high and low skill) and it is possible for the planner to distinguish between them, then the zero taxation result is overturned. For other examples see Correia (1996) and Jones et al. (1997).

Money and Growth

Since the neoclassical growth model satisfies the assumptions of the convex economy studied by Debreu (1959), it is impossible to find an equilibrium in which a non-interest earning asset (for example, money) has positive value in equilibrium. In order to introduce money, the neoclassical growth model has been modified in a variety of ways. One of the first attempts corresponds to Sidrauski's (1967) analysis of a monetary model. Sidrauski studied the case in which money enters the utility function, as a reduced form that captures the services provided by money balances. In Sidrauski's formulation (adapted to discrete time), the consumer problem is

$$\max \sum_{t=0}^{\infty} \beta^t u(c_t, m_{t+1}/p_t)$$

subject to

$$c_t + \frac{m_{t+1}}{p_t} + x_t + \frac{B_{t+1}}{p_t} \leq w_t + q_t k_t + \frac{m_t}{p_t} + \frac{(1+i_t)B_t}{p_t} + \frac{M_{t+1} - M_t}{p},$$

where m_t is nominal money balances chosen by the household, M_t is the economy-wide per capita money supply (that the individual takes as given), $p_{t.}$ is the price level, B_t is the nominal value of one period bonds purchased at time $t - 1$, and $(1 + i_t)$ is the gross nominal interest rate. The specification of the budget constraint reflects the assumption that the government exogenously increases the stock of money through lump-sum transfers.

The first order conditions for this problem are (imposing the standard equilibrium conditions)

$$u_1(c_t, m_{t+1}/p_t) = \lambda_t, \tag{9a}$$

$$u_2(c_t, m_{t+1}/p_t) = \lambda_t \frac{i_{t+1}}{1 + i_{t+1}}, \tag{9b}$$

$$\lambda_t = \beta \lambda_t [1 - \delta_k + z f'(k_{t+1})], \tag{9c}$$

and feasibility. In this version of the model, money is superneutral in the steady state. In the steady state Eq. (9c) reduces to Eq. (5a) and, hence, the rate of money growth has no impact on the long-run level of output. This result is not robust. If labour is supplied elastically, inflation has (in general) real effects through its impact on the marginal rate of substitution between real money balances and leisure. The one case in which money is still neutral is when the utility function is separable in real money balances (see Fischer 1979).

In an economy in which nominal money balances grow at the (gross) rate $1 + \pi$, the nominal interest rate is given by

$$1 + i = (1 + \rho)(1 + \pi),$$

and satisfies the Fisher equation. Friedman (1969) argued that since money is costless to produce, its optimal level should be such that individuals are satiated. This corresponds to $u_2(c_t\ ,\ m_{t+1}/p_t) = 0$. Inspection of Eq. (9b) shows that the optimal quantity of money requires that the nominal interest rate be 0. This can be implemented by engineering a deflation (that is, setting $1 + \pi = (1 + \rho)^{-1}$) or by keeping the price level constant and paying interest on money holdings.

In general, in the non-separable case, the Friedman rule needs to be modified (see Turnovsky and Brock 1980).

Fertility and Growth

The neoclassical growth model can be easily extended to the case of exogenous population growth and exogenous technical change. It has also been used to understand the interplay between economic forces and fertility decisions (see Barro and Becker 1989; Becker and Barro 1988).

To illustrate the relationship between growth and fertility, assume that individuals live for just one period and that each agent gives birth to η offspring. The utility function of a member of generation t is given by

$$U_t = u(c_t) + \beta\eta_t^{(1-\varphi)}U_{t+1}, 0 \le \varphi \le 1,$$

where η_t is the number of children. When $\varphi > 0$, these preferences display imperfect altruism as increases in the number of children result in lower marginal contribution of the last child to utility.

It is assumed that each child costs υ units of labour, and the per capita labour endowment is normalized to 1. The planner's problem for this economy can be expressed as

$$\max \sum_{t=0}^{\infty} \beta^t N_t u(c_t),$$

subject to

$$c_t + \eta_t(a + k_{t+1}) \le zF(k_t, 1 - \eta_t\upsilon) + (1 - \delta_k)k_t, k_0 > 0, N_{t+1} \le N_t\eta_t^{(1-\varphi)}, N_0 = 1$$

Thus, from a formal point of view, endogenous fertility plays the role of another good, N_t, which is 'produced' with a linear technology with current fertility as its only input. This is a special case of a two-sector model. Barro and Becker showed that if the utility function is of the form $u(c) = c^{\sigma}$ – a standard specification – the model can have multiple steady states, with some stable and some unstable.

The model has been used to study the effect of changes in child mortality on fertility (see Doepke 2005), the impact of introducing social security (see Boldrin and Jones 2005), and the relationship between fertility, growth and human capital (see Manuelli and Seshadri 2007b). In general, the ability of the model to match the evidence depends on the specific parameterization used, and finding the appropriate specification is an active area of research.

Finite Lifetimes

What are the properties of the neoclassical growth model if economic agents have short – relative to the economy – horizons? The simplest case is study an economy in which individuals live for two periods, and have preferences defined over first-and second-period consumption. This model was originally analysed by Diamond (1965), and an excellent textbook treatment can be found in Azariadis (1993).

Each agent inelastically offers one unit of labour in his first period, and $e \le 1$ units in his second period. The representative agent problem is

$$\max U\left(c_t^t, c_{t+1}^t\right)$$

subject to

$$c_t^t + (1 + r_{t+1})^{-1}c_{t+1}^t \le w_t + (1 + r_{t+1})^{-1}w^{t+1}e,$$

where c_t^j denotes consumption at time t of an individual born in period j, and w_t is the wage rate. Feasible allocations satisfy

$$c_t^j + c_t^{t-1} + x_t \le zF(k_t, 1 + e), k_{t+1} \le (1 - \delta_k)k_t + x_t,\ t = 0, 1, \ldots$$

where, as before, we assume that F is homogeneous of degree 1.

Since the solution to an individual optimization problem is completely summarized (in the two period setting) by its saving function, let

$$s_t = s(w_t, w_{t+1}, r_{t+1}) \tag{10}$$

denote saving by a member of generation t. Firms, as in the case of infinite horizons, are assumed to solve static problems. Equilibrium input prices, satisfy the appropriate version of (7).

An equilibrium in this economy consists of sequences of capital stocks and prices such that individuals and firms optimize and markets clear. A simple (and intuitive) condition that characterizes all the equilibria is the requirement that saving by the young at time t equal the capital stock at the beginning of period $t + 1$.

Formally, this corresponds to

$$k_{t+1} = s(\overline{w}(k_t), \overline{w}(k_{t+1}), \overline{r}(k_{t+1})), \tag{11}$$

where,

$$\overline{w}(k) \equiv zF_2(k, 1 + e), \overline{r}(k) = zF_1(k, 1 + e) - \delta_k.$$

For a given k_0, any sequence that satisfies (11) and that does not violate other feasibility conditions (for example, $k_t \geq 0$) is an equilibrium sequence of capital stocks. The other components of an equilibrium (for example, consumption and prices) can be readily obtained from the household and firm optimization problems.

Even though this set-up (with only one type of consumer) appears very close to the infinite horizon model, its implications are quite different. An (incomplete) list of the most interesting properties includes the following:

1. Even if $e = 0$ (young individuals are net savers), and if both consumption goods are normal, the equilibrium need not be unique. A sufficient condition for uniqueness is that the two goods be gross substitutes. This corresponds to the saving function being an increasing function of the interest rate.
2. If $e = 0$ and saving is increasing in the interest rate, Eq. (11) can be solved for k_{t+1}. Let the solution be denoted $k_{t+1} = G(k_t)$. Then, if $G'(0) > 1$, then this map can have and odd number $(2j + 1)$ of nontrivial steady states, of which $j + 1$ are asymptotically stable and j are unstable. If $G'(0) < 1$ there may be an even number of nontrivial steady states.
3. If $e = 0$ and saving is not increasing in the interest rate, Eq. (11) can be solved for k_{t+1} only locally. The major impact of this is that stable steady states need not be separated by unstable steady states.
4. Equilibrium paths of capital may display cycles and, depending on the specification, chaotic dynamics.
5. Equilibria – even stationary equilibria – need not be optimal.

This last result shows that when the individual horizon differs from the economy's horizon, then optimal saving at the individual level need not imply optimality in the aggregate, even in the absence of the standard arguments (for example, externalities) for market failure.

To illustrate what can go wrong, consider an economy in which U is strictly quasi-concave and that, in a stationary equilibrium, the stock of capital is such that $\overline{r}(\overline{k}) = zF_1(\overline{k}, 1) - \delta_k < 0$. Let the levels of consumption in young and old age be denoted $(\overline{c}_1, \overline{c}_2)$. The key condition is that the gross interest rate be less that the gross rate of population growth, which is assumed to be 1 in this example. Consider next the problem of maximizing the utility of a given generation subject to the constraint that allocations be constant and the stock of capital also remains constant. Let k^* be the solution to

$$\max U(c_1, c_2)$$

subject to

$$c_1 + c_2 \leq zF(k, 1) - \delta_k k.$$

Let the solution of this problem be (c_1^*, c_2^*, k^*). Given that k^* is such that $zF_1(k^*,1) - \delta_k = 0$, it follows that $k^* < \overline{k}$. Since $(\overline{c}_1, \overline{c}_2, \overline{k})$ is feasible, it must be the case that. $U(c_1^*, c_2^*) > U(\overline{c}_1, \overline{c}_2)$. Thus all generations, starting with generation 1, are better off under this alternative allocation. What about the initial old? Since they only care about consumption they are also better off as fewer resources are allocated to investment.

To summarize, when individual horizons are shorter than the economy's horizon, even the simplest specification of the neoclassical growth model can result in very complicated equilibrium paths.

Concluding Comments

For many years, the neoclassical growth model has been the workhorse of researchers interested in fluctuations and growth. The model is not without weaknesses. Perhaps the most important is its inability to explain long-run growth: in the steady state the growth rate is exogenous. Endogenous growth models – versions of which are very close to the neoclassical growth model – can be used to understand the effects of policies and shocks on long-run growth. Currently, there are isolated attempts to integrate both views. This has been done for versions of the models that assume convex technologies. For example, endogenous growth models have been used to eliminate the need for arbitrary detrending in the study of business fluctuations (see, for example, Jones et al. 2005). The versions of the models that have been studied so far are, of necessity, the simplest ones. It is too early to tell whether the integration of the two strands will succeed.

A large literature on endogenous growth departs from the assumption of convex technologies and no external effects. This body of research views innovation as a form of public good, and emphasizes the role of institutions (for example, how property rights are protected) in determining growth. Since these assumptions amount to departures from the convexity assumptions of the neoclassical model, competitive equilibria are no longer optimal, and this alternative view suggests that a variety of interventions are needed to attain optimality. Thus, the major difference relies on the presence (or absence) of departures from the assumption that technologies form a convex cone.

If the neoclassical growth model is narrowly interpreted (as in this article) as assuming that government policies are exogenous (and markets are competitive), then it follows that the fundamental cause of cross-country differences in output are differences in policies. More recently, the analysis of the determinants of development has emphasized the role of (endogenous) institutions and geography. Endogenizing the institutional structure seems like a natural next step in the development of the theory. However, serious theoretical limitations of our understanding of social choice theory in dynamic settings has limited progress so far. The direct role of geography is easily incorporated into the framework. However, to the extent that the geographic dimension is viewed as influencing (or determining) institutions and or policies, the same limitations apply.

In summary, the neoclassical growth model is still the basic framework to study questions that require understanding differences across countries, regions or individuals, in the *level* of some economic variable. The main challenge for future research is to develop a theory of social choices (policy choices) that is consistent with the dynamic framework.

See Also

▶ Neoclassical Growth Theory

Bibliography

Azariadis, C. 1993. *Intertemporal macroeconomics*. Cambridge: Blackwell Publishers.

Barro, R.J., and G.S. Becker. 1989. Fertility choice in a model of economic growth. *Econometrica* 57: 481–501.

Becker, G.S., and R.J. Barro. 1988. A reformulation of the economic theory of fertility. *Quarterly Journal of Economics* 103: 1–25.

Ben Porath, Y. 1967. The production of human capital and the life cycle of earnings. *Journal of Political Economy* 75: 352–365.

Bils, M., and P. Klenow. 2000. Does schooling cause growth? *American Economic Review* 90: 1160–1183.

Boldrin, M., and L.E. Jones. 2005. Fertility and social security. Staff Report No. 359, Federal Reserve Bank of Minneapolis.

Boldrin, M., and L. Montrucchio. 1986. On the indeterminacy of capital accumulation paths. *Journal of Economic Theory* 40: 26–39.

Brock, W.A., and L.J. Mirman. 1972. Optimal economic growth and uncertainty. *Journal of Economic Theory* 4: 479–513.

Burmeister, E. 1980. *Capital theory and dynamics*. Cambridge: Cambridge University Press.

Cass, D. 1965. Optimum growth in an aggregative model of capital accumulation. *Review of Economic Studies* 32: 233–240.

Chamley, C. 1986. Optimal taxation of capital income in general equilibrium with infinite lifetimes. *Econometrica* 54: 607–622.

Cooley, T.F. 1995. *Frontiers of business cycle research*. Princeton: Princeton University Press.

Correia, I. 1996. Should capital be taxed in the steady state? *Journal of Public Economics* 60: 147–151.

Debreu, G. 1954. Valuation equilibrium and Pareto optimum. *Proceedings of the National Academy of Sciences* 40: 588–592.

Debreu, G. 1959. *The theory of value*. New Haven/London: Yale University Press.

Diamond, P.A. 1965. National debt in a neoclassical growth model. *American Economic Review* 55: 1126–1150.

Doepke, M. 2005. Child mortality and fertility decline: Does the Barro–Becker model fit the facts? *Journal of Population Economics* 18: 337–366.

Donaldson, J.B., and R. Mehra. 1983. Stochastic growth with correlated production shocks. *Journal of Economic Theory* 29: 282–312.

Fischer, S. 1979. Capital accumulation on the transition path in a monetary optimizing model. *Econometrica* 47: 1433–1439.

Friedman, M. 1969. The optimum supply of money. In *The optimum supply of money and other essays*, ed. M. Friedman. Chicago: Aldine.

Jones, L.E., R.E. Manuelli, and P.E. Rossi. 1997. On the optimal taxation of capital income. *Journal of Economic Theory* 73: 93–117.

Jones, L.E., R.E. Manuelli, and H. Siu. 2005. Fluctuations in convex models of endogenous growth II: business cycle properties. *Review of Economic Dynamics* 8: 805–828.

Judd, K.J. 1985. Redistributive taxation in a perfect foresight model. *Journal of Public Economics* 28: 59–84.

Klenow, P., and A. Rodríguez-Clare. 1997. The neoclassical revival in growth economics: Has it gone too far? In *Macroeconomics annual 1997*, ed. B. Bernanke and J. Rotenberg. Cambridge, MA: MIT Press.

Koopmans, T.J. 1965. On the concept of optimal economic growth. In *The econometric approach to development planning*. Chicago: Rand McNally.

Manuelli, R.E., and A. Seshadri. 2007a. Human capital and the wealth of nations. Working paper, University of Wisconsin.

Manuelli, R.E., and A. Seshadri. 2007b. Explaining international fertility differences. Working paper, University of Wisconsin.

McKenzie, L.W. 1986. Optimal economic growth, Turnpike theorems and comparative dynamics. In *Handbook of mathematical economics*, ed. K.J. Arrow and M.D. Intriligator, Vol. 3. Amsterdam: North-Holland.

Prescott, E.J., and R. Mehra. 1980. Recursive competitive equilibrium: the case of homogeneous households. *Econometrica* 48: 1365–1379.

Ramsey, F.P. 1928. A mathematical theory of saving. *Economic Journal* 28: 543–559.

Sidrauski, M. 1967. Inflation and economic growth. *Journal of Political Economy* 75: 796–810.

Stokey, N.L., and R.E. Lucas. (with E.C. Prescott). 1989. Recursive methods in economic dynamics. Cambridge, MA: Harvard University Press.

Turnovsky, S.J., and W.A. Brock. 1980. Time consistency and optimal government policies in perfect Foresight equilibrium. *Journal of Public Economics* 13: 183–212.

Neoclassical Synthesis

Olivier Jean Blanchard

Abstract

The term ‘neoclassical synthesis’ appears to have been coined by Paul Samuelson to denote the consensus view of macroeconomics which emerged in the mid-1950s in the United States. This synthesis remained the dominant paradigm for another 20 years, in which most of the important contributions, by Hicks, Modigliani, Solow, Tobin and others, fit quite naturally. The synthesis had, however, suffered from the start from schizophrenia in its relation to microeconomics, which eventually led to a serious crisis from which it is only now re-emerging. I describe the initial synthesis, the mature synthesis, the crisis and the new emerging synthesis.

Keywords

Aggregate demand; Animal spirits; Consumption function; Countercyclical fiscal policy; Dynamic stochastic general equilibrium (DSGE) models; Econometrics; Excess demand and supply; Hicks, J.; Imperfect competition; Imperfect financial markets; Interest rates; Intertemporal utility maximization; Investment function; IS–LM model; Keynesianism; Life-cycle hypothesis; Liquidity constraints; Lucas, R.; Market clearing; Microfoundations; Modigliani, F.; Monetary policy; MPS–FMP model; Neoclassical synthesis; Neoclassical theory of investment; New classical synthesis; New Keynesian synthesis; New neoclassical synthesis; Nominal rigidities; Permanent-income hypothesis; Phillips curve; Pigou effect; Price and wage adjustment; Public debt; Rational behaviour; Real business cycles; Samuelson, P.; Stagflation; tâtonnement; Wage rigidity; Wealth effects

JEL Classifications

E1

The term ‘neoclassical synthesis’ appears to have been coined by Paul Samuelson to denote the consensus view of macroeconomics which emerged in the mid-1950s in the United States. In the third edition of *Economics* (1955, p. 212), he wrote:

> In recent years 90 per cent of American Economists have stopped being ‘Keynesian economists’ or ‘anti-Keynesian economists’. Instead they have worked toward a synthesis of whatever is valuable in older economics and in modern theories of income determination. The result might be called neoclassical economics and is accepted in its broad outlines by all but about 5 per cent of extreme left wing and right wing writers.

Unlike the old neoclassical economics, the new synthesis did not expect full employment to occur under laissez-faire; it believed, however, that, by proper use of monetary and fiscal policy, the old classical truths would come back into relevance.

This synthesis was to remain the dominant paradigm for another 20 years, in which most of the important contributions, by Hicks, Modigliani, Solow, Tobin and others, were to fit quite naturally. Its apotheosis was probably the large econometric models, in particular the MPS model developed by Modigliani and his collaborators, which incorporated most of these contributions in an empirically based and mathematically coherent model of the US economy. The synthesis had, however, suffered from the start from schizophrenia in its relation to microeconomics. This schizophrenia was eventually to lead to a serious crisis from which it is only now reemerging. I describe in turn the initial synthesis, the mature synthesis, the crisis and the new emerging synthesis.

The Initial Synthesis

The post-war consensus was a consensus about two main beliefs. The first was that the decisions of firms and of individuals were largely rational, and as such amenable to study using standard methods from microeconomics. Modigliani, in the introduction to his collected papers, stated it strongly:

> [One of the] basic themes that has dominated my scientific concern [has been to integrate] the main building blocks of the General Theory with the more established methodology of economics, which rests on the basic postulate of rational maximizing behavior on the part of economic agents. . .' (1980, p. xi)

The faith in rationality was far from blind: animal spirits were perceived as the main source of movements in aggregate demand through investment. For example, the possibility that corporate saving was too high and not offset by personal saving was considered a serious issue, and discussed on empirical rather than theoretical grounds.

This faith in rationality did not, however, extend to a belief in the efficient functioning of markets. The second main belief was indeed that prices and wages did not adjust very quickly to clear markets. There was broad agreement that markets could not be seen as competitive. But, somewhat surprisingly given the popularity of imperfect competition theories at the time, there was no attempt to think in terms of theories of price and wage setting, with explicit agents setting prices and wages. Instead, the prevailing mode of thinking was in terms of tâtonnement, with prices adjusting to excess supply or demand, along the lines of the dynamic processes of adjustment studied by Samuelson in his *Foundations of Economic Analysis* (1947). The Phillips curve, imported to the United States by Samuelson and Solow in 1960, was in that context both a blessing and a curse. It gave strong empirical support to a tâtonnement-like relation between the rate of change of nominal wages and the level of unemployment, but it also made less urgent the need for better microeconomic underpinnings of market adjustment. Given the existence of a reliable empirical relation and the perceived difficulty of the theoretical task, it made good sense to work on other and more urgent topics, where the marginal return was higher.

These twin beliefs had strong implications for the research agenda as well as for policy. Because prices and wages eventually adjusted to clear markets, and because policy could avoid prolonged disequilibrium anyway, macroeconomic research could progress along two separate lines. One could study long-run movements in output, employment and capital, ignoring business cycle fluctuations as epiphenomena along the path and using the standard tools of equilibrium analysis: 'Solving the vital problems of monetary and fiscal policy by the tools of income analysis will validate and bring back into relevance the classical verities' (Samuelson 1955, p. 360). Or one could instead study short-run fluctuations around that trend, ignoring the trend itself. This is indeed where most of the breakthroughs had been made by the mid- 1950s. Work by Hicks (1937) and Hansen (1949), attempting to formalize the major elements of Keynes's informal model, had led to the IS–LM model. Modigliani (1944) had made clear the role played by nominal wage rigidity in the Keynesian model. Metzler (1951) had shown the importance of wealth effects, and the role of government debt. Patinkin (1956) had clarified the structure of the macroeconomic model, and the relation between the demands for goods, money and bonds, in the case of flexible prices and wages. There was general agreement that, except in unlikely and exotic cases, the IS curve was downward sloping and the LM curve upward sloping. Post-war interest rates were high enough – compared with pre-war rates – to make the liquidity trap less of an issue. There was still, however, considerable uncertainty about the effect of interest rates on investment, and thus about the slope of the IS relation. The assumption of fixed nominal wages made by Keynes and early Keynesian models had been relaxed in favour of slow adjustment of prices and wages to market conditions. This was not seen, however, as modifying substantially earlier conclusions. The 'Pigou effect' (so dubbed by Patinkin in 1948), according to which low enough prices would increase real money and wealth, was not considered to be of much practical significance. Only activist policy could avoid large fluctuations in economic activity.

Refinements of the model were not taken as implying that the case for policy activism was any less strong than Keynes had suggested. Because prices and wages did not adjust fast enough, active countercyclical policy was needed to keep the economy close to full employment. Because prices and wages, or policies themselves, eventually got the economy to remain not far from its growth path, standard microeconomic principles of fiscal policy should be used to choose the exact mix of fiscal measures at any point in time. The potential conflict between their relative efficacy in terms of demand management, and their effect on the efficiency of economic allocation, were considered an issue but not a major problem. Nor was the fact that the market failure which led to short-run fluctuations in the first place was not fully understood or even identified.

The ground rules for cyclical fiscal policy were laid in particular by Samuelson in a series of contributions (1951, for example). Countercyclical fiscal policy was to use both taxes and spending; in a depression, the best way to increase demand was to increase both public investment and private investment through tax breaks, so as to equalize social marginal rates of return on both. Where the synthesis stood on monetary policy is less clear. While the potential of monetary policy to smooth fluctuations was generally acknowledged, one feels that fiscal policy was still the instrument of predilection, that policy was thought of as fiscal policy in the lead with accommodating monetary policy in tow.

The Mature Synthesis

For the next 20 years the initial synthesis was to supply a framework in which most macroeconomists felt at home and in which contributions fitted naturally. As Lucas remarks in his critique of the synthesis, 'those economists, like Milton Friedman, who made no use of the framework, were treated with some impatience by its proponents' (1980, p. 702). The research programme was largely implied by the initial synthesis, the emphasis on the behavioural components of IS–LM and its agnostic approach to price and wage adjustment; to quote Modigliani, 'the Keynesian system rests on four basic blocks: the consumption function, the investment function, the demand and the supply of money, and the mechanisms determining prices and wages' (1980, p. xii). Progress on many of these fronts was extraordinary; I summarize it briefly as these developments are reviewed in more depth elsewhere in this dictionary.

The failure of the widely predicted post-war over-saving to materialize had led to a reassessment of consumption theory. The theory of intertemporal utility maximization progressively emerged as the main contender. It was developed independently by Friedman (1957) as the 'permanent income hypothesis' and Modigliani and collaborators (1954 in particular) as the 'life cycle hypothesis'. The life-cycle formulation, modified to allow for imperfect financial markets and liquidity constraints, was, however, to dominate most of empirical research. Part of the reason was that it emphasized more explicitly the role of wealth in consumption, and, through wealth, the role of interest rates. Neither wealth effects nor interest rate effects on consumption had figured prominently in the initial synthesis.

Research on the investment function was less successful. Part of the difficulty arose from the complexity of the empirical task, the heterogeneity of capital, and the possibility of substituting factors *ex ante* but not *ex post.* Many of the conceptual issues were clarified by work on growth, but empirical implementation was harder. Part of the difficulty, however, came from the ambiguity of neoclassical theory about price behaviour, about whether firms could be thought of as setting prices or whether the slow adjustment of prices implied that firms were in fact output constrained. The 'neoclassical theory of investment' developed by Jorgenson and collaborators (for example, Hall and Jorgenson 1967) was ambiguous in this respect, assuming implicitly that price is equal to marginal cost, but estimating empirical functions with output rather than real wages.

Research on the demand for and supply of money was extended to include all assets. Solid foundations for the demand for money were given by Tobin (1956) and Baumol (1952), and the theory of finance provided a theory of the demand for all assets (Tobin 1958). The expectations hypothesis, which alleviated the need to estimate full demand and supply models of financial markets, was thoroughly tested and widely accepted as an approximation to reality.

In keeping with the initial synthesis, work on prices and wages was much less grounded in theory than work on the other components of the Keynesian model. While research on the microeconomic foundations of wage and price behaviour was proceeding (Phelps 1972 in particular), it was poorly integrated in empirical wage and price equations. To a large extent, this block of the Keynesian synthesis remained throughout the period the ad hoc but empirically successful Phillips curve, respecified through time to allow for a progressively larger effect of past inflation on current wage inflation.

All these blocks, together with work on growth theory, were largely developed in relation with and then combined in macroeconometric models, starting with the models estimated by Klein (for example, Goldberger and Klein 1955). The most important model was probably the MPS–FMP model developed by Modigliani and collaborators. This model, while maintaining the initial IS–LM Phillips curve structure of its ancestors, showed the richness of the channels through which shocks and policy could affect the economy. It could be used to derive optimal policy, show the effects of structural changes in financial markets, and so on. By the early 1970s the synthesis appeared to have been highly successful and the research programme laid down after the war to have been mostly completed. Only a few years later, however, the synthesis was in crisis and fighting for survival.

The Crisis and the Reconstruction

The initial trigger for the crisis was the failure of the synthesis to explain events. The scientific success of the synthesis had been largely due to its empirical success, especially during the Kennedy and the first phase of the Johnson administrations in the United States. As inflation increased in the late 1960s, the empirical success and, in turn, the theoretical foundations of the synthesis were more and more widely questioned. The more serious blow was, however, the stagflation of the mid-1970s in response to the increases in the price of oil: it was clear that policy was not able to maintain steady growth and low inflation. In a clarion call against the neoclassical synthesis, Lucas and Sargent (1978) judged its predictions to have been an 'econometric failure on a grand scale'.

One cannot, however, condemn a theory for failing to anticipate the shape and the effects of shocks which have not been observed before; few theories would pass such a test and, as long as the events can be explained after the fact, there is no particular cause for concern. In fact, soon thereafter models were expanded to allow for supply shocks such as changes in the price of oil. It became clear, however, that while the models could indeed be adjusted *ex post,* there was a more serious problem behind the failure to predict the events of the 1970s. To quote again from the polemical article by Lucas and Sargent, 'That the doctrine on which [these predictions] were made is fundamentally flawed is simply a matter of fact' (1978, p. 49). The 'fundamental flaw' was the asymmetric treatment of agents as being highly rational and of markets as being inefficient in adjusting wages and prices to their appropriate levels. The tension between the treatment of rational agents and that of myopic impersonal markets had been made more obvious by the developments of the 1960s, and the representation of consumers and firms as highly rational intertemporal decision makers. It was further highlighted by the research on fixed price equilibria, which went to the extreme of taking prices as unexplained and solving for macroeconomic equilibrium under non-market clearing. That research made clear, in a negative way, that progress could be made only if one understood why markets did not clear, why prices and wages did not adjust.

The solution proposed by Lucas and others in the 'new classical synthesis' was thoroughly unappealing to economists trained in the neoclassical synthesis. It was to formalize the economy as if markets were competitive and clearing instantaneously. The 'as if' assumption seemed objectionable on a priori grounds, in that direct evidence on labour and goods markets suggested important departure from competition; it also appeared to many to be an unpromising approach if the goal was to explain economic fluctuations and unemployment. Soon papers by Fischer (1977) and Taylor (1980) showed that one could replace the Phillips curve by a model of explicit nominal price and wage setting and still retain most of the traditional results of the neoclassical synthesis. These papers led the way to a major overhaul and reconstruction, and by the mid-1990s a new synthesis had emerged, a synthesis now dubbed the 'new neoclassical synthesis' (Goodfriend and King 1997) or the 'new Keynesian synthesis' (for example, Clarida et al. 1999). This new synthesis is described in more detail elsewhere in this dictionary, and I shall limit myself to a few remarks and comparisons between the old and the new. Like the old synthesis, the new synthesis has two major features: on the one hand, optimizing behaviour by firms, consumers and workers; on the other, the presence of distortions, most importantly nominal rigidities. In contrast to the old synthesis, however, the distortions are introduced explicitly, and price and wage behaviour is derived from optimizing behaviour by price and wage setters. These distortions imply that, as in the old synthesis, monetary policy and fiscal policy have a major role to play.

Like the old synthesis, the new synthesis is derived from microfoundations, utility maximization by consumers, and profit maximization by firms. But, while models in the old synthesis used theory as a loose guide to empirical specifications and allowed the data to determine the ultimate specification, models in the new synthesis remain much closer to their microfoundations. Dynamics are derived from the model itself, and the implied behavioural equations, rather than being estimated, are typically derived from assumptions about underlying technological and utility parameters. These more explicit microfoundations allow for a more careful welfare analysis of the implications of policy than was possible with the old models.

The models in the new synthesis are referred to as 'dynamic stochastic general equilibrium', or DSGE, models. Because they are typically difficult to solve, even the larger models are smaller than the models of the old synthesis, and their formalization of markets such as those for goods and labour remains primitive compared with the spirit of the formalizations in the old models. Improvements both in the formalization of these markets and in numerical techniques are, however, allowing for steadily richer and larger models.

To parallel the quotation from Samuelson given at the beginning, it is fair to say that the new neoclassical synthesis is attracting wide support, although less so than the old one. Some researchers, particularly those in the 'real business cycle' tradition, are sceptical about the importance of nominal rigidities in fluctuations. Others find the rationality assumptions embodied in the new synthesis to be too strong, and the methodology too constraining to capture the complexity present in the data.

Nevertheless, DSGE models are increasingly used to guide policy. Many challenges remain, for example in capturing the relevant distortions in goods, labour, financial, and credit markets, or in using econometrics to assess the fit of both the specific components and the overall model to reality. Progress is rapid, however. When I wrote the first version of this contribution in 1991, the emergence of a new synthesis appeared uncertain, and at best far in the future. In updating this contribution, I am struck by the progress that has taken place since then, and by the speed at which progress continues to be made today.

See Also

- ▶ Friedman, Milton (1912–2006)
- ▶ Hicks, John Richard (1904–1989)
- ▶ Klein, Lawrence R. (Born 1920)
- ▶ Lucas, Robert (Born 1937)
- ▶ Microfoundations
- ▶ Modigliani, Franco (1918–2003)
- ▶ Patinkin, Don (1922–1955)
- ▶ Phillips Curve (New Views)
- ▶ Samuelson, Paul Anthony (1915–2009)
- ▶ Tobin, James (1918–2002)

Bibliography

Baumol, W. 1952. The transactions demand for cash. *Quarterly Journal of Economics* 66: 545–546.

Clarida, R., J. Gali, and M. Gertler. 1999. The science of monetary policy: A New Keynesian perspective. *Journal of Economic Literature* 37: 1661–1707.

Fischer, S. 1977. Long-term contracts, rational expectations, and the optimal money supply rule. *Journal of Political Economy* 85: 191–205.

Friedman, M. 1957. *A theory of the consumption function*. New York: NBER.

Goldberger, A., and L. Klein. 1955. *An econometric model of the United States, 1929–1952*. Amsterdam: North-Holland.

Goodfriend, M., and R. King. 1997. The new neoclassical synthesis and the role of monetary policy. In *NBER macroeconomics annual 1997*, ed. B. Bernanke and J. Rotemberg. Cambridge: MIT Press.

Hall, R., and D. Jorgenson. 1967. Tax policy and investment behavior. *American Economic Review* 57: 391–414.

Hansen, A. 1949. *Monetary theory and fiscal policy*. New York: McGraw-Hill.

Hicks, J. 1937. Mr Keynes and the 'classics': A suggested interpretation. *Econometrica* 5: 147–159.

Lucas, R. 1980. Methods and problems in business cycle theory. *Journal of Money, Credit and Banking* 12: 696–715.

Lucas, R., and T. Sargent. 1978. After Keynesian macroeconomics. In *After the Phillips curve: Persistence of high inflation and high unemployment.* Boston: Federal Reserve of Boston.

Metzler, L. 1951. Wealth, saving and the rate of interest. *Journal of Political Economy* 59: 93–116.

Modigliani, F. 1944. Liquidity preference and the theory of interest and money. *Econometrica* 12: 45–88.

Modigliani, F. 1980. *Collected papers.* Vol. 1: Essays in macroeconomics. Cambridge, MA: MIT Press.

Modigliani, F., and R. Brumberg. 1954. Utility analysis and the consumption function: An interpretation of cross section data. In *Post-Keynesian economics*, ed. K. Kurihara. New Brunswick: Rutgers University Press.

Patinkin, D. 1948. Price flexibility and full employment. *American Economic Review* 38: 543–564.

Patinkin, D. 1956. *Money, interest and prices*. New York: Harper and Row.

Phelps, E. 1972. *Inflation policy and unemployment theory. London*: *Macmillan*.

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

Samuelson, P. 1951. Principles and rules in modern fiscal policy: A neoclassical reformulation. In *Money, trade and economic growth: Essays in honor of John Henry Williams*, ed. H. Waitzman. New York: Macmillan.

Samuelson, P. 1955. *Economics*. 3rd ed. New York: McGraw-Hill.

Taylor, J. 1980. Aggregate dynamics and staggered contracts. *Journal of Political Economy* 88: 1–23.

Tobin, J. 1956. The interest-elasticity of transactions demand for cash. *Review of Economics and Statistics* 38: 241–247.

Tobin, J. 1958. Liquidity preference as behavior towards risk. *Review of Economic Studies* 25: 65–86.

Neo-ricardian Economics

Heinz D. Kurz and Neri Salvadori

Abstract

This article deals with the revival of the classical theory of value and distribution, championed by Piero Sraffa. The general rate of profits and relative prices are shown to be determined exclusively in terms of the given system of production and real wages (or the share of wages). Prices generally depend on

income distribution. So does the cost-minimizing technique. The 'quantity of capital' cannot be ascertained independently of prices and thus the rate of profits. Techniques cannot generally be ordered monotonically with the rate of profits. Marginalist ideas regarding input proportions and input prices therefore cannot generally be sustained.

Keywords

Actual vs. normal values; Austrian economics; Cantillon, W.; Capital accumulation; Capital theory; Circular flow of production; Classical distribution theories; Classical economics; Cost-minimizing behaviour; Division of labour; Economic growth; Endogenous growth; Intertemporal equilibrium theory; Labour theory of value; Labour's share of income; Laws of capitalism; Long-period positions; Marginalist theory of value and distribution; Marx, K.; Methodological individualism; Mill, J.; Neo-Ricardian economics; New growth theory; Petty, W.; Physical real cost; Profits; Quesnay, F.; Reswitching; Ricardo, D.; Robinson, J.; Romer, P.; Say's Law; Simultaneous equations; Smith, A.; Social surplus; Sraffa, P.; Stratification; Technical change; Torrens, R.; Uniform rate of profits; Wicksell effects

JEL Classifications

B5

The term 'neo-Ricardian economics', as it is understood today, can mean several things. It was coined in the aftermath of the publication of *The Works and Correspondence of David Ricardo*, edited by Piero Sraffa with the collaboration of Maurice H. Dobb (Ricardo 1951/73), and the publication of Sraffa's *Production of Commodities by Means of Commodities* (Sraffa 1960). One meaning of the term simply refers to these facts and interprets Sraffa's work in the way Sraffa himself saw it: as a return to the 'standpoint of the old classical economists from Adam Smith to Ricardo, [which] has been submerged and forgotten since the advent of the "marginal" method' (Sraffa 1960, p. v; see Smith 1776, and Ricardo 1951/73). However, the term was first used by Marxist economists to distinguish Sraffa's approach to the theory of value and distribution, which explained relative prices and income distribution strictly in material terms (that is, quantities of commodities and labour), from the Marxist one, which starts from labour values (see Rowthorn 1974). In some contributions Sraffa's analysis is described in a derogatory manner as a 'peanut theory of profits' and rejected together with marginalist (or 'neoclassical') theory as a variant of 'vulgar economics', dealing with 'appearances' only, whereas Marxist theory is taken to investigate 'the real relations of production in bourgeois society' (Marx 1867, p. 85n). Neoclassical economists in turn occasionally (see, for example, Hahn 1982) applied the term to the analysis of those critics who, in the so-called Cambridge controversies on the theory of capital, had attacked marginalism, especially its long-period version, showing it to be logically flawed (see Kurz and Salvadori 1995, ch. 14). Because of the nationalities of the critics – especially Joan Robinson, Nicholas Kaldor, Piero Sraffa, Pierangelo Garegnani and Luigi Pasinetti – they also spoke of an 'Anglo-Italian school'.

Such an unfortunate diversity of meanings may reflect a misunderstanding both of Sraffa's achievement and of the relation of his analysis to that of Marxist and marginalist economics respectively. What Sraffa in fact provides is a reformulation of the *classical* approach to the problem of value and distribution that sheds the weaknesses of its earlier formulations and builds upon their strengths. Put briefly, profits and all property incomes (such as interest and land rents) are explained in terms of the *social surplus* left over after the necessary means of production and the wages in the support of workers have been deducted from the gross outputs produced during a year. As Ricardo had stressed: 'Profits come out of the surplus produce' (*Works*, vol. 2, pp. 130–1; cf. vol. 1, p. 95). Therefore, instead of 'neo-Ricardian economics' it would be more appropriate to speak of that part of classical economics that deals with value and distribution. As is well known, this part was designed to constitute the foundation of all other economic analysis, including the investigation of capital accumulation and technical progress, of development and growth, of social transformation and structural change, and of taxation and public debt. The pivotal role of the theory of value and distribution in the classical authors can be inferred from the fact that it is typically developed at the beginning of their major works. By rectifying this part, Sraffa revived interest in classical economics. In addition to this constructive task Sraffa also pursued a critical task: the propositions of his book were explicitly 'designed to serve as the basis for a critique of [the marginal theory of value and distribution]' (1960, p. vi).

In the following we first summarize the achievements of Sraffa and his followers with respect to the constructive task. We then turn to the criticism of marginalist theory. In conclusion, we point out some of the problems that are currently being tackled by scholars working in the classical tradition.

Reformulating the Classical Theory of Value and Distribution

The concern of the classical economists, especially Smith and Ricardo, was the laws governing the emerging capitalist economy, characterized by the stratification of society into three classes: workers, landowners, and the rising class of capitalists; wage labour as the dominant form of the appropriation of other people's capacity to work; an increasingly sophisticated division of labour within and between firms; the coordination of economic activity through a system of interdependent markets in which transactions were mediated through money; and significant technical, organizational and institutional change. In short, they were concerned with an economic system incessantly in motion. How to analyse such a system? The ingenious device of the classical authors to see through the complexities of the modern economy consisted in distinguishing between the 'actual' values of the relevant variables – the distributive rates and prices – and their 'normal' values. The former were taken to reflect all kinds of influences, many of an accidental or temporary nature, about which no general propositions were possible, whereas the latter were conceived of as expressing the persistent, non-accidental and nontemporary factors governing the economic system, which could be systematically studied.

The method of analysis adopted by the classical economists is known as the method of 'long-period positions' of the economy. Any such position is the situation towards which the system is taken to gravitate as the result of the self-seeking actions of agents, thereby putting into sharp relief the fundamental forces at work. In conditions of free competition the resulting long-period position is characterized by a *uniform rate of profits* (subject perhaps to persistent inter-industry differentials reflecting different levels of risk and of agreeableness of the business; see Kurz and Salvadori 1995, ch. 11) and uniform rates of remuneration for each particular kind of primary input. Competitive conditions were taken to engender *cost-minimizing behaviour* of profit-seeking producers.

Alfred Marshall (1920) had interpreted the classical economists as essentially early and somewhat crude demand and supply theorists, with the demand side in its infancy. It was this interpretation and the underlying continuity thesis in economics that Sraffa challenged. As he showed, the classical economists' approach to the theory of value and distribution was fundamentally different from the later marginalist one, and explained profits in terms of basically two data: (*a*) the system of production in use and (*b*) a given real wage rate (or, alternatively, a given share of wages). Profits (and rents) were thus conceived of as a *residual* income. Whereas in marginalist theory wages and profits are treated symmetrically, in classical theory they are treated *asymmetrically.* On a still deeper methodological level the divide between the classical and the later marginalist authors could hardly be more pronounced. While the classical authors took the economic system to exist independently of the single agent and actually exert a considerable influence upon the latter depending upon the role ascribed to him as worker, capitalist or landowner, the marginalist authors advocated one version or another of 'methodological individualism', which takes a set of assumedly optimizing agents who exist independently of the system as a whole and who shape the system rather than the other way round.

Let us now examine more closely the scope, content and analytical structure of classical theory. The classical economists proceeded essentially in two steps. In the first step they isolated the kinds of factors that were seen to determine income distribution and the prices supporting that distribution in specified conditions, that is, *in a given place and time*. The theory of value and distribution was designed to identify *in abstracto* the dominant factors at work and to analyse their interaction. In the second step they turned to an investigation of the causes which *over time* affected systematically the factors at work from within the economic system. This was the realm of the classical analysis of capital accumulation, technical change, economic growth and socio-economic development.

It is another characteristic feature of the classical approach to profits, rents and relative prices that these are explained essentially in terms of magnitudes that can, in principle, be observed, measured or calculated. The *objectivist* orientation of classical economics has received its perhaps strongest expression in a famous proclamation by William Petty, who was arguably its founding father. Keen to assume what he called the '"physician's" outlook', Petty in his *Political Arithmetick*, published in 1690, stressed that he was to express himself exclusively 'in Terms of *Number*, *Weight* or *Measure*' (Petty 1986, p. 244). And James Mill noted significantly that '*The agents of production are the commodities themselves* . . .. They are the food of the labourer, the tools and the machinery with which he works, and the raw materials which he works upon' (Mill 1826, p. 165, emphasis added). According to Sraffa the classical authors advocated essentially a concept of *physical real cost*. Man cannot create matter, man can only change its form and move it. Production involves destruction, and the real cost of a commodity consists in the commodities destroyed in the course of its production. This concept differs markedly from the later marginalist concepts, with their emphasis on 'psychic cost', reflected in such notions as 'utility' and 'disutility'.

In line with what may be called their 'thermodynamic' view, the classical authors saw production as a *circular flow*. This idea can be traced back to William Petty and Richard Cantillon, and was most effectively expressed by François Quesnay (1759) in the *Tableau économique*: commodities are produced by means of commodities. This is in stark contrast with the view of production as a one-way avenue leading from the services of original factors of production via some intermediate products to consumption goods, as was entertained by the 'Austrian' economists.

Why then did the classical economists fail to elaborate a consistent theory of value and distribution on the basis of the twin concepts of (*a*) physical real costs and (*b*) a circular flow of production? According to Sraffa (see Kurz and Salvadori 2005) a main, if not *the* main, reason consisted in a mismatch between highly sophisticated analytical concepts on the one hand and inadequate tools available to the classical authors to deal with them on the other. More specifically, the tool needed in order to bring to fruition an analysis based on these twin concepts was simultaneous equations: knowledge of how to solve them and how to discover what their properties are. This indispensable tool (alas!) was not at their disposal. They therefore tried to solve the problems they encountered in a roundabout way, typically by first identifying an 'ultimate standard of value' by means of which *heterogeneous* commodities could be rendered *homogeneous*. Several authors, including Smith, Ricardo and Marx, had then reached the conclusion that 'labour' was the standard they sought and had therefore arrived in one way or another at some version of the labour theory of value. This preserved the objectivist character of the theory by taking as data, or known quantities, only measurable things, such as amounts of commodities actually produced and amounts actually used up, including the means of subsistence in the support of workers. This was understandable in view of the unresolved tension between concepts and tools. However, with production as a circular flow, even labour values cannot be known independently of solving a system of simultaneous equations. Hence the route via labour values was not really a way out of the impasse in which the classical authors found themselves: it rather landed them right in that impasse again. Commodities were produced by means of commodities and there was no way to circumnavigate the simultaneous equations approach.

What made it so difficult, if not impossible, for the classical authors to see that the theory of value and distribution could be firmly grounded in the concept of physical real cost? Given their primitive tools of analysis, they did not see that the information about the system of production in use and the quantities of the means of subsistence in support of workers was all that was needed in order to determine *directly* the system of necessary prices and the rate of profits. Sraffa understood this as early as November 1927, as we can see from his hitherto unpublished papers kept at Trinity College Library, Cambridge (UK), with respect to what he called his 'first' (without a surplus) and 'second' (with a surplus) 'equations'.

We may start with James Mill's aforementioned case with three kinds of commodities, tools (t), raw materials (m), and the food of the labourer (f). Production in the three industries may then be depicted by the following system of quantities

$$\begin{array}{l} T_t \bigoplus M_t \bigoplus F_t \rightarrow T \\ T_m \bigoplus M_m \bigoplus F_m \rightarrow M \\ T_f \bigoplus M_f \bigoplus F_f \rightarrow F \end{array} \tag{1}$$

where T_i, M_i and F_i designate the inputs of the three commodities (employed as means of production *and* means of subsistence) in industry $i(i = t, m, f)$, and T, M and F total outputs in the three industries; the symbol $\bigoplus$ indicates that all inputs on the LHS of →, representing production are required to generate the output on its RHS. Invoking classical concepts, Sraffa called these relations 'the methods of production and productive consumption' (1960, p. 3). In the hypothetical case in which the economy is just viable, that is, able to reproduce itself without any surplus (or deficiency), we have $T = \Sigma_i T_i$, $M = \Sigma_i M_i$, and $F = \Sigma_i F_i$.

From this schema of reproduction and reproductive consumption we may directly derive the corresponding system of 'absolute' or 'natural' values, which expresses the idea of physical real cost-based values in an unadulterated way. Denoting the value of one unit of commodity i by p_i, $p_i(i = 1, m, f)$ we have

$$\begin{array}{l} T_t p_t + M_t p_m + F_t p_f = T p_t \\ T_m p_t + M_m p_m + F_m p_f = M p_m \\ T_f p_t + M_f p_m + F_f p_f = F p_f \end{array} \tag{2}$$

These linear equations are homogeneous and therefore only relative prices can be determined. Further, only two of the three equations are independent of one another. This is enough to determine the two relative prices. Alternatively, it is possible to fix a standard of value whose price is *ex definitione* equal to unity. This provides an additional (non-homogeneous) equation without adding a further unknown, and allows one to solve for the remaining dependent variables.

A numerical example illustrates the important finding that the given sociotechnical relations rigidly fix relative values:

$$\begin{array}{ll} & \textit{Values} \\ 2p_t + 15p_m + 20p_f = 17p_t & p_t = 3p_m \\ 5p_t + 7p_m + 4p_f = 28p_m & p_m = \dfrac{2}{3}p_f \\ 10p_t + 6p_m + 11p_f = 35p_f & p_f = \dfrac{1}{2}p_t \end{array}$$

These values depend exclusively on necessities of production. They are the only ones that allow the initial distribution of resources to be restored. Apparently, the value of one commodity may be 'reduced' to a certain amount of another commodity needed directly or indirectly in the production of the former. For example, one might reduce one unit of commodity t to an amount needed of commodity m. Hence one might say that each of the three commodities could serve as a 'common measure' and that, for example, commodities t and f exchange for one another in the proportion 1:2 because commodity t 'contains' or 'embodies' twice as much of commodity m as commodity f.

There is no need even to talk about labour values at this stage of the argument. The same applies to the next stage, which refers to a system with a surplus and given commodity (or real) wages advanced at the beginning of the production period. In conditions of free competition the surplus will be distributed in terms of a *uniform* rate of profits on the 'capitals' advanced in the different industries.

We start again from the system of quantities consumed productively and produced (1), but now we assume that $T \geq \Sigma_i T_i$, $M \geq \Sigma_i M_i$, and $F \geq \Sigma_i F_i$ where at least with regard to one commodity the strict inequality sign holds. In conditions of free competition 'normal' prices, or 'prices of production', have to satisfy the following system of price equations:

$$\begin{aligned}(T_t p_t + M_t p_m + F_t p_f)(1+r) &= T p_t \\ (T_m p_t + M_m p_m + F_m p_f)(1+r) &= M p_m \\ (T_f p_t + M_f p_m + F_f p_f)(1+r) &= F p_f\end{aligned} \qquad (3)$$

The case of a uniform rate of physical surplus across all commodities contemplated by David Ricardo and Robert Torrens

$$\frac{T - \Sigma_i T_i}{\Sigma_i T_i} = \frac{M - \Sigma_i M_i}{\Sigma_i M_i} = \frac{F - \Sigma_i F_i}{\Sigma_i F_i} = r \qquad (4)$$

denotes a very special constellation: in it the general rate of profits, r, equals the uniform material rate of produce. *Here we see the rate of profits in the commodities themselves, as having nothing to do with their values.* In this case only two of the Eq. (3) are linearly independent so that Eq. (4) determines the rate of profits, and Eq. (3), following the same procedure used for Eq. (2), determine relative prices. In general, the rates of physical surplus will be different for different commodities. Unequal rates of commodity surplus do not, however, by themselves imply unequal rates of profit across industries.

In this case there are three numbers, each of which substituted for r in Eq. (3) makes them linearly dependent on one another with respect to prices. It is possible to show that, when the highest real number among such numbers is substituted for r, the corresponding relative prices are positive, whereas when any of the other numbers is substituted for r some relative prices are negative. Since a negative relative price has no economic meaning in the present context, we can assert that there is a single solution which is relevant from an economic point of view. Fixing a standard of value provides a fourth equation and no extra unknown, so that the system of equations can be solved.

The important point to note here is the following. With the real wage rate given and paid at the beginning of the periodical production cycle, the problem of the determination of the rate of profits consists in distributing the surplus product in proportion to the capital advanced in each industry. Obviously,

> such a proportion between two aggregates of heterogeneous goods (in other words, the rate of profits) cannot be determined before we know the prices of the goods. On the other hand, we cannot defer the allotment of the surplus till after the prices are known, for…the prices cannot be determined before knowing the rate of profits. *The result is that the distribution of the surplus must be determined through the same mechanism and at the same time as are the prices of commodities*. (Sraffa 1960, p. 6; emphasis added)

This passage shows that the idea which underlies Marx's so-called 'transformation' of labour values into prices of production (see Marx 1894, part 2) cannot generally be sustained. Marx had proceeded in two steps; Ladislaus von Bortkiewicz (1906/7, essay 2, p. 38) aptly dubbed his approach 'successivist' (as opposed to 'simultaneous'). In a first step Marx had assumed that the general rate of profits is determined independently of, and prior to, the determination of prices as the ratio between the labour value of the social surplus and that of social capital, consisting of 'constant capital' (means of production) and 'variable capital' (wages or means of subsistence). In a second step he had then used this rate to calculate prices.

So far we have assumed that real wages are given in kind at some level of subsistence. The classical economists, however, saw clearly that wages may rise above mere sustenance of labourers, which makes necessary a new wage concept. This case had made Ricardo adopt a *share* concept of wages and establish the inverse relationship between the share of wages in the product and the rate of profits: ‘The greater the *portion of the result of labour* that is given to the labourer, the smaller must be the *rate* of profits, and vice versa’ (*Works*, vol. 8, p. 194; emphasis added). The concept of ‘proportional wages’, as Sraffa called it, was then adopted by Marx in terms of a given rate of surplus value. Sraffa also adopted the concept, albeit with two important changes. First, when workers participate in the sharing out of the surplus product, the original classical idea of wages being entirely paid out of social capital can no longer be sustained. After some deliberation Sraffa decided to treat wages as a whole as paid out of the product. Second, he did not express the share of wages in terms of labour but as the ratio of total wages to the net product expressed in terms of normal prices, *w*. These changes necessitated reformulating the price equations by taking explicitly into account the amounts of labour expended in the different industries, L_i($i = t, m, f$), because wages are taken to be paid in proportion to these amounts, and by defining these amounts as fractions of the total annual labour of society, that is, $L_t + L_m + L_f = 1$. In addition, it is assumed, following the classical economists, that differences in the quality of labour have been previously reduced to equivalent differences in quantity, so that each unit of labour receives the same wage rate (see Kurz and Salvadori 1995, ch. 11). We may now formulate the corresponding system of production equations again for the case of the three kinds of commodities mentioned by Mill, where now the quantities represented by *Ti*, *Mi* and *Fi* refer exclusively to the inputs of the three commodities employed as means of production. We get (on the assumption that wages are paid *post factum*)

$$\begin{aligned}\left(T_t p_t + M_t p_m + F_t p_f\right)(1+r) + L_t w &= T p_t\\ \left(T_m p_t + M_m p_m + F_m p_f\right)(1+r) + L_m w &= M p_m\\ \left(T_f p_t + M_f p_m + F_f p_f\right)(1+r) + L_f w &= F p_f\end{aligned} \tag{5.1}$$

With the net product taken as standard of value, we have in addition that

$$(T - \Sigma_i T_i)p_t + (M - \Sigma_i M_i)p_m + (F - \Sigma_i F_i)p_f = 1.$$

Taking one of the distributive variables, the share of wages *w* (or the rate of profits *r*) as given, allows one to determine the remaining variables: *r* (or *w*) and the prices of commodities.

Using this approach, Sraffa was able to show that, whereas the wage rate as a function of the rate of profits is necessarily decreasing (but does not need to be so if commodities are produced jointly), any relative price as a function of the rate of profits typically does not follow a simple rule: the function can alternately be increasing or decreasing, and can pass through unity a number of times (but such a number is constrained by the overall number of commodities involved). This fact is important also because the problem of the choice of technique from among several alternatives can be studied by following substantially the same argument. Suppose, for instance, that commodity *t* can be produced also with process

$$T'_t \oplus M'_t \oplus F'_t \oplus L'_t \to T'$$

Then we can add to system (5.1) the equation

$$\left(T'_t p_t + M'_t p_m + F'_t p_f\right)(1+r) + L'_t w = T' p'_t \tag{5.2}$$

with the further unknown p'_t.The study of the ratio p'_t/p_t allows one to say when it is profitable to use the old process and when the new one: if p'_t/p_t is smaller than 1, the new process will be chosen by cost-minimizing producers; if it is larger than 1, the old process will be retained, whereas the two processes can coexist in case $p'_t/p_t = 1$ Obviously, if the new process is chosen and has replaced the old one, and if it is assumed that the rate of profits is unchanged, then Eq. (5.1) give way to the following equations, serving as the new system

$$\begin{aligned}&\left(T'_t p'_t + M'_t p'_m + F'_t p'_f\right)(1+r) + L'_t w' = T' p' t\\&\left(T_m p'_t + M_m p'_m + F_m p'_f\right)(1+r) + L_m w' = M p' m\\&\left(T_f p'_t + M_f p'_m + F_f p'_f\right) \times (1+r) + L_f w' = F p'_f\end{aligned} \tag{6.1}$$

In this new system prices and the wage are different $\left(p'_j \neq p_j \, and \, w' \neq w\right)$ but they are not so when $p'_j/p_t = 1$ in system (5). If we now evaluate the old process in terms of the prices and wage of the new system by combining system (6.1) and the equation

$$\left(T_t p'_t + M_t p'_m + F_t p'_f\right)(1+r) L_t w' = T p_t \tag{6.2}$$

we can calculate again the ratio p'_t/p_t and the property that prices and the wage in the two systems coincide when $p'_t/p_t = 1$ is enough to prove that p'_t/p_t is larger (lower) than 1 for a given r in system (6) if and only if it is so in system (5). Hence the comparison between the new process and the old one can be indifferently done at the prices of either the old system or the new system.

In the following a system involving a number of processes equal to the number of commodities involved, each producing a different commodity, is called a *technique*, and a technique which is chosen at a given income distribution is called a *cost-minimizing technique* at that income distribution. The fact that a relative price can pass through unity at several income distributions implies that a technique can be cost-minimizing at different values of the rate of profits, with other techniques being cost minimizing in the interval in between. This fact has been called *reswitching*; it played an important role in the criticism of neoclassical theory.

In the above it has for simplicity been assumed that there is only single production, that is, only circulating capital. While the circulating part of the capital goods advanced in production contributes entirely and exclusively to the output generated, that is, 'disappears' from the scene, so to speak, the fixed part of it contributes to a sequence of outputs over time, that is, after a single round of production its items are still there – older but still useful. For a discussion of joint production, fixed capital and scarce natural resources, see Kurz and Salvadori (1995).

Critique of Marginalist Theory

The passage quoted above from Sraffa (1960, p. 6) contains the key to his critique of the long-period marginalist concept of capital. This concept hinges crucially on the possibility of defining the 'quantity of capital', whose relative scarcity and thus marginal productivity was taken to determine the rate of profits, independently of the rate of profits. However, according to the logic of Sraffa's above argument the rate of profits and the quantity (that is, value) of social capital ($\Sigma_i T_i p_t + \Sigma_i M_i p_m + \Sigma_i F_i pf$) can only be determined simultaneously.

We may approach the issues under consideration by first discussing what are known as 'Wicksell effects'. The term was introduced by Joan Robinson (1953, p. 95) during a debate in the theory of capital (see Kurz and Salvadori 1995, ch. 14). We distinguish between *price Wicksell effects* and *real Wicksell effects* (henceforth PWE and RWE). A PWE relates to a change in relative prices corresponding to a change in income distribution, given the system of production in use. A RWE relates to a change in technique, with the fact taken into account that at the income distribution at which two techniques are both cost-minimizing (one being so at higher, the other at lower levels of the rate of profits) both techniques have the same prices. The 'changes' under consideration refer to comparisons of long-period equilibria.

Marginalist theory contends that both effects are invariably positive. A *positive* PWE means that with a rise (fall) in the rate of interest prices of consumption goods will tend to rise (fall) relative to those of capital goods. The reason given is that consumption goods are said to be produced more capital intensively than capital goods: consumption goods emerge at the end of the production process, whereas capital goods are intermediate products that gradually 'mature' towards the final product. The higher (lower) is the rate of interest the less (more) expensive are the intermediate products in terms of a standard consisting of a (basket of) consumption good(s). At the macro level of a stationary economy (in which the net product contains only consumption goods) this implies that with a rise in the rate of interest the value of the net social product rises relatively to the value of the aggregate of capital goods employed. Clearly, seen from the marginalist perspective, a positive PWE with regard to the relative price of the two aggregates under consideration involves a negative relationship between the aggregate capital-to-net output ratio on the one hand and the interest rate on the other. Let $K/Y = \mathbf{xp}(r)/\mathbf{yp}(r)$ ($\mathbf{x}$ is the row vector of capital goods, $\mathbf{y}$ the row vector of net outputs, and $\mathbf{p}(r)$ the column vector of prices (in terms of the consumption vector) which depends on r) designate the capital-output ratio, then the marginalist message is:

$$\frac{\partial(K/Y)}{\partial r} \leq 0$$

Since for a given system of production the amount of labour is constant irrespective of the level of the rate of interest, also the ratio of the value of the capital goods and the amount of labour employed, or capital–labour ratio, K/L, would tend to fall (rise) with a rise (fall) in the rate of interest,

$$\frac{\partial(K/L)}{\partial r} \leq 0 \tag{7}$$

This is the first claim marginalist authors put forward. The second is that RWEs are also positive. A *positive* RWE means that with a rise (fall) in the rate of interest cost-minimizing producers switch to methods of production that generally exhibit higher (lower) labour intensities, 'substituting' for the 'factor of production' that has become more expensive – 'capital' (labour) – the one that has become less expensive – labour ('capital'). Hence (7) is said to apply also in this case. The assumed positivity of the RWE underlies the marginalist concept of a demand function for labour (capital) that is inversely related to the real wage rate (rate of interest).

Careful scrutiny of the marginalist argument has shown that it cannot generally be sustained: there is no presumption that PWEs and RWEs are invariably positive. In fact there is no presumption that techniques can be ordered monotonically with the rate of interest (Sraffa 1960). Reswitching implies that, even if PWEs happen to be positive, RWEs cannot always be positive. As Mas-Colell (1989) stressed, the relationship between K/L and r can have almost any shape whatsoever. In the intervals in which K/L is an increasing function of r we say that there is *capital reversal*. It implies that, if the neoclassical approach to value and distribution is followed, the 'demand for capital' is not decreasing, and therefore the resulting equilibrium, provided there is one, is not stable. Hence the finding that PWEs and RWEs need not be positive challenges the received doctrine of the working of the economic system, as it is portrayed by conventional economic theory with its reference to the 'forces' of demand and supply (see Pasinetti 1966; Garegnani 1970; see also Harcourt 1972; Kurz and Salvadori 1995, ch. 14; 1998c).

Current Work in the Classical Tradition

In more recent times authors working in the classical tradition, as it was revived by Sraffa, have focused attention on a large number of problems. First, there has been a lively interest in generalizing the results provided by Sraffa on joint production, fixed capital, and land. Then the approach was extended to cover renewable and exhaustible resources and to allow for the more realistic case of costly disposal, which leads to the concept of negative prices of products that have to be disposed of. There is also a renewed interest in the problem of economic growth and development. Freed from the straightjacket of Say's Law, which can be said to be an implication of the finding that conventional equilibrium analysis cannot be sustained, there is no presumption that the economy will consistently follow a full-capacity path of economic expansion. Hence the problem of different degrees and modes of utilization of productive capacity and the role of effectual demand (Adam Smith) have to be analysed. This avenue has opened up avenues for cross-fertilization between classical economics on the one hand, and Keynesian economics, based on the principle of effective demand, and evolutionary economics, concerned with complex dynamics, on the other (see Coase 1976; Nelson 2005). This fact is also highlighted in comparisons with the so-called new growth theory, and allows one to better understand the latter's merits and demerits (see Kurz and Salvadori 1998a, ch. 4; 1999).

In the 1960s and 1970s the long-period versions of marginalist theory revolving around the concept of a uniform rate of return on capital were called into question on logical grounds. While many marginalist authors accepted this criticism, some of them contended that intertemporal equilibrium theory, the 'highbrow version' of neoclassicism, was not affected by it (see especially Bliss 1975; Hahn 1982). This claim has more recently been subjected to close scrutiny (see Garegnani 2000, Schefold 2000, and the special issue of *Metroeconomica*, vol. 56(4), 2006). While the criticism of the long-period versions of marginalist theory is irrefutable, as authors from Paul Samuelson to Andreu Mas-Colell have admitted, surprisingly this has not prevented the economics profession at large from still using this theory. This is perhaps so because in more recent years the way of theorizing in large parts of mainstream economics has fundamentally changed. Whether this change is a response to the criticism need not concern us here. It suffices to draw the reader's attention to a statement by Paul Romer in one of his papers on endogenous growth in which he self-critically pointed out a slip in his earlier argument. The error he had committed, he wrote, 'may seem a trifling matter in an area of theory that depends on so many other short cuts. After all, if one is going to do violence to the complexity of economic activity by assuming that there is an aggregate production function, how much more harm can it do to be sloppy about the difference between rival and nonrival goods?' (Romer 1994, pp. 15–16) Once economic theory has taken the road indicated, criticism becomes a barren instrument. Indeed, why should someone who seeks to provide 'microfoundations' in terms of a representative agent with an infinite time horizon find fault with the counter-factual but attractive assumption that there is only a single (capital) good?

See Also

- ▶ Capital Theory
- ▶ Capital Theory (Paradoxes)
- ▶ Classical Growth Model
- ▶ Classical Distribution Theories
- ▶ Classical Production Theories
- ▶ Reswitching of Technique
- ▶ Ricardo, David (1772–1823)
- ▶ Smith, Adam (1723–1790)
- ▶ Sraffa, Piero (1898–1983)
- ▶ Sraffian Economics
- ▶ Sraffian Economics (New Developments)

Bibliography

Bliss, C. 1975. *Capital theory and the distribution of income*. Amsterdam: North-Holland.

Bortkiewicz, L von. 1906/7. Wertrechnung und preisrechnung im marxschen system. *Archiv für Sozialwissenschaft und Sozialpolitik* 23(1906):1–50 (essay 1), 25 (1907), 10–51 (essay 2) and 445–88 (essay 3).

Coase, R. 1976. Adam Smith's view of man. *Journal of Law and Economics* 19: 529–546.

Garegnani, P. 1970. Heterogeneous capital, the production function and the theory of distribution. *Review of Economic Studies* 37: 407–436.

Garegnani, P. 1987. Surplus approach to value and distribution. In *The new palgrave: A dictionary of economics*, ed. J. Eatwell, M. Milgate, and P. Newman, vol. 4. London: Macmillan.

Garegnani, P. 2000. Savings, investment and the quantity of capital in general intertemporal equilibrium. In *Critical essays on Piero Sraffa's legacy in economics*, ed. H. Kurz. Cambridge: Cambridge University Press.

Hahn, F. 1982. The neo-ricardians. *Cambridge Journal of Economics* 6: 353–374.

Harcourt, G. 1972. *Some cambridge controversies in the theory of capital*. Cambridge: Cambridge University Press.

Kurz, H., eds. 2000. *Critical essays on Piero Sraffa's legacy in economics*. Cambridge: Cambridge University Press.

Kurz, H., and N. Salvadori. 1995. *Theory of productio: A long-period analysis*. Cambridge: Cambridge University Press.

Kurz, H., and N. Salvadori. 1998a. *Understanding 'classical' economics: Studies in long-period theory*. London: Routledge.

Kurz, H., and N. Salvadori, eds. 1998b. *The elgar companion to classical economics*. Vol. 2. Cheltenham/Northhampton: Edward Elgar.

Kurz, H., and N. Salvadori. 1998c. Reverse capital deepening and the numeraire: A note. *Review of Political Economy* 10: 415–426.

Kurz, H., and N. Salvadori. 1999. Theories of 'endogenous' growth in historical perspective. In *Contemporary economic issues. Proceedings of the eleventh World congress of the international economic association, volume 4: Economic behaviour and design*, ed. M. Sertel. London: Macmillan.

Kurz, H., and N. Salvadori. 2005. Representing the production and circulation of commodities in material terms: On Sraffa's objectivism. *Review of Political Economy* 17: 414–441.

Marshall, A. 1920. *Principles of economics*. 8th ed. London: Macmillan.

Marx, K. 1867. *Capital*. Vol. 1, 1954. Moscow: Progress Publishers.

Marx, K. 1894. *Capital*. Vol. 3, 1959. Moscow: Progress Publishers.

Mas-Colell, A. 1989. Capital theory paradoxes: Anything goes. In *Joan Robinson and modern economic theory*, ed. R. Feiwel. London: Macmillan.

Mill, J. 1826. *Elements of political economy*, 3rd edn, reprinted 1844. London: Baldwin, Cradock, and Joy.

Nelson, R. 2005. *Technology, institutions, and economic growth*. Cambridge, MA/London: Harvard University Press.

Petty, W. 1986. *The economic writings of Sir William Petty*. New York: Kelley.

Quesnay, F. 1759. In *Quesnay's tableau economique*, ed. M. Kuczynski and R. Meek, 1972. London: Macmillan.

Pasinetti, L. 1966. Changes in the rate of profit and switches of techniques. *Quarterly Journal of Economics* 80: 503–517.

Ricardo, D. 1951/73. *The works and correspondence of David Ricardo*, 11 vols, ed. P. Sraffa with the collaboration of M. Dobb. Cambridge: Cambridge University Press. (In the text referred to as *Works*, volume number.)

Robinson, J. 1953. The production function and the theory of capital. *Review of Economic Studies* 21: 81–106.

Romer, P. 1994. The origins of endogenous growth. *Journal of Economic Perspectives* 8(1): 3–22.

Rowthorn, R. 1974. Neo-classicism, neo-Ricardianism and Marxism. *New Left Review* 86: 63–87.

Schefold, B. 2000. Paradoxes of capital and counter-intuitive changes of distribution in an intertemporal equilibrium model. In *Critical essays on Piero Sraffa's legacy in economics*, ed. H. Kurz. Cambridge: Cambridge University Press.

Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*, 1976. Oxford: Oxford University Press.

Sraffa, P. 1960. *Production of commodities by means of commodities: Prelude to a critique of economic theory.* Cambridge: Cambridge University Press.

neo-Ricardianism

Murray Milgate

The term neo-Ricardianism appeared in the literature in the 1970s to describe work in economic theory undertaken in the spirit of Piero Sraffa's *Production of Commodities by Means of Commodities*. The original impulse to the invention of this category came from certain modern Marxists who were anxious to distinguish their own arguments from anything that might have been contained in Sraffa's book. To the extent that Sraffa himself spoke of his work as a return to the standpoint 'of the old classical economists from Adam Smith to Ricardo' (1960, p. v), there is some basis for the designation. Its relationship to Marxism was then supposedly settled with the observation that 'the Marxian theory of *value* ought to be understood as a *critique* rather than a development of Ricardo's theory' (Medio 1972, p. 313). This line of argument was taken up by Rowthorn (1974) in what remains perhaps the benchmark case of a modern Marxist critique of neo-Ricardianism.

Since it is its alleged depreciation of the contributions of Marx that draws Marxist criticism upon Sraffa's work, it is evident that much of the modern Marxist hostility to neo-Ricardianism has historical roots. The work on the theory of Ricardo and Marx by Bortkiewicz, for example, concluded with an argument which held that as far as formal theory was concerned Marx added nothing to what was already to be found in Ricardo. Rowthorn cites this against neo-Ricardianism (1974, p. 29). Moreover Dmitriev, in his return to Ricardo, reached similar conclusions, and even attempted to provide a synthesis between that approach and the theory of marginal utility. How closely Sraffa might be said to follow these arguments is open to question, but certainly some 'neo-Ricardians' have been said (not without justification) to follow them quite closely. In this latter context the reader might consult the work of Steedman (1977 and 1982).

In more mainstream circles the term is also used to describe (and criticize) the same group of theorists. This is the manner in which it is used by Hahn. The purpose of these critics of Sraffa is not so much to separate Sraffa from Marx as it is to argue that 'there is no correct neo-Ricardian proposition which is not contained in the set of propositions which can be generated by orthodoxy' (Hahn 1982, p. 353). It is worth noting that this last idea is shared by some Marxists (see, for example, Rowthorn 1974, pp. 26–7).

See Also

- ▶ British Classical Economics
- ▶ Marxism
- ▶ Natural and Normal Conditions
- ▶ Ricardo, David (1772–1823)
- ▶ Sraffa, Piero (1898–1983)
- ▶ Sraffian Economics

Bibliography

De Vivo, G. 1982. Notes on Marx's critique of Ricardo. *Contributions to Political Economy* 1: 87–99.

Hahn, F. 1982. The neo-Ricardians. *Cambridge Journal of Economics* 6: 356–374. As reprinted in F. Hahn, *Equilibrium and macroeconomics*. Oxford: Blackwell, 1984.

Medio, A. 1972. Profits and surplus value: Appearance and reality in capitalist production. In *A critique of economic theory*, ed. E.K. Hunt and J.G. Schwartz. Harmondsworth: Penguin.

Rowthorn, B. 1974. Neo-classicism, neo-Ricardianism and Marxism. *New Left Review*. As reprinted in B. Rowthorn, *Capitalism, conflict and inflation*, 14–47. London: Lawrence & Wishart, 1980.

Sraffa, P. 1960. *Production of commodities by means of commodities*. Cambridge: Cambridge University Press.

Steedman, I. 1977. *Marx after Sraffa*. London: New Left Books.

Steedman, I. 1982. Marx on Ricardo. In *Classical and Marxian political economy*, ed. I. Bradley and M. Howard. London: Macmillan.

Net Product

Paolo Varri

The net product of a nation is the total amount of all commodities and services produced in that nation in a given period of time in excess of the commodities and services that have been required for its production. This definition coincides with the notion of wealth first introduced by Adam Smith (1776). The main difference between the modern concept and the original one concerns wages that we now consider as part of the net product but were initially (and until Marx) included among the advances to be reproduced.

The idea of a net product, literally *'produit net'*, as a final result of the economic activity of a whole nation, initially emerged among the French Physiocrats and received a first assessment in Quesnay's *Tableau Economique*, where agriculture is considered to be the only activity capable of creating a surplus, over and above the commodities used in production, as opposed to manufacture, which is believed to transform simply what is already in existence. The concept of net product is at the basis of what is now known as the (classical) surplus approach to economics. The structure of this approach emerges in its bare essentials in Ricardo (1815) where corn is assumed to be the only input and output of the economy. The net product of this economy is then simply the difference between total corn production and the amount of corn advanced as subsistence wages and as means of production.

The notion of a net product immediately leads to what Ricardo considered the fundamental problem of economics: the explanation of the laws of distribution. How are all the different conflicting claims on the net product of the nation eventually composed? Ricardo's answer is that profits emerge at the end as a residual after rents have been determined according to the decreasing fertility of land. The average rate of profit may then be calculated as a physical ratio of quantities of corn.

The extension of the Ricardian corn model to a multi-commodity system has remained for more than a century an unsolved problem in the history of economic thought. It is only after Sraffa (1960) presented his scheme where commodities are produced by means of commodities and labour that the seminal and far reaching approach of Ricardo emerged clearly.

Sraffa's scheme is based on the same vision of production as a circular process able to reproduce all the commodities used in production and to provide the net product as a surplus like the original Ricardian model, but it includes also all the industrial interdependences of modern economies. In this way Sraffa is able to define the net product and to deal with its distribution following the same logical steps of Ricardo.

Let us consider, for simplicity's sake, only the case of a system of single-product industries (but Sraffa analyses also fixed capital, non-produced means of production and general joint production). Using matrix notation and calling A the square matrix of physical commodity inputs and B the diagonal matrix of commodity productions, the (column) vector of net product y is then defined as

$$y = (B - A)s$$

where s is the (column) sum vector. Provided total wages are exogenously given in physical terms as a vector w, profits may still be defined as the residual vector

$$p = y - w.$$

But, of course, in the general case, unless the system happens to be in its standard proportions, the average rate of profit cannot be calculated in physical terms. Sraffa shows that its correct determination may only be obtained by solving simultaneously a new system of prices that replace and generalize the Ricardian labour theory of value.

See Also

▶ Produit Net

Bibliography

Ricardo, D. 1815. An essay on the influence of a low price of corn on the profits of stock. In *The works and correspondence of David Ricardo*, vol. IV, ed. P. Sraffa. Cambridge: Cambridge University Press, 1951.

Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*. Oxford: Clarendon Press, 1976.

Sraffa, P. 1960. *Production of commodities by means of commodities: Prelude to a critique of economic theory.* Cambridge: Cambridge University Press.

Network Formation

Matthew O. Jackson

Abstract

A brief introduction and overview of models of the formation of networks is given, with a focus on two types of model. The first views networks as arising stochastically, and uses random graph theory, while the second views the links in a network as social or economic relationships chosen by the involved parties, and uses game theoretic reasoning.

Keywords

Clustering; Degree distributions; Graph theory; Myerson value; Network formation; Pairwise stability; Random graphs; Small worlds

JEL Classifications

D85

A growing literature in economics examines the formation of networks and complements a rich literature in sociology and recently emerging literatures in computer science and statistical physics. Research on network formation is generally motivated by the observation that social structure is important in a wide range of interactions, including the buying and selling of many goods and services, the transmission of job information, decisions on whether to undertake criminal activity, and informal insurance networks.

Networks are often modelled using tools and terminology from graph theory. Most models of networks view a network as either a non-directed or a directed graph; which type of graph is more appropriate depends on the context. For instance, if a network is a social network of people and links represent friendships or acquaintances, then it would tend to be non-directed. Here the people would be modelled as the nodes of the network and the relationships would be the links. (In terms of a graph, the people would be vertices and the relationships would be edges.) If, instead, the network represents citations from one article to another, then each article would be a node and the links would be directed, as one article could cite another. While many social and economic relationships are reciprocal or require the consent of both parties, there are also enough applications that take a directed form, so that both non-directed and directed graphs are useful as modelling tools.

Models of how networks form can be roughly divided into two classes. One derives from random graph theory, and views an economic or social relationship as a random variable. The other views the people (or firms or other actors involved) as exercising discretion in forming their relationships, and uses game theoretic tools to model formation. Each of these techniques is discussed in turn.

Models of Random Networks

Bernoulli Random Graphs

Some of the earliest formal models used to understand the formation of networks are random graphs: the canonical example is that of a pure Bernoulli process of link formation (for example, see the seminal study of Erdös and Rényi 1960). For instance, consider a network where the (non-directed) link between any two nodes is formed with some probability p (where $1 > p > 0$), and this process occurs independently across pairs of nodes. While such a random method of forming links allows any network to potentially emerge, some networks are much more likely to do so than others. Moreover, as the number of nodes becomes large, there is much that can be deduced about the structure the network is likely to take, as a function of p. For instance, one can examine the probability that the resulting network will be connected in the sense that one can find a path (sequence of links) leading from any given node to any other node. We can also ask what the average distance will be in terms of path length between different nodes, among other things. As Erdös and Rényi showed, such a random graph exhibits a number of ‘phase’ transitions as the probability of forming links, p, is varied in relation to the number of nodes, n; that is, resulting networks exhibit different characteristics depending on the relative sizes of p and n.

Whether or not such a uniformly random graph model is a good fit as a model of network formation, it is of interest because it indicates that networks with different densities of links might tend to have very different structures and also provides some comparisons for network formation processes more generally. Some of the basic properties that such a random graph exhibits can be summarized as follows. When p is small in relation to n, so that $p < 1/n$ (that is, the average number of links per node is less than one), then with a probability approaching 1 as n grows the resulting graph consists of a number of disjointed and relatively small components, each of which has a tree-like structure. (A component of a network is a subgraph, so that each node in the subgraph can be reached from any other node in the subgraph via a path that lies entirely in the subgraph, and there are no links between any nodes in the subgraph and any nodes outside the subgraph.) Once p is large enough in relation to n, so that $p > 1/n$, then a single ‘giant component’ emerges; that is, with a probability approaching 1 the graph consists of one large component, which contains a nontrivial fraction of the nodes, and all other components are vanishingly small in comparison. Why there is just one giant component and all other components are of a much smaller order is fairly intuitive. In order to have two ‘large’ components each having a nontrivial fraction of n nodes, there would have to be no links between any node in one of the components and any node in the other. For large *n*, it becomes increasingly unlikely to have two large components with absolutely no links between them. Thus, nontrivial components mesh into a giant component, and any other components must be of a much smaller order. As *p* is increased further, there is another phase transition when *p* is proportional to $log(n)/n$. This is the threshold at which the network becomes 'connected' so that all nodes are path-connected to each other and the network consists of a single component. Once we hit the threshold at which the network becomes connected, we also see further changes in the diameter of the network as we continue to increase *p* relative to *n*. (The diameter is the maximal distance between two nodes, where distance is the minimal number of links that are needed to pass from one node to another.) Below the threshold, the diameter of a giant component is of the order of $log(n)$, then at the threshold of connectedness it hits $log(n)/loglog(n)$, and it continues to shrink as *p* increases.

Similar properties and phase transitions have been studied in the context of other models of random graphs. For example, Molloy and Reed (1995), among others (see Newman 2003), have studied component size and connectedness in a 'configuration model'. There, a set of nodes is given together with the number of links that each node should have, and then links are randomly formed to leave each node with the pre-specified number of links.

Clustering and Markov Graphs

Although the random graphs of Erdös and Rényi are a useful starting point for modelling network formation, they lack many characteristics observed in most social and economic networks. This has led to a series of richer random graph-based models of networks. The most basic property that is absent from such random networks is that the presence of links tends to be correlated. For instance, social networks tend to exhibit significant clustering. Clustering refers to the following property of a network. If we examine triples of nodes so that two of them are each connected to the third, what is the frequency with which those two nodes are linked to each other? This tends to be much larger in real social networks than one would see in a Bernoulli random graph. On an intuitive level, models of network formation where links are formed independently tend to look too much like 'trees', while observed social and economic networks tend to exhibit substantial clustering, with many more cycles than would be generated at random (see Watts 1999, for discussion and evidence).

Frank and Strauss (1986) identified a class of random graphs that generalize Bernoulli random graphs, which they called 'Markov graphs' (also referred to as p^* networks). Their idea was to allow the chance that a given link forms to be dependent on whether or not neighbouring links are formed. Specific interdependencies require special structures, because, for instance, making one link dependent on a second, and the second on the third, can imply some interdependencies between the first and third. These sorts of dependencies are difficult to analyse in a tractable manner, but nevertheless some special versions of such models have been useful in statistical estimation of networks.

Small Worlds

Another variation on a Bernoulli network was explored by Watts and Strogatz (1998) in order to generate networks that exhibit both relatively low distances (in terms of minimum path length) between nodes and relatively high clustering – two features that are present in many observed networks but not in the Bernoulli random graphs unless the number of links per node (p (n − 1)) is extremely high. They started with a very structured network that exhibits a high degree of clustering. Then, by randomly rewiring enough (but not too many) links, one ends up with a network that has a small average distance between links but still has substantial clustering. While such a rewiring process results in networks that exhibit some of the features of social networks, it leads to networks that miss out on other basic characteristics that are present in many social networks. For example, the nodes of such a network tend to be too similar in terms of the number of links that they each have.

Degree Distributions

One fundamental characteristic of a social network is a network's degree distribution. The degree of a node is the number of links it has, and the degree distribution keeps track of how varied the degree is across the nodes of the network. That is, the degree distribution is simply the frequency distribution of degrees across nodes. For instance, in a friendship network some individuals might have only a few friends while other individuals might have many, and then the degree distribution quantifies this information.

Price (1965) examined a network of citations (between scientific articles), and found that the degree distribution exhibited 'fat tails' compared with what one would observe in a Bernoulli random graph; that is, there was a higher frequency of articles that had many citations and a higher frequency of articles that had no citations than should be observed if citations were generated independently. In fact, many social networks exhibit such fat tails, and some have even been thought to exhibit what is known as a 'scale-free' degree distribution or said to 'follow a power law'. A scale-free distribution is one where the frequency of degrees can be written in the form $f(d) = ad^{-b}$, for some parameters a and b, where d is the degree and $f(d)$ is the relative frequency of nodes with degree d. Such distributions date to Pareto (1896), and have been observed in a variety of other contexts ranging from the distribution of wealth in a society to the relative use of words in a language. Price (1976) adapted ideas from Simon (1955) to develop a random link formation process that produces networks with such degree distributions. A similar model was later studied by Barabási and Albert (2001), who called the process of link formation 'preferential attachment'. The idea is that nodes gain new links with probabilities that are proportional to the number of links they already have (which is closely related to a lognormal growth process). In a system where new nodes are born over time, this process generates scale-free degree distributions.

A simple preferential attachment model also has its limitations. One is that most social networks do not in fact have degree distributions that are scale-free. Observed degree distributions tend to lie somewhere between the extremes of a scale-free distribution and that corresponding to an independent Bernoulli random graph (sometimes known as a Poisson random graph for its approximate degree distribution). Second, the preferential attachment model fails to produce the type of clustering observed in many social networks, just as Bernoulli random graphs do. This has led to the construction of hybrid models that allow for richer sets of degree distributions, as well as clustering and correlation in degrees, and allows for the structural fitting of random graph based network formation models to data (for example, see Jackson and Rogers 2007, and the discussion there).

Strategic Models of Network Formation

Strategic models of network formation have emerged from the economics literature, and offer a very different perspective from that seen in random graph models, and a complementary set of insights (see Jackson 2006, for comparison and discussion). The starting point for a game theoretic approach is to assume that the nodes are active discretionary agents or players who get payoffs that depend on the social network that emerges. For example, if nodes are countries and links are political alliances, or nodes are firms and links are trading or collaboration agreements, then the relationships are entered into with some care and thought. Even in modelling something like a friendship network, while individuals might not be directly calculating costs and benefits from the relationship, they do react to how enjoyable or worthwhile the relationship is and might tend to spend more effort or time in relationships that are more beneficial and avoid ones that are less so. Different social networks lead to different outcomes for the involved agents (for example, different trades, different access to information or favours, and so on). Links are then formed at the discretion of the agents, and various equilibrium notions are used to predict which networks will form. This differs from the random models not only in that links result as a function of decisions rather than at random, but also in that there are natural costs and benefits associated with networks which then allow a welfare analysis.

Some of the first models to bring explicit utilities and choice to the formation of social links were in the context of modelling the trade-offs between ‘strong’ and ‘weak’ ties (links) in labour contact networks. Such models by Boorman (1975) and Montgomery (1991) explored a theory, due to Granovetter (1973), about different strengths of social relationships and their role in finding employment. Granovetter observed that when individuals obtained jobs through their social contacts, while they sometimes did so through strong ties (people whom they knew well and interacted with on a frequent basis), they also quite often obtained jobs through weak ties (acquaintances whom they knew less well and/or interacted with relatively infrequently). This led Granovetter to coin the phrase ‘the strength of weak ties’. Boorman’s article and Montgomery’s articles provided explicit models where costs and benefits could be assigned to strong and weak ties, and trade-offs between them could be explored.

In a very different setting, another use of utility functions involving networks emerged in the work of Myerson (1977). Myerson analysed a class of cooperative games that were augmented with a graph structure. In these games the only coalitions that could produce value are those that are pathwise connected by the graph, and so such graphs indicate the possible cooperation or communication structures. This approach led Myerson to characterize a variation on the Shapley value, now called the Myerson value, which was a cooperative game solution concept for the class of cooperative games where constraints on coalitions were imposed by a graph structure. Although the graphs in Myerson’s analysis are tools to define a special class of cooperative games, they allow the graph structure to influence the allocation of societal value among a set of players. Aumann and Myerson (1988), recognizing that different graph structures led to different allocations of value, used this to study a game where the graph structure was endogenous. They studied an extensive form game where links are considered one by one according to some exogenous order, and formed if both agents involved agree. While that game turns out to be hard to analyse even in three-person examples, it was an important precursor to the more recent economic literature on network formation.

In contrast to the cooperative game setting, Jackson and Wolinsky (1996) explicitly considered networks, rather than coalitions, as the primitive. Thus, rather than deducing utilities indirectly through a cooperative game on a graph, they posited that networks were the primitive structure and agents derived utilities based on the network structure in place. So, once a social network structure is in place, one can then deduce what the agent’s payoffs will be. Using such a formulation where players’ payoffs are determined as a function of the social network in place, it is easy to model network formation using game theoretic techniques.

Pairwise Stability

In modelling network formation from a game theoretic perspective, one needs to have some notion of equilibrium or stable networks. Since it is natural to require mutual consent in many applications, standard Nash equilibrium based ideas are not very useful. For instance, consider a game where each agent simultaneously announces which other agents he or she is willing to link to. It is always a Nash equilibrium for each agent to say that he or she does not want to form any links, anticipating that the others will do the same. Generally, this allows for a multiplicity of equilibria, many of which make little sense from a social network perspective. Even equilibrium refinements (such as undominated Nash or perfect equilibrium) do not avoid this problem. Given that it is natural in a network setting for the agents prospectively forming a link to be able to communicate with each other, they should also be able to coordinate with each other on the forming of a link. An approach taken by Jackson and Wolinsky (1996) is to define a stability notion that directly incorporates the mutual consent needed to form links. Jackson and Wolinsky (1996) defined the following notion of ‘pairwise stability’: a network is pairwise stable if (i) no player would be better off if he or she severed one of his or her links, and (ii) no pair of players would both benefit (with at least one of the pair seeing a strict benefit) from adding a link that is not in the network. The requirement that no player wishes to delete a link that he or she is involved in implies that a player has the discretion to unilaterally terminate relationships that he or she is involved in. The second part of the definition captures the idea that if we are at a network where the creation of a new link would benefit both players involved, then the network g is not stable, as it will be in the players' interests to add the link.

Pairwise stability is a fairly permissive stability concept – for instance, it does not consider deviations where players delete some links and add others at the same time. While pairwise stability is easy to work with and often makes fairly pointed predictions, the consideration of further refinements can make a difference. A variety of refinements and alternative notions have been introduced, including allowing agents to form and sever links at the same time, allowing coalitions of agents to add and sever links in a coordinated fashion, or behaviour where agents anticipate how the formation of one link might influence others to form further links (see Jackson 2004, for discussion and references). There are also dynamic models (for example, Watts 2001) in which the possibility of forming links arises (repeatedly) over time, and agents might 'tremble' when they form links (see Jackson 2004, for references). These various equilibrium/stability concepts have different properties and are appropriate in different contexts.

With pairwise stability, or some other solution in hand, one can address a series of questions. One fundamental question is whether, from society's point of view, efficient or optimal networks will be stable when agents form links with their selfish interests in mind. Given that transfers are being considered here, one natural definition of an 'efficient' or 'optimal' network is one that maximizes the total value or the sum of utilities of all agents in the society. Another basic question is to ask whether in situations where no efficient network is pairwise stable, is it possible for some sort of intervention (for example, in the form of taxing or subsidizing links), to lead efficient networks to form.

A Connections Model of Social Networks

One stylized example from Jackson and Wolinsky (1996) gives some feeling for the issues involved in the above questions and is useful for illustrating the relationship between efficient and pairwise stable networks. Jackson and Wolinsky called this example the 'symmetric connections model', in which the links represent social relationships between players such as friendships. These relationships offer benefits in terms of favours, information, and so on, and also involve some costs. Moreover, players benefit from having indirect relationships. A 'friend of a friend' produces benefits or utility for a player, although of a lesser value than the direct benefits that come from a 'friend'. The same is true of 'friends of a friend of a friend', and so forth. Benefit deteriorates in the 'distance' of the relationship, as represented by a factor δ between 0 and 1, which indicates the benefit from a direct relationship between two agents and is raised to higher powers for more distant relationships. For instance, in the network where player 1 is linked to 2, 2 is linked to 3, and 3 is linked to 4; player 1 gets a benefit of δ from the direct connection with player 2, an indirect benefit of δ^2 from the indirect connection with player 3, and an indirect benefit of δ^3 from the indirect connection with player 4. For $\delta < 1$ this leads to a lower benefit from an indirect connection than a direct one. Players also pay some cost c for maintaining each of their direct relationships (but not for indirect ones). Once the benefit parameter, δ, and the cost parameter, $c > 0$ are specified, it is possible to determine each agent's payoff from every possible network, allowing a characterization of the pairwise stable networks as well as the efficient networks. The efficient network structures are the complete network if $c < \delta - \delta^2$, a 'star' (a network where one agent is connected to each other agent and there are no other connections) encompassing all nodes if $\delta - \delta^2 < c < \delta + \frac{(n-2)}{2}\delta^2$, and the empty network if $\delta + \frac{(n-2)}{2}\delta^2 < c$. The idea is that if costs are very low it will be efficient to include all links in the network, because shortening any path leads to higher payoffs. When the link cost is at an intermediate level, then the unique efficient network structure is to have all players arranged in a star network, since such a structure has the minimal number of links ($n - 1$) needed to connect all individuals, and yet still has all nodes within at most two links from one another. Once links become so costly that a star results in more cost than benefit, then the empty network is efficient. One can also examine a directed version of such a model, as in Bala and Goyal (2000), who find related results, but with some differences that depend on whether both agents or just one of the agents enjoys the benefits from a directed link.

Inefficiency of Stable Networks

The set of pairwise stable networks does not always coincide with the efficient ones, and sometimes do not even intersect with the set of efficient networks. For instance, if the cost of a link is greater than the direct benefit ($c > \delta$), then relationships are only valuable to a given agent if they generate indirect benefits as well as direct ones. In such a situation a star is not pairwise stable since the centre player gets benefit of the direct value from each of his or her links, which is less than the cost of each of those links. This model of social networks makes it obvious that there will be situations where individual incentives are not aligned with overall societal benefits.

As it will generally be the case that in economic and social networks there are some sort of externalities present, since two agents' decisions of whether or not to form a relationship can affect the well-being of other agents, one should expect that there will be situations where the networks formed through the selfish decisions of the agents do not coincide with those that are efficient from society's perspective. In such situations, it is natural to ask whether intervention in the form of transfers among agents might help align individual and overall societal incentives to form the right network. For instance, in the connections model, it would make sense to have the peripheral agents in a star pay the centre of the star in order to maintain their links. The peripheral agents benefit much more from the relationship with the centre agent than vice versa, as the centre agent provides access to many indirect agents. Although a simple set of transfers can align individual and overall incentives in the connections model, it is impossible to always correct this tension between individual incentives and overall efficiency by taxing and subsidizing agents for the links they form (even in a complete information setting). The fact that there are very simple, natural network settings where no 'reasonable' set of transfers can help rectify the disparity stability and efficiency was shown in Jackson and Wolinsky (1996). Without providing details, the impossibility of reconciling stability and efficiency stems from the following considerations: from any given network, there are many other networks that can be reached. In fact, if there are n nodes, then there are $n\,(n - 1)/2$ possible links that can be added to or deleted from any given network. In order to ensure that a given efficient network is pairwise stable, payoffs to all neighbouring networks have to be configured so that no agent finds it in his or her interest to delete a link and no two agents find it in their interests to add a link. It is impossible to assign all the necessary taxes and subsidies in such a way that (i) the transfers are feasible (and are not given to unattached agents), (ii) identical agents are treated identically, and (iii) it is always the case that at least one efficient network is pairwise stable.

Much more has been learned about the relationship between stable and efficient networks and possible transfers to ensure that efficient networks form. For instance, one can characterize some classes of settings where the efficient networks and the stable ones coincide (see Jackson and Wolinsky 1996). One can also design transfers that ensure that some efficient network is stable by treating agents unequally (for example, taxing or subsidizing them differently even though the agents are identical in the problem as shown by Dutta and Mutuswami 1997). Another important point was made by Currarini and Morelli (2000), who showed that if agents bargain over the division of payoffs generated by network relationships at the time when they form link, then in a nontrivial class of settings equilibrium networks are efficient. While the conclusions hinge on the structure of the link-formation-bargaining game, and in particular on an asymmetry in bargaining power across the agents, such a result tells us that it can be important to model the formation of the links of a network together with any potential bargaining over payoffs or transfers. Further study in this area shows how the types of transfers needed to reach efficient networks relate to the types of network externalities that are present in the setting.

Small Worlds and Strategic Network Formation

Beyond understanding the relationship between stable and efficient networks, strategic models of network formation have also shed light on some empirical regularities and helped predict which networks will arise in settings of particular interest. For instance, strategic models of network formation provide substantial insight into the 'small-worlds' properties of social networks: the simultaneous presence of high clustering (a high density of links on a local level) and short average path length between nodes (see Jackson 2006, for references). The reasoning is based on a premise that different nodes have different distances from each other, either geographically or according to some other characteristic, such as profession, tastes, and so on. The low cost of forming links to other nodes that are nearby then naturally explains high clustering. High benefits from forming links that bridge disparate parts of the network, due to the access and indirect connections that they bring, naturally explain low average path length.

Networks and Markets

There is a rich set of studies of markets and networks from an economics perspective, including models that explicitly examine whether or not buyers and sellers have incentives to form an efficient network of relationships (for example, Kranton and Minehart 2001). The incentives to form efficient networks depend on the setting and which agents bear the cost of forming relationships. In some settings competitive forces lead to the right configuration of links, and in others buyers and sellers over-connect in order to improve their relative bargaining positions. Other studies focus on the context of specific markets, such as labour markets, where people benefit from connections with neighbours who provide information about job opportunities (see Ioannides and Loury 2004, for an overview and references).

In addition to studies of networks of relationships between buyers and sellers, firms also form relationships amongst themselves that affect their costs and the sets of products they offer. Such oligopoly settings where network formation is important (see Bloch 2004, for a recent survey), again provide a rich set of results regarding the structure of networks that emerge, and contrasts between settings where efficient networks naturally emerge and others where only inefficient networks are formed.

Network formation has also been studied in the context of many other applications, including risk-sharing in developing countries, social mobility, criminal activity, international trade and banking deposits.

Finally, there have been a number of experiments on network formation, using human subjects. These examine a variety of questions, ranging from how forward-looking agents are when they form social ties, to whether or not agents overcome coordination problems when forming links, to whether there are pronounced differences between network formation when links can be formed unilaterally as opposed to when they require mutual consent, to whether efficient networks will tend to result and how that depends on symmetries or asymmetries in the efficient network structure (see Falk and Kosfeld 2003, for some discussion and references).

See Also

- ▶ Business Networks
- ▶ Learning and Information Aggregation in Networks
- ▶ Mathematics of Networks
- ▶ Power Laws
- ▶ Psychology of Social Networks
- ▶ Social Networks in Labour Markets

Bibliography

Aumann, R., and R. Myerson. 1988. Endogenous formation of links between players and coalitions: an application of the Shapley value. In *The shapley value*, ed. A. Roth. Cambridge: Cambridge University Press.

Bala, V., and S. Goyal. 2000. A non-cooperative model of network formation. *Econometrica* 68: 1181–1230.

Barabási, A., and R. Albert. 1999. Emergence of scaling in random networks. *Science* 286: 509–512.

Bloch, F. 2004. Group and network formation in industrial organization: a survey. In *Group formation in economics; networks, clubs and coalitions*, ed. G. Demange and M. Wooders. Cambridge: Cambridge University Press.

Bollobás, B. 2001. *Random graphs*. 2nd ed. Cambridge: Cambridge University Press.

Boorman, S. 1975. A combinatorial optimization model for transmission of job information through contact networks. *Bell Journal of Economics* 6: 216–249.

Currarini, S., and M. Morelli. 2000. Network formation with sequential demands. *Review of Economic Design* 5: 229–250.

Dutta, B., and S. Mutuswami. 1997. Stable networks. *Journal of Economic Theory* 76: 322–344.

Erdös, P., and A. Rényi. 1960. On the evolution of random graphs. *Publication of the Mathematical Institute of the Hungarian Academy of Sciences* 5: 17–61.

Falk, A., and M. Kosfeld. 2003. *It's all about connections: Evidence on network formation*. Mimeo: University of Zurich.

Frank, O., and D. Strauss. 1986. Markov graphs. *Journal of the American Statistical Association* 81: 832–842.

Granovetter, M. 1973. The strength of weak ties. *American Journal of Sociology* 78: 1360–1380.

Ioannides, Y.M., and L.D. Loury. 2004. Job information networks, neighborhood effects and inequality. *Journal of Economic Literature* 42: 1056–1093.

Jackson, M.O. 2004. A survey of models of network formation: stability and efficiency. In *Group formation in economics: Networks, clubs and coalitions*, ed. G. Demange and M. Wooders. Cambridge: Cambridge University Press.

Jackson, M.O. 2006. The economics of social networks. In *Chapter 1, volume 1 in Advances in economics and econometrics, theory and applications: ninth world congress of the econometric society*, ed. R. Blundell, W. Newey, and T. Persson. Cambridge: Cambridge University Press.

Jackson, M.O., and B.W. Rogers. 2007. Meeting strangers and friends of friends: how random are socially generated networks? *American Economic Review* 97: 890–915.

Jackson, M.O., and A. Wolinsky. 1996. A strategic model of social and economic networks. *Journal of Economic Theory* 71: 44–74.

Kranton, R., and D. Minehart. 2001. A theory of buyer–seller networks. *American Economic Review* 91: 485–508.

Molloy, M., and B. Reed. 1995. A critical point for random graphs with a given degree sequence. *Random Structures & Algorithms* 6: 161–179.

Montgomery, J. 1991. Social networks and labor market outcomes. *American Economic Review* 81: 1408–1418.

Myerson, R. 1977. Graphs and cooperation in games. *Math Operations Research* 2: 225–229.

Newman, M. 2003. The structure and function of complex networks. *SIAM Review* 45: 167–256.

Page, F., M. Wooders, and S. Kamat. 2005. Networks and farsighted stability. *Journal of Economic Theory* 120: 257–269.

Price, D.J.S. 1965. Networks of scientific papers. *Science* 149: 510–515.

Price, D.J.S. 1976. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science* 27: 292–306.

Simon, H. 1955. On a class of skew distribution functions. *Biometrika* 42: 425–440.

Watts, A. 2001. A dynamic model of network formation. *Games and Economic Behavior* 34: 331–341.

Watts, D.J. 1999. *Small Worlds: The dynamics of networks between order and randomness*. Princeton: Princeton University Press.

Watts, D.J., and S. Strogatz. 1998. Collective dynamics of 'small-world' networks. *Nature* 393: 440–442.

Network Goods (Empirical Studies)

Neil Gandal

Abstract

A network effect exists if the consumption benefits of a good or service increase with the total number of consumers who purchase compatible products. A growing empirical literature examines technological adoption of products with network effects. The early literature mainly addressed the question of whether network effects are indeed significant; this work typically employed reduced form models. Later literature employed structural methodology, which can address aspects of firm strategy, such as incentives to provide compatible products. Key issues in the empirical work on network industries are examined.

Keywords

Hedonic prices; Logit models of individual choice; Network effect (empirical studies); Product differentiation

JEL Classifications

L13

A network effect exists if the consumption benefits of a good or service increase with the total number of consumers who purchase compatible products. The literature distinguishes between direct and indirect network effects.

In the case of a direct (or physical) network effect, an increase in the number of consumers on the same network raises the consumption benefits for everyone on the network. Communication networks such as telephone and e-mail networks are examples of goods with direct network effects.

A network effect can also arise in a setting with a ‘hardware/software’ system. Here, the benefits of the hardware good increase when the variety of compatible software increases. An indirect (or virtual) network effect arises endogenously in this case because an increase in the number of users of compatible hardware increases the demand for compatible software. Since software goods are typically characterized by economies of scale, the increase in demand leads to increases in the supply of software varieties. Examples of settings where virtual network effects arise include consumer electronics such as CD players and compact discs, computer operating systems and applications programs, and television sets and programming.

Given the dramatic growth of the internet and information technology industries, and the importance of interconnection in these networks, it is not surprising that there is a large theoretical literature on competition in industries with network goods. Important questions in this literature include

- the examination of the private and social incentives to attain compatibility;
- the trade-off between standardization and variety;
- modelling the dynamics of competition between competing networks; and
- how the private and social choice among competing incompatible networks differs when there are both early and late adopters.

See Farrell and Klemperer (2007) for further discussion.

Although relatively small, a growing empirical literature has developed to examine technological adoption of products with network effects. In this short article, I briefly discuss this literature. The empirical work can be organized by the issues addressed and the methodology employed. The primary issue addressed by the early literature is whether network effects are indeed significant; this work typically employed reduced form models. The article first surveys early work in this genre, then examines papers that employed structural methodology. The main advantage of this methodology is that it can address aspects of firm strategy, such as incentives to provide compatible products. The article closes by examining key issues in empirical work on network industries.

Early Work: Indirect Evidence of Network Effects

Greenstein (1993), Gandal (1994, 1995), and Saloner and Shepard (1995) provide early evidence that the value of the ‘hardware’ good depends on the variety of compatible complementary software. (Shy 2001, surveys many of the empirical papers discussed in this article in greater detail than space permits here.)

Software for the IBM 1400 mainframe could not run on succeeding generations of IBM mainframes while software for the IBM 360 could run on succeeding models. Greenstein (1993) finds that, other things being equal, a firm with an IBM 1400 was no more likely than any other firm to purchase an IBM mainframe when making a future purchase. On the other hand, a firm with an IBM 360 was more likely to purchase an IBM mainframe than a firm that did not own an IBM 360. This result can be interpreted as a demand for compatible software.

Gandal (1994) estimates hedonic (quality-adjusted) price equations for spreadsheets to examine whether spreadsheet programs that were compatible with Lotus – the de facto standard – command a premium. The results – that consumers place a positive value on compatibility – suggest (a) direct network effects because people want to share files and (b) indirect network effects because compatible software enables the transfer of data among a variety of software programs. Gandal (1995) extends the analysis to database management software (DMS) and multiple standards and finds that only the Lotus file compatibility standard is significant in explaining price variations, suggesting that indirect network effects are important in the DMS market.

Saloner and Shepard (1995) test for network effects in the automated teller machine (ATM) industry. In particular, they test whether banks with a larger expected number of ATM locations will adopt the ATM technology sooner. Since expected network size is not an observable variable, they use the number of branches as a proxy. The results suggest that banks with more branches will adopt earlier, which is consistent with virtual network effects.

Structural Models: Explicitly Modelling the Complementary Goods Market

Because hedonic price equations are a reduced form, rather than a structural model, parameter estimates associated with compatibility in Gandal (1994, 1995) may be capturing demand effects or supply effects or some combination of both. In other words, are consumers really willing to pay a premium for compatibility or is the marginal cost of compatibility relatively high? In the case of software, fixed costs of providing characteristics are quite significant, while marginal production costs associated with the characteristics are typically very small; they primarily include duplication of digital material. Hence, in these papers the estimated hedonic price coefficients on compatibility indeed measure consumer willingness to pay for compatibility.

Nevertheless, reduced form models are not suitable for examining business strategies or conducting counterfactuals. Gandal et al. (2000) develop a dynamic structural model of consumer adoption and software entry, and use the model to estimate the feedback from hardware to software and vice versa in the CD industry. The advantage of the structural methodology is that it enables researchers to assess business strategies as well as examine conduct counterfactuals. In the case of business strategies, Gandal et al. (2000) show that a five per cent reduction in price would have had the same effect as a ten per cent increase in CD variety in terms of increasing sales of CD players. They also show that, if it had been possible to make CD players compatible with LPs, compatibility could have accelerated the adoption process by more than a year. This is just a ‘thought experiment’ for CD players, but it has policy relevance for other systems like HDTV.

Rysman (2004) develops a structural model to examine the importance of network effects in the market for Yellow Pages. The model includes a consumer adoption equation, advertiser demand for space, and a firm’s profit maximizing behaviour. He finds that consumers value advertising and advertisers value consumer adoption, suggesting virtual network effects.

In several recent papers, advances in the estimation of discrete choice models of product differentiation – see Berry (1994) and Berry et al. (1995) – have also been employed when testing for indirect network effects in differentiated product markets. Ohashi and Clements (2005), for example, use a logit model to test for indirect network effects in the US video game market.

Key Issues in Empirical Work

As in most fields, empirical work is typically limited by the available data. A key problem exists when one tries to estimate network effects in homogeneous product industries using time series data. For many network industries, technological progress drives down prices and costs. Hence an increase in the number of users on a network might be due to a network effect or to falling prices (see Gowrisankaran and Stavins 2004, for further discussion). In order to estimate these effects, one must have additional data.

Gandal et al. (2000), for example, have data on the number of available compact disc titles at each point in time. Hence, in their model the two main effects that lead to greater adoption of CD players – lower prices of the hardware good and network effects due to increases in the number of titles – are measured separately. Nevertheless, that is only a start, since both of these variables are typically endogenous. Identification in Gandal et al. (2000) was possible only because there were data on the fixed costs of entering the CD production industry over time. These data were used as an instrument for CD (title) availability. Additionally, case studies indicated that the CD player industry was quite competitive, leading the authors to assume that the price of CD players was exogenous. Without both of these assumptions, it would not have been possible to identify the model.

Additionally, there is the thorny issue of pricing in dynamic models of competition in network industries. Since hardware firms may want to subsidize early adopters in order to build up a network advantage and then (perhaps) charge a higher price when the installed base grows, pricing issues are dynamic; firms will take into account (current and expected future) network size when choosing their prices. Park (2004) develops a dynamic structural model of competition in an oligopolistic market with network effects that addresses the dynamic pricing issues; he then estimates the model for VCRs. To the best of my knowledge, this is the only empirical paper that deals explicitly with dynamic pricing issues.

A similar issue arises in dynamic models of competition in network industries when firms make investment in quality over time. Markovich (2001) examines the trade-off between standardization and variety in a dynamic setting using numerical methods. With suitable data one might be able to use her framework to empirically examine investment incentives and pricing decisions in a dynamic setting with network effects.

Finally, there is a budding empirical literature on standardization via committees. Papers include Simcoe (2006), who examines the standardization process in various committees of the Internet Engineering Task Force, and Gandal et al. (2006), who examine firms' incentives to participate in Telecommunication Industry Association standardization meetings.

See Also

- ▶ Hedonic Prices
- ▶ Network Goods (Theory)

Bibliography

Berry, S. 1994. Estimating discrete-choice models of product differentiation. *RAND Journal of Economics* 25: 334–347.

Berry, S., J. Levinsohn, and A. Pakes. 1995. Automobile prices in market equilibrium. *Econometrica* 63: 841–890.

Farrell, J., and P. Klemperer. 2007. Coordination and lock-in: Competition with switching costs and network effects. In *Handbook of industrial organization*, vol. 3, ed. M. Armstrong and R. Porter. Amsterdam: North-Holland.

Gandal, N. 1994. Hedonic price indexes for spreadsheets and an empirical test for network externalities. *RAND Journal of Economics* 25: 160–170.

Gandal, N. 1995. A Selective survey of the literature on indirect network externalities. *Research in Law and Economics* 17: 23–31.

Gandal, N., M. Kende, and R. Rob. 2000. The dynamics of technological adoption in hardware/software systems: The case of compact disc players. *RAND Journal of Economics* 31: 43–61.

Gandal, N., N. Gantman, and D. Genesove. 2006. Intellectual property and standardization committee participation in the U.S. modem industry. In *Standards and public policy*, ed. S. Greenstein and V. Stango. Cambridge: Cambridge University Press.

Gowrisankaran, G., and J. Stavins. 2004. Network externalities and technology adoption: Lessons from electronic payments. *RAND Journal of Economics* 35: 260–276.

Greenstein, S. 1993. Did installed base give an incumbent any (measurable) advantages in federal computer procurement? *RAND Journal of Economics* 24: 19–39.

Markovich, S. 2001. *Snowball: The evolution of dynamic markets with network externalities*. Mimeo: Tel Aviv University.

Ohashi, H., and M. Clements. 2005. Indirect network effects and the product cycle: Video games in the U.S., 1994–2002. *Journal of Industrial Economics* 53: 515–542.

Park, S. 2004. Quantitative analysis of network externalities in competing technologies: The VCR case. *Review of Economics and Statistics* 86: 937–945.

Rysman, M. 2004. Competition between networks: A study of the market for Yellow Pages. *Review of Economic Studies* 71: 483–512.

Saloner, G., and A. Shepard. 1995. Adoption of technologies with network externalities: An empirical examination of the adoption of automated teller machines. *RAND Journal of Economics* 26: 479–501.

Shy, O. 2001. *The economics of network industries*. Cambridge: Cambridge University Press.

Simcoe, T. 2006. Committees and the creation of technical standards. In *Standards and public policy*, ed. S. Greenstein and V. Stango. Cambridge: Cambridge University Press.

Network Goods (Theory)

Paul Klemperer

Abstract

Network effects arise where current users of a good gain when additional users adopt it (classic examples are telephones and faxes). The effects create multiple equilibria and fierce competition between incompatible networks; users' expectations are crucial in determining which network succeeds. Early choices, such as the QWERTY typewriter keyboard, lock in the market; new entry, especially against established networks with proprietary technology, is often nearly impossible. Incompatible networks can induce efficient 'competition for the market', but more often create biases and inefficiencies. Policymakers should scrutinize markets where firms deliberately choose incompatibility.

Keywords

Compatible products; Competition for the market; Competition policy; Coordination; Entry; Excess early power; Excess inertia; Excess momentum; Herding; Indirect network effects; Intellectual property; Lock-in; Market share; Microsoft; Multiple equilibria; Network effects; Network externality; Penetration pricing; Pre-announcements; Product variety; Proprietary technology; QWERTY; Standards; Switching costs; Tipping

JEL Classifications

L13

Direct network effects arise if each user's payoff from the adoption of a good, and his incentive to adopt it, increase as more others adopt it; that is, if adoption by different users is complementary. For example, telecommunications users gain directly from more widespread adoption, and telecommunications networks with more users are also more attractive to non-users contemplating adoption.

Indirect network effects arise if adoption is complementary because of its effect on a related market. For example, users of hardware may gain when other users join them, not because of any direct benefit, but because it encourages the provision of more and better software.

Extensive case studies and more formal econometric evidence document significant network effects in many areas including, for example, telecommunications, radio and television, computer hardware and software, applications software and operating systems (including Microsoft's), securities markets and exchanges (including Ebay), and credit cards (see, for example, Gabel 1991; Rohlfs 2001; Shy 2001; and the article on network goods (empirical studies) in this dictionary).

Usually adoption prices do not fully internalize the network effects, so there is a positive externality from adoption. A single network product therefore tends to be under-adopted at the margin – this issue was the main focus of the early literature (see, for example, Leibenstein 1950; Rohlfs 1974). However, if two networks compete, then adopting one network means not adopting the other, which dilutes or reverses the externality.

More interestingly – and what is the starting point for the more recent literature – network effects create incentives to 'herd' with others. In a static (simultaneous-adoption) game there are often multiple equilibria, so expectations are crucial, and self-fulfilling. Likewise, a dynamic (sequential-adoption) game exhibits positive feedback or 'tipping' – a network that looks like succeeding will *as a result* do so (see, for example, David 1985; Arthur 1989; Arthur and Rusczcynski 1992).

How well competition among incompatible networks works depends dramatically on how adopters form expectations and coordinate their choices. If adopters smoothly coordinate on the best deals, vendors face strong pressure to offer them. Competition may then be unusually fierce because all-or-nothing competition neutralizes horizontal differentiation – since adopters focus not on matching a product to their own tastes but on joining the expected winner.

However, coordination is not easy. With simultaneous adoption, adopters may fail to coordinate at all and 'splinter' among different networks, or may coordinate on a different equilibrium from the one that is best for them – for example, each adopter may expect others to choose a low-quality product because it is produced by a firm that was successful in the past. Furthermore, consensus standard-setting (informally or through standards organizations) can be painfully slow when different adopters prefer different coordinated outcomes (see Bulow and Klemperer 1999). Coordination through contingent contracts is possible in theory (see, for example, Dybvig and Spatt 1983; Segal 1999), but seems uncommon in practice.

When adoption is sequential, we see *early instability and later lock-in* (see, for example, Arthur 1989) – this corresponds to the multiple equilibria that arise with simultaneous adoption. Because early adoptions influence later ones, long-term behaviour is determined largely by early events, whether accidental or strategic. In theory, at least, fully sequential adoption achieves the efficient outcome if it is best for all adopters, but more generally early adopters' preferences count for more than later adopters': this is 'excess early power'. Note that 'excess early power' does not depend on 'excess inertia', that is, on incompatible transitions being too hard *given ex post* incompatibility. (Both 'excess inertia', and its opposite, 'excess momentum', are theoretically possible; see Farrell and Saloner 1985.)

Firms promoting incompatible networks compete to win the pivotal early adopters, and so achieve *ex post* dominance and monopoly rents. Strategies such as penetration pricing and pre-announcements (see, for example, Farrell and Saloner 1986) are common. History, and especially market share, matter because an installed base both directly means a firm offers more network benefits and boosts expectations about its future sales. Such 'Schumpeterian' competition 'for the market' can neutralize (or even overturn) excess early power if promoters of networks that will be more efficient later on set low penetration prices in anticipation of this (see Katz and Shapiro 1986a). More commonly, though, late developers struggle while networks that are preferred by early pivotal customers thrive.

So early preferences and early information are likely to be excessively important in determining long-term outcomes. For example, whether or not the Dvorak typewriter keyboard is really much better than QWERTY (as David 1985, contends), there clearly was a chance in the 1800s that a keyboard superior to QWERTY would later be developed, and it is not clear what could have persuaded early generations of typists to wait, or to adopt diverse keyboards, *if* that was socially desirable. So it seems unlikely that the market gave a very good test of whether or not waiting was efficient. (Liebowitz and Margolis 1990, and Liebowitz 2002, contest both the details of the QWERTY example and the claim that network effects are significant more generally, but at least the second view is probably a minority one.)

Despite the possibility of competition for the market passing *ex post* rents through to earlier buyers, incompatibility often reduces efficiency and harms consumers in several ways.

Incompatibility means that consumers are faced with either a segmented market with low network benefits, or – if the market does 'tip' all the way to one network – with reduced product variety and without the option value from the possibility that a currently inferior technology might later become superior. Product variety is more sustainable if niche products are compatible with the mainstream, and so don't force users to sacrifice network effects.

These direct costs of poor coordination by adopters may be exacerbated by weaker incentives for vendors to offer good deals. For example, if a firm like Microsoft is widely believed to have the ability to offer the highest quality, it may never bother to do so: the fact that everyone expects Microsoft to recapture the market if it ever lost any one cohort of customers (or lost any one cohort of providers of complementary products) means everyone rationally chooses Microsoft even if it never actually produces high quality or offers a low price (see Katz and Shapiro 1992).

Ex post rents are often not fully dissipated by *ex ante* competition, especially if expectations fail to track relative surplus. Worse, the rent dissipation that does occur may be wasteful, such as socially inefficient marketing. At best, *ex ante* competition induces 'bargain-then-rip-off' pricing (low to attract business, high to extract surplus) but this distorts buyers' quantity choices and gives them artificial incentives to be or appear pivotal.

Furthermore, outcomes are biased in favour of a proprietary technology (for example, Microsoft's) whose single owner has the incentive to market it strategically over 'open' unsponsored alternatives (for example, Linux) – see, for example, Katz and Shapiro (1986b). As discussed above, outcomes are also often biased in favour of networks that are more efficient early on, and are generally biased in favour of established firms on whom expectations focus. The last bias implies entry with proprietary network effects is often nearly impossible (and frequently much too hard from the social viewpoint even *given* incompatibility). And this in turn makes it easier to recoup profits after predatory behaviour that eliminates a rival, and so encourages such predation.

So while incompatibility does not necessarily damage competition, it often does, and firms may therefore also dissipate further resources creating and defending incompatibility.

If firms offer compatible products, then consumers don't need to buy from the same firm to enjoy full network benefits, and (differentiated) products will be better matched with customers. Consumers will be willing to pay more for these benefits, and this may encourage firms to choose compatibility. But compatibility often intensifies competition and nullifies the competitive advantage of a large installed base, whereas proprietary networks tend to make competition all-or-nothing, with the advantage going to large firms, and may completely shut out weaker firms. So large firms and those who are good at steering adopters' expectations may prefer their products to be incompatible with rivals' (see, for example, Katz and Shapiro 1985; Bresnahan 2001), and may be able to use their intellectual property to enforce this.

Competition with incompatible network effects is closely related to other forms of competition when market share is important, especially competition when consumers have switching costs (see, for example, Klemperer 1995; Farrell and Klemperer 2007; and the companion-piece to this article, switching costs), and has similar broader implications (for example, for international trade, see Froot and Klemperer 1989).

Because competition 'for the market' differs greatly from conventional competition 'in the market', and especially because capturing consumers' and complementors' expectations can be so profitable, competition policy needs to be vigilant against predatory or exclusionary tactics by advantaged firms, including deliberately creating incompatibility by misusing intellectual property protection. Thus, for example, the network effect by which more popular operating systems attract more applications software took centre stage in both the US and European Microsoft cases (see, for example, Bresnahan 2001). And because coordination is often important and difficult, institutions such as standards organizations matter, and government procurement policy takes on more significance than usual.

In summary, network effects *can* involve efficient competition for larger units of business – 'competition for the market' – but very often make competition, especially entry, less effective. So I, and others, recommend that public policymakers should have a cautious presumption in favour of compatibility, and should look particularly carefully at markets where incompatibility is strategically chosen rather than inevitable.

Farrell and Klemperer (2007) contains a recent and comprehensive survey of network effects.

See Also

- ▶ Network Goods (Empirical Studies)
- ▶ Switching Costs

The views expressed here are personal and should not be attributed to the UK Competition Commission or to any of its individual Members other than myself. Furthermore, although some observers thought some of the behaviour discussed warranted regulatory investigation, I do not intend to suggest that any of it violates any applicable rules or laws.

Bibliography

Arthur, W.B. 1989. Competing technologies, increasing returns, and lock-in by historical events. *Economic Journal* 99: 116–131.

Arthur, W.B., and A. Rusczcynski. 1992. Dynamic equilibria in markets with a conformity effect. *Archives of Control Sciences* 37: 7–31.

Bresnahan, T. 2001. Network effects in the Microsoft case. Discussion Paper No. 0051, Stanford Institute for Economic Policy Research, Stanford University.

Bulow, J., and P.D. Klemperer. 1999. The generalized war of attrition. *American Economic Review* 89: 175–189.

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

Dybvig, P.H., and C.S. Spatt. 1983. Adoption externalities as public goods. *Journal of Public Economics* 20: 231–247.

Farrell, J., and P.D. Klemperer. 2007. Coordination and lock-in: Competition with switching costs and network effects. In *Handbook of industrial organization*, vol. 3, ed. M. Armstrong and R. Porter. Amsterdam: North-Holland.

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

Farrell, J., and G. Saloner. 1986. Installed base and compatibility: Innovation, product preannouncements, and predation. *American Economic Review* 76: 940–955.

Froot, K.A., and P.D. Klemperer. 1989. Exchange rate pass-through when market share matters. *American Economic Review* 79: 637–654.

Gabel, H.L. 1991. *Competitive strategies for product standards*. New York: McGraw-Hill.

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

Katz, M.L., and C. Shapiro. 1986a. Product compatibility choice in a market with technological progress. *Oxford Economic Papers* 38: 146–165.

Katz, M.L., and C. Shapiro. 1986b. Technology adoption in the presence of network externalities. *Journal of Political Economy* 94: 822–841.

Katz, M.L., and C. Shapiro. 1992. Product introduction with network externalities. *Journal of Industrial Economics* 40: 55–83.

Klemperer, P.D. 1995. Competition when consumers have switching costs. *Review of Economic Studies* 62: 515–539.

Leibenstein, H. 1950. Bandwagon, snob and veblen effects in the theory of consumers' demand. *Quarterly Journal of Economics* 64: 183–207.

Liebowitz, S. 2002. *Re-thinking the network economy: The true forces that drive the digital marketplace*. New York: American Management Association.

Liebowitz, S.J., and S.E. Margolis. 1990. The fable of the keys. *Journal of Law and Economics* 33: 1–25.

Rohlfs, J. 1974. A theory of interdependent demand for a communications service. *Bell Journal of Economics* 5: 16–37.

Rohlfs, J. 2001. *Bandwagon effects in high technology industries*. Cambridge, MA: MIT Press.

Segal, I. 1999. Contracting with externalities. *Quarterly Journal of Economics* 114: 337–388.

Shy, O. 2001. *The economics of network industries*. Cambridge: Cambridge University Press.

Neumann, Franz (1900–1954)

J. Vichniac

In 1942, Franz Neumann, a German legal theorist, completed one of the most influential books written on national socialism. Entitled *Behemoth*, it helped set the agenda for scholarship on this subject in the post-war period. Franz Neumann was born in 1900 in Kattowitz on the Polish-German border into an assimilated Jewish family. He served briefly in the German army in World War I and participated in the soldiers' councils that sprung up at the end of the war. He then went on the study in Breslau, Leipzig, Rostock and finally

Frankfurt, where he completed an undergraduate degree in labour law. During the Weimar period, he lived in Berlin, teaching at the Deutsche Hochschule für Politik and practising law. At the same time, he became involved in the Social Democratic Party, serving as a legal adviser. It was this activity which led to his arrest in April 1933 after the Nazi seizure of power. He was able to escape to London a month later, and under the tutelage of Harold Laski he completed a doctorate in political science at the London School of Economics. He found exile in England uncongenial, however, and in 1936 he emigrated to the United States, where he joined the Institut für Sozialforschung which had moved from Frankfurt to Columbia University. There, in the company of other exiles such as Herbert Marcuse, Max Horkheimer, Erich Fromm, Theodor Adorno, Karl August Wittfogel and others he wrote *Behemoth*. When the United States entered World War II, Neumann along with Barrington Moore, Jr., Herbert Marcuse, Leonard Krieger and Carl Schorske, worked in the Office of Strategic Services. He later went on to work in the State Department until the end of the War. In the late 1940s, he returned to Columbia University where he became a professor in political science, a position that he held until 1954 when he died in a tragic car accident.

In *Behemoth*, Neumann analyses the rise of German national socialism as well as the nature of the Nazi regime in power. His explanation of why Germay was attracted to national socialism hinges on Germany's position in the world economic order in the interwar period. With the onset of the Depression in the 1930s, he argues, German businessmen, in search of markets, became committed to imperialism and foreign conquest. This policy was unacceptable to German Social Democracy and therefore could not be pursed within the confines of the Weimar Republic. Nor could this could be done under a restoration of the monarchy. German business, therefore, supported the Nazi seizure of power, Neumann argues, because totalitarian political power was needed to fortify monopoly capitalism.

Once in power, the Nazi elite consisted of four groups: big industry, the party, the bureaucracy, and the armed forces. It was the first two, big industry and the party, that in large part determined policy. The Nazi state was unlike any other state in history. In it, the traditional distinctions between civil society and the state were dissolved. The rule of law was completely abandoned and the German masses, according to Neumann, were kept under control through a policy of persuasion and terror. Neumann believed that Germany would have to be defeated on the battlefield and monopoly capitalism destroyed before it could become a peaceful nation among others.

Neumann's analysis of Nazism has had a profound influence on a generation of scholars working in the Marxist tradition. As archival material has become available, further work has been done on the actual workings of the Nazi state. His mode of analysis, however, continues to dominate the thinking in this area. But for other scholars Neumann's analysis has been controversial ever since its appearance. It is the economic determinism of his explanation that is at the heart of the problem for many historians. They argue that Neumann ignores the importance of individuals, specifically Hitler, and downplays the importance of ideology in explaining the workings of the Nazi state. This creates particular problems for his treatment of anti-semitism when he argues that the German people were 'the least Anti-Semitic of all' and that anti-Jewish policies were adopted only because they were functionally useful to the Nazi state (1942; 1966, p. 121). The Jews, he wrote, would never be killed because they were useful scapegoats for the regime. This is not the only prediction that turned out to be wrong. He believed that the masses would rise up after the end of the War and that the reconstruction of a democratic Germany could not be built on the foundation of middle class support. Still others have criticized the link he made between big business and the national socialism, arguing that the business community was not instrumental in bringing Hitler to power. Yet, despite the problems with this analysis, Neumann never attempted to revise *Behemoth* during the remaining years of his life.

His work at Columbia, in the early 1950s, however, showed a shift in emphasis in his concerns.

Neumann wrote a series of essays grouped under the title *The Democratic and Authoritarian State* which were published and edited by Herbert Marcuse after his death. In these essays, he was concerned with analysing the conflict between political power and political liberty. They were part of a larger project, a comprehensive study of dictatorships that he was unable to complete.

Selected Works

1928. Gesellschaftliche und staatliche Verwaltung der monopolistischen Unternehmungen. *Die Arbeit* 7: 393–406.

1930. Die soziale Bedeutung der Grundrechte der Weimarer Verfassung. *Die Arbeit* 9: 569–582.

1942. *Behemoth: The structure and practice of national socialism*. New York: Oxford University Press. Ist Harper paperback ed. New York: Harper & Row, 1966.

1953. The social sciences. In R. Crawford et al. *The cultural migration: The European scholar in America*, 4–26. Philadelphia: University of Philadelphia Press.

1957. *The democratic and authoritariam state*. London: Free Press.

Bibliography

Hugnes, Stuart H. 1968. Franz Neumann between marxism and liberal democracy. *Perspectives in American History* 2: 446–462.

Jay, M. 1973. *The dialectical imagination: A history of the Frankfurt School and the Institute of Social Research*. Boston: Little, Brown.

Neuroeconomics

John Dickhaut and Aldo Rustichini

Abstract

Neuroeconomics aims at improving the science of major economic phenomena such as the formation of prices and the design and performance of institutions. A revised model of choice is expected, based on the behaviour of the neuronal structures of the brain. Researchers are tackling issues such as determining how fundamental constructs like probabilities and payoffs are reflected in neuronal activity; disentangling the processing of inputs to choice from the act of choice; isolating learning, impulsive and analytic components of neuronal behaviour; and distinguishing how context affects the processing of the brain and subsequent levels of trust and cooperation in exchange.

Keywords

Allais paradox; Choice; Ellsberg paradox; Experimental economics; Learning; Mixed strategy equilibrium; Neuroeconomics; Preference reversals; Prisoner's Dilemma; Probability; Regret; Reputation; Trust; Ultimatum game

JEL Classifications

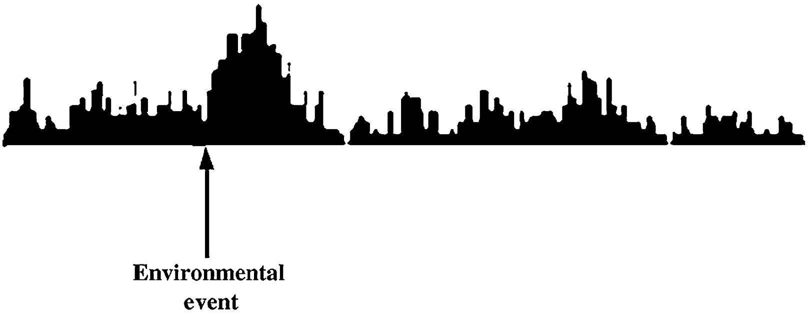
C9

The fundamental unit of activity of the brain is the neuron. It ingests nutrients, receives chemical signals from other neurons, and fires (produces electro-chemical action potentials), which results in sending chemical signals (that is, neurotransmitters) to other neurons. Human brains are estimated to have as many as 100 billion neurons. A first task of neuroeconomics is to accumulate information about the behaviour of collections of neurons and how they interact to produce economic choices.

Research Methods

Research methods employed include single neuron recordings of non-human primates, often macaque monkeys, brain scans (such as functional magnetic resonance imaging, fMRI) of humans and comparative studies of lesioned and normal patients.

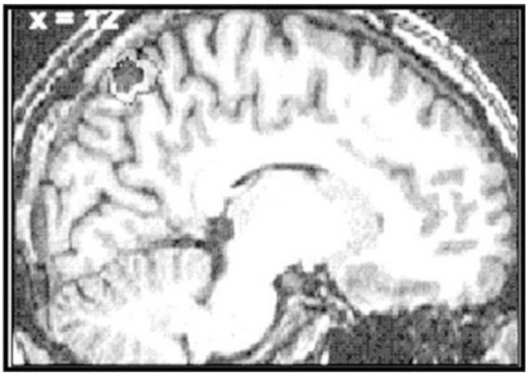
Neuroeconomics, Fig. 1



Single Cell Recording

Only in rare instances is it possible to target specific neurons of living human beings (for example, when someone is having open brain surgery). Because many brain structures of non-humans correspond to human brain structures, it is possible to use results from non-human studies to postulate neuronal structures that function in human brains making economic choices. The method for making observations of a neuron's behaviour using monkeys is single cell recording. In this approach specific groups of neurons are targeted. Electrodes are implanted in individual neurons in the group. When a neuron fires, an electrical impulse is sent to a recording device.

Figure 1 shows a typical result for a specific neuron in a targeted group of neurons. The distance along the horizontal axis represents the number of seconds into the experimental trial. In this picture an experimental event such as the receipt of reward occurred roughly one fifth of the way through the experimental trial. The vertical axis represents the sum of activations for this neuron at each particular time over a set of experimental trials; here there is much activation immediately after the experimental event when looking across trials.



Neuroeconomics, Fig. 2 *Source:* Dehaene et al. (2003)

Imaging

In studying the human brain researchers employ scanning, for example, functional magnetic resonance imaging (fMRI). fMRI surrounds the economic agent with a strong magnetic field. When specific neurons are engaged in a task, capillaries near those neurons carry more oxygenated blood than capillaries surrounding neurons not engaged in the task. fMRI assesses where such oxygenated blood is. These assessments can be represented in an image indicating areas of the brain that activate differentially. A typical scan produces an image like that in Fig. 2. The image shows the implicit activation in the superior parietal lobe (upper-left darkened spot of image) when a subject performs certain numerical operations. The whitened area surrounding the darkened spot suggests the increasing activation around the location.

An fMRI captures brain activity at a much coarser level than single unit recording; it cannot isolate some brain structures in humans to the same degree as single unit recording can isolate neuronal activation in monkeys. fMRI allows investigators time resolution in milliseconds.

A related type of scanning is positron emission tomography (PET). In PET studies subjects are injected with radioactive isotopes. Activated neurons in the brain recruit more blood than other neurons and thus brain areas with more positron emissions indicate where more blood is flowing. These areas are then highlighted to produce an image similar to that in Fig. 2.

Using Lesioned and Normal Subjects

Another type of study involves using lesioned (subjects with damaged brain areas) and normal subjects. When normal subjects perform differently on tasks from lesioned subjects, it is evidence consistent with the hypothesis that the area in question is responsible for the differential performance.

Skin Conductance

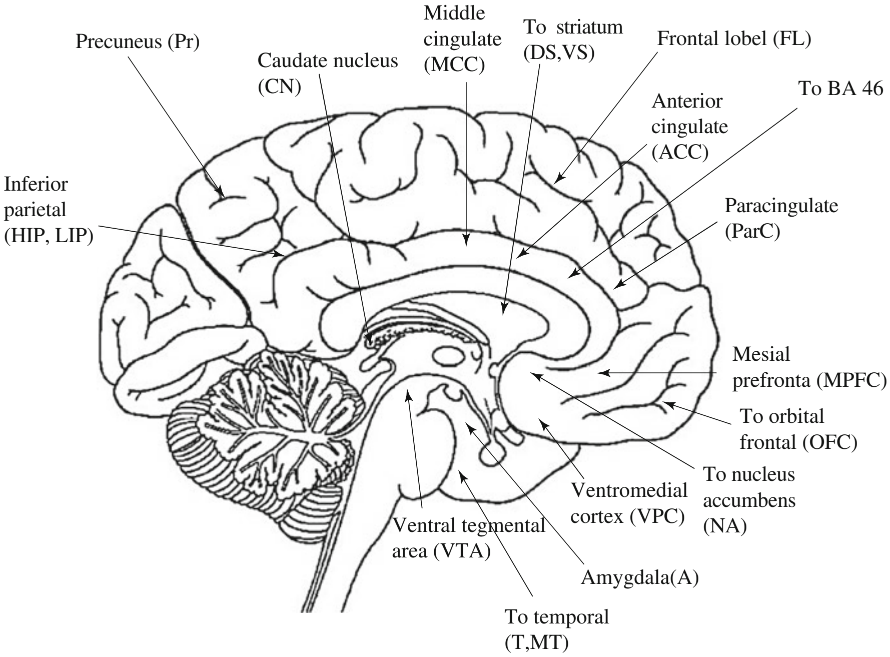
Skin conductance (SCR) measures the ability of skin to conduct electricity (conductance increases with sweat secretion). Generally measures such as SCR and heart rate (HR) have been used to proxy behaviour in the emotional part of the brain. Brain structures associated with emotion send signals to both the heart and the sweat glands.

Figure 3 is intended to assist the reader in identifying brain areas mentioned in the discussion. The image depicts a cross-section (a sagittal view) of the brain taken at the midline of the brain. Approximate locations of brain structures are provided. Where the word 'To' appears in the figure it means the brain part is behind the cross-section at that location.

A critical question about such research concerns what we have learned so far about the economic behaviours of humans (in relation to monkeys) using these methods. The remainder of this article suggests several answers to this question.

Results Related to Games Against Nature

1. The monkey brain has mechanisms that are sensitive to environmental differences in probabilities (relative frequencies) and payoffs. Typically, neuroeconomists with neuroscience backgrounds use a reinforcement perspective. For example, no representation of a probabilistic process is made. Rather, a subject learns probabilities through repeated exposure to outcome feedback. One important set of findings using this paradigm reveals a collection of neurons responsible for detecting differences in economic information in the environment.



Neuroeconomics, Fig. 3

Tremblay and Schultz (1999) used single cell recording to demonstrate that a region of the brain, the ventromedial prefrontal cortex (VPC), has some very specialized neurons. These VPC neurons are differentially activated for different reward expectations in macaque monkeys. The experimenters established that monkeys reveal a preference for different food and liquid items. For example, they were able to establish that a raisin was stochastically preferred to a piece of apple, which was stochastically preferred to cereal. Then they established that, when the raisin and pieces of apple were alternated as rewards, the VPC neurons activated more for raisins than for apples; on the other hand when apples and cereal were the rewards, the same neurons activated more for the pieces of apple.

Fiorillo et al. (2003) showed monkeys were differentially sensitive to differences in probabilities of stimuli. The researchers employed five different visual cues, each of which yielded a reward with different probabilities, 0, .25, .5, .75 and 1.00. Neurons in the ventral tegmental area (VTA) showed higher activation immediately after cues the more likely the cue was to yield a reward. At the actual time of reward the same neurons activated more the less likely it was that the reward would follow.

2. *The findings regarding how monkeys come to know probabilities and payoffs have implications for how humans come to know probabilities and payoffs.* Brain areas such as the VTA are so small that it is not easy to detect them in humans using fMRI. Knutson et al. (2003) exploited neuroanatomy to show that VTA neurons send neural information to the nucleus accumbens (NA) and mesial prefrontal cortex (MPFC) The results from using fMRI indicate that the NA is sensitive to differential gains and that the MPFC encodes differences in probabilities. Thus, Knutson et al., without directly assessing the behaviour of human VTA neurons, were able to look downstream to infer an informational role for these neurons.
3. *Researchers have begun to incorporate results in experiments with feedback into a testable dynamic theory of choice.* The diagnostic role of VTA neurons in relating expectation to outcome serves as a basis for a particular dynamic model of choice, the actor–critic model (Schultz et al. 1997). In the model the critic assesses the difference between expectation and outcome, the difference forms the basis for evaluating the stimuli in the experiment and for revising the probability for the next choice. Berns et al. (2001) showed that parts of this model are appropriate to human behaviour when they looked specifically at how predictable sequences of squirts of water and juice activate brains of human subjects as compared with unpredictable ones. Areas more activated for unpredicted areas than predicted areas included the NA and the orbital frontal cortex (OFC), clusters of neurons also downstream from the VTA. O'Doherty et al. (2004) pinpointed differential activation associated with the actor, dorsal striatum (DS), and the critic, ventral striatum (VS).
4. *Emotions can play a beneficial role in choice.* Bechara and Damasio (2005) invented the Iowa gambling task (IGT) to assess the role of emotions in choice. In earlier studies, emotions had been shown to be associated with activation in the OFC and the amygdala (A). In the IGT subjects sampled 100 times from four decks of cards and subjects received the reward that showed up on the face of the card drawn. Two of the decks were bad decks, resulting in occasional high losses as well as a low long-run payoff. Two were good decks, which produced moderate gains and an occasional moderate loss, but yielded long-run gains. To show that emotions aided choice, Bechara and Damasio report using three sets of subjects – subjects with damage to the VPC area of the brain, subjects with damage to the A, and normal subjects. None of the subjects knew the composition of the decks, but as they performed the task they received feedback; hence the potential for learning the composition of the decks. Neuronal firing was implicitly detected using skin conductance and heart rate (SCR, HR).

Normal and VPC damaged subjects showed SCR and HR increases when the card was observed, but A damaged subjects showed no response. Furthermore, while learning the task normal subjects developed ‘anticipatory’ SCRs, that is, their SCR measurement increased as their hand neared the choice of a bad deck even though supplemental evidence showed no awareness of the bad deck. This anticipatory response was not detected in either of the groups with brain damage. The final piece of evidence corroborating that emotions play a positive role in choice is that subjects with brain damage made poorer choices in the task.

5. *A decision itself consists of more than just a choice. There is a neuronal modification of sensory inputs, a choice, and various neuronal communications to muscular structures that reveal the choice.* Shadlen and Newsome (2001) used a task in which a monkey sees moving dots presented on a screen. A portion of the dots had direction determined randomly and a portion had a fixed direction right or left. The monkey’s choice involved making an eye movement, a saccade, to the right or left signifying the net direction of movements in the dots. The monkey was rewarded if correct.

   Suppose there is a small net movement of dots to the right. When the monkey first sees the dots, they are registered on the retinas of the monkey’s eyes. These signals are transferred through the optic chasm back to the occipital lobe and then to secondary areas of the visual cortex (MT). This processing takes place encoding and partially preserving various aspects of the stimuli, including colour, size and background, but most importantly the direction of movement of the dots on the screen. MT enervates (sends signals to) the lateral interior parietal cortex (LIP); however the LIP does not just preserve the signals in MT but summarizes the net activation between groups of neurons in the MT, in particular the difference in activation in neurons representing movement of dots from right to left. The LIP then sends signals which direct the muscle movements of the eye.

   Such a structure seems somewhat removed from probability and value as they might be expected to be seen in economic choice. Platt and Glimcher (1999) provided the work that helps make the linkage clear. Using single unit recording, they placed electrodes in the LIP. A monkey indicated choices by making eye movements to the left or right. When appropriate a movement to the left yielded a juice squirt of .01 ml while a movement to the right yielded .03 ml. The monkey was signalled the appropriate direction of eye movement by different coloured fixation points in the middle of the monkey’s computer screen. The fixation point signalled left and right with .5 probability. This set-up allowed the investigators to compute the expected payoff at different levels of information (before and after showing the fixation point) to the monkeys. Results revealed a collection of neurons in the LIP that responded monotonically to increases in expected payoff. Thus, in a task with computable expected payoffs, the LIP registers how differences in expected payoff enter into the decision process.

6. *The implicit processes of traditional choice theory tend to be evoked when subjects deal with numerical representations of outcomes.* In results 1–5a reinforcement paradigm is involved, and many findings are the result of repeated trials with subjects bringing no knowledge of the stimuli to the task. For example, in Fiorillo et al. (2003), monkeys experienced one signal at a time and seconds later learned whether a reward occurred. On the other hand economic theory often assumes there can be a structured and often numerical representation of the choice problem. In experiments conducted by economists, physical objects such as dice, urns filled with different-coloured marbles, and wheels of fortune with different-coloured segments have been used to convey probabilistic information. At times subjects have been simply told numbers that represent probabilities that the experimenters would like them to believe were the true probabilities.

   Furthermore, because decision theory describes the relationship between choices that are available to the decision maker given no changes in subject's endowments, studies done by experimental economists have often provided no feedback after every choice; but rather, a randomly selected choice is played only after a set of choices have been made. In this sense experimental economics has traditionally been concerned with choice, while experiments conducted by neuroscientists are often concerned with learning. Such traditional types of experimental economics studies have unearthed a large number of regularities including the Allais and Ellsberg paradoxes and preference reversals, and in some studies expected utility is supported.

Dickhaut et al. (2003) had subjects make binary choices between gambles. For example, the subject could choose between a certainty gamble and a risky gamble (or two risky gambles). Probabilities were represented to subjects as the number of balls of particular colours that could be drawn from an urn, and after a set of choices was made one or more of the subject's designated choices was played. In the study the balls were drawn from a real urn. The study showed that context plays a role in how the brain functions during choice. For risky gambles comparison brain areas such as the frontal lobe (FL) and parietal (P) are relatively more activated than the OFC and nearby areas. Thus, context alters how parts of the brain come into play in choice and simultaneously how analytical functions of the brain are recruited.

Employing this paradigm, Rustichini et al. (2005) added ambiguous and partially ambiguous gambles. They uncovered key aspects of the choice process that are involved when subjects work with explicit probabilistic representations and payoffs. Subjects behaved as if they were employing cut-offs to distinguish between numerical magnitudes; it was also shown that the closer the gamble evaluated was to the cut-off the more difficult the judgment (that is, the longer was the reaction time). Areas of major activation found by Rustichini et al. included P, precuneus (Pr) and Brodman area 6. Rustichini et al. raised the possibility that such cut-off rules operate as approximate calculations like those found by Dehaine et al. when subjects compare numbers to a criterion. In monkeys Dehaine et al. isolated the horizontal inferior parietal (HIP) area as an area capable of making relational comparisons.

Within the classical paradigm Hsu et al. (2005) studied ways in which the brain processed information differently under ambiguity and risk. Using three different approaches to approximating ambiguous and risky tasks, they identified the A and OFC as areas in which ambiguity and risk are differentially processed. In supplemental materials the authors reported inferior parietal activation, which is consistent with giving the subjects both verbal and numerical representations of the choice.

Leland and Grafman (2005) also studied the traditional type of economic tasks. Their study was constructed along the lines of the Bechara and Damasio (2005) studies since they used normal subjects and subjects with brain damage to the VPC. There was no difference between the performance of these groups on these traditional types of tasks, which is consistent with the proposition that people recruit areas other than orbital frontal cortex in performing these tasks.

Another study that examined economic behaviour in a more traditional choice context is McClure et al. (2004), who studied whether agents have a propensity to discount hyperbolically. They found that the evaluation of immediate payoffs produced relatively more VPC activation, but for all decisions (those involving immediate and non-immediate payoffs) a broader set of areas including the Pr and the P areas was activated.

Camille et al. (2004) examined the degree to which normal and subjects with VPC lesions incorporate regret into their choices. In this study regret is the maximum difference in payoffs that exists between two choices. Camille et al. reported that normal subjects are much more likely to incorporate regret into their choices. The authors found that normal and lesioned subjects both incorporated expected value into their choices. In this study subjects saw gambles represented explicitly in terms of payoffs and probabilities. Feedback was provided after every choice. The results of this study and Hsu et al.'s results imply that some of the more analytic processes implied by Dickhaut et al. and Rustichini et al. can be at work in these studies, but that there is emerging a potentially delicate interplay between the reward areas and the analytical areas of the brain.

Results Related to Game Theory

7. *Monkeys' neuronal activity encodes mixed strategies.* Dorris and Glimcher (2004) extended the examination of the behaviour of monkeys to consider how a monkey plays against different strategies of the computer in a game with a mixed strategy equilibrium. Results reveal that monkeys are capable of adjusting their mixed strategies approximately optimally to the mixed strategies played by the computer. Dorris and Glimcher examined the behaviour of LIP neurons and found that they reflected the mixed strategy of the monkeys.
8. *Games with other agents are consistent with a theory of mind.* In typical game theory experiments it is customary to attempt to give players a complete description of the game, from which strategic behaviour ensues. Then it is assumed that individual players generate beliefs contingent on their beliefs about others' strategies. Given this perspective of how choice proceeds, technically it becomes useful to have an experimental design that attempts to ensure that every player has the chance to fully anticipate the other players' actions prior to any moves made by any of the players. Often this common knowledge approach is approximated by representation of a game matrix in a simultaneous-play game or a game tree in a sequential game.

   Neuroscientists have isolated the paracingulate cortex (ParC) as a location associated with the ability to understand another person's deception. Utilizing this perspective, McCabe et al. (2001) investigated whether this area was implicated in cooperative games such as the trust game. They uncovered increased ParC activity when subjects knew they were playing against a person as opposed to a computer, and also found increased ParC activity for cooperative as opposed to non-cooperative players. Sanfey et al. (2003) further examined the McCabe results by employing the ultimatum game and Prisoner's Dilemma games. They preprogrammed a set of outcomes for the subjects to play against. The experimenters attempted to lead subjects to believe they were playing against computers for one set of outcomes and against real people for the other set. These differences in procedure yielded some differences in the brain areas activated. McCabe et al. (2001) found P activation that is not reported by Sanfey et al. However, Sanfey et al. found temporal (T), FL and Pr activation in addition to ParC activation.
9. *Economic reputation building is identifiable at a neuronal level.* King-Casas et al. (2005) used fMRI to scan pairs of subjects in a trust game repeated ten periods. The researchers were able to show that activations in the middle cingulate cortex (MCC) of a sender (when an amount is invested) were coterminous with activations of the anterior cingulate cortex (ACC) of the receiver in the game when the receiver saw the money sent. The receiver's intent to reciprocate was reflected in activation of the receiver's caudate nucleus (CN). Initially this activation lagged the receipt of the investment by approximately eight seconds, but with repeated play the activation precedes receiver's knowledge of the investment by approximately eight seconds. In this way the authors implicitly measured the way economic reputation is built by the sender in the receiver's brain in the trust game.
10. *The brain has mechanisms that reveal individuals enjoy punishing norm violators.* De Quervain et al. (2003) examined the neuronal basis of costly punishment. They allowed the sender to penalize the receiver when the receiver did not reciprocate, but at a cost to the sender. They found evidence consistent

with the assumption that the sender was comparing the costs of punishment with a derived benefit (satisfaction) from punishing. The locus of the derived benefit from punishing was reflected in behaviour of the CN and the VPC, the area in which the authors argued the evaluations take place.

Conclusion

Neuroeconomics has moved the economics from the discussion of useful fictions regarding choice to the direct examination of the structures in the human brain that are making the choices. Evidence to date suggests that the underpinnings of modern-day *homo economicus* are reflected in brain structures that exist in both monkeys and humans and in both Robinson Crusoe and multi-agent settings, and findings are emerging on which a more informed model of choice and exchange can be formulated using brain function as the underpinning.

See Also

- ▶ Altruism in Experiments
- ▶ Behavioural Game Theory
- ▶ Evolutionary Economics
- ▶ Experimental Methods in Economics
- ▶ Market Institutions
- ▶ Trust in Experiments
- ▶ Uncertainty

Acknowledgment This material is based upon work supported by the National Science Foundation under Grant No. 0452477.

Bibliography

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

Bechara, A., and A.R. Damasio. 2005. The somatic marker hypothesis: A neural theory of economic decisions. *Games and Economic Behavior* 52: 336–372.

Berns, G.S., S.M. McClure, G. Pagnoni, and P.R. Montague. 2001. Predictability modulates human brain response to reward. *Journal of Neuroscience* 21: 2793–2798.

Camille, N., G. Coricelli, J. Sallet, P. Pradat, J.R. Duhamel, and A. Sugrue. 2004. The involvement of the orbitofrontal cortex in the experience of regret. *Science* 304: 1167–1170.

De Quervain, J., U. Fischbacher, V. Treyer, M. Schellhammer, U. Schnyder, A. Buck, and E. Fehr. 2003. The neural basis of altruistic punishment. *Science* 425: 785–791.

Dehaene, S., M. Piazza, P. Pinel, and L. Cohen. 2003. Three parietal circuits for number processing. *Cognitive Neuropsychology* 20: 487–506.

Dickhaut, J., K. McCabe, J. Nagode, A. Rustichini, and J. Pardo. 2003. The impact of the certainty context on the process of choice. *Proceedings of the National Academy of Science* 6: 3536–3541.

Dorris, M.C., and P.W. Glimcher. 2004. Activity in posterior parietal cortex is correlated with relative subjective desirability of action. *Neuron* 44: 365–378.

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

Fiorillo, C.D., P. Tobler, and W. Schultz. 2003. Discrete coding of reward probability and uncertainty by dopamine neurons. *Science* 299: 1898–1902.

Hsu, M., M. Bhat, R. Adolphs, D. Tranel, and C.F. Camerer. 2005. Neural systems responding to degrees of uncertainty in human decision-making. *Science* 310: 1680–1683.

King-Casas, B., D. Tomlin, C. Anen, C.F. Camerer, S.R. Quartz, and P.R. Montague. 2005. Getting to know you: Reputation and trust in a two-person economic exchange. *Science* 308: 78–83.

Knutson, B., G.W. Fong, S.M. Bennett, C.M. Adams, and D. Hommer. 2003. A region of mesial prefrontal cortex tracks monetarily rewarding outcomes: Characterization with rapid event-related fMRI. *NeuroImage* 18: 263–272.

Leland, J.W., and J. Grafman. 2005. Experimental tests of the somatic marker hypothesis. *Games and Economic Behavior* 52: 386–409.

McCabe, K., D. Houser, L. Ryan, V. Smith, and T. Trouard. 2001. A functional imaging study of cooperation in two person reciprocal exchange. *Proceedings of the National Academy of Sciences of the United States of America* 98: 11832–11835.

McClure, S.M., D.I. Laibson, G. Lowenstein, and J. Cohen. 2004. Separate neural systems value immediate and delayed monetary rewards. *Science* 306: 503–507.

O' Doherty, J., P. Dayan, J. Schultz, R. Deichmann, K. Friston, and R. Dolan. 2004. Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* 304: 452–454.

Platt, M.L., and P.W. Glimcher. 1999. Neural correlates of decision variables in parietal cortex. *Nature* 400: 233–238.

Rustichini, A., J. Dickhaut, P. Ghirardato, K. Smith, and J.V. Pardo. 2005. A brain imaging study of the choice procedure. *Games and Economic Behavior* 52: 257–282.

Sanfey, A.G., J.K. Rilling, J.A. Aronson, L.E. Nystrom, and J.D. Cohen. 2003. The neural basis of economic decision-making in the ultimatum game. *Science* 300: 1755–1758.

Schultz, W., P. Dayan, and P.R. Montague. 1997. A neural substrate of prediction and reward. *Science* 275: 1593–1597.

Shadlen, M.N., and W.T. Newsome. 2001. Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *Journal of Neurophysiology* 86: 1916–1936.

Tremblay, L., and W. Schultz. 1999. Relative reward preference in primate orbitofrontal cortex. *Nature* 398: 704–708.

Neutral Taxation

Arnold C. Harberger

Abstract

Formally, neutral taxation is taxation falling on something that is in completely inelastic supply, with the tax being so designed as not to affect resource allocation either within or among the affected categories or between them and the other activities not subject to the tax. To minimize deadweight loss, the Ramsey rule says that, the more demand-elastic a good is, the less it should be taxed. But in practice, given ignorance about demand elasticities, uniform low-rate, broad-based taxation reliably reduces deadweight loss and implies neutrality on the part of the state between citizens' preferred actions within the rule of law.

Keywords

Deadweight loss; Efficiency vs. equity; Elasticity; Harberger, A. C.; Land tax; Neutral taxation; Optimal taxation; Ramsey rule taxation; Uniform taxation; Value-added tax

JEL Classifications

H2

One can detect in the literature of economics two important lines of thinking on the subject of neutral taxation. One emphasizes economic efficiency (i.e. the elimination of deadweight loss) as the objective in terms of which the neutrality of taxation is defined. The other emphasizes the generality of a tax as itself imparting the quality of neutrality. Two examples, each with a long history in economic thinking, illustrate the main lines of the distinction.

On the one hand we have the taxation of land rents or land values. It builds on the notion (not precisely true in fact) that each piece or plot of land is totally fixed in supply, with the consequence that any tax levied upon it will ultimately be paid out of its pure economic rent.

On the other hand we have the relatively modern idea of a general tax on value added, the tax being applied at a uniform rate on all activities in the economy. Here there is no thought that the underlying resources are fixed in each activity; quite to the contrary, mobility among the various taxed activities is taken for granted for most of the resources on whose product the tax will fall.

It is easy enough by making artful assumptions to bring these two notions very close together. For example we can *assume* that no manmade improvements to the soil are possible, or alternatively that the tax assessors can always distinguish between 'the intrinsic and immutable qualities of the soil', on which tax is then duly assessed, and the manmade improvements thereon or accretions thereto, on which (under our convenient assumption) no tax is either assessed or paid. Similarly, we can *assume* for the value added tax that there are just three basic resources in the economy – land, labour and capital – and that each of them is fixed in supply. Therefore a uniform tax on the marginal product of any one of them will be neutral, striking the factor equally regardless of the end use to which it is applied, and leaving the factor (because of the assumed zero-elasticity of its supply) no untaxed haven (not even leisure) to which it might choose to escape.

The above assumptions make it easy to define neutral taxation for a Dictionary. (Neutral taxation is taxation falling on something that is in completely inelastic supply, with the tax being so designed as not to affect resource allocation either within or among the affected categories or between them and the other activities not subject to the tax.) But it would probably not add much to the usefulness of the Dictionary.

To be truly useful, I believe, a definition of neutral taxation should be able to throw away such artificial crutches as the two assumptions presented above. It should be able to live in the real world, where we know that the relevant supply elasticities are rarely zero, but where we do not feel at all sure about their magnitudes nor how they vary as between the short, middle and long run. It should be able to cope with reality that, for tax policy at least, the objects of tax do not have an independent essence as commodities; rather, a commodity subject to tax is whatever the tax law (including the regulations and practices followed in enforcing that law) defines it to be. And finally it should come to grips with the serious claims that can be made for considering equality (among the affected activities) in the applicable tax rate to be an attribute whose presence connotes neutrality and whose absence creates a presumption of non-neutrality.

Economics has come the farthest in responding to the first of the desiderata expressed above. Deadweight loss is a concept completely familiar to the discipline, as is the idea of minimizing the deadweight loss of raising a certain amount of tax revenue subject to given constraints. A clear line of thinking runs from Ramsey in the 1920s through Hotelling in the 1930s, Meade in the 1940s, Corlett and Hague and Lipsey and Lancaster in the 1950s, Harberger in the 1960s, to the modern writers on optimal taxation of whom Atkinson, Diamond, Dixit, Mirrlees, and Stiglitz are a representative few. Flowing through this strand of thought are the related ideas (a) that uniform taxation is not always neutral; (b) that the special condition under which uniform taxation of a subset of commodities or activities minimizes the deadweight loss of raising a given amount of revenue from that subset is met when the equilibrium quantity (or activity level) of *each* member of the taxed subset would respond in the same proportion to a (hypothetical) uniform tax on all goods or activities that are *not* in the taxed subset; and (c) that whenever the condition stated in (b) is *not* met then instead of uniform taxation the minimization of deadweight loss requires higher-than-average taxation on goods whose quantities would fall as a result of a (hypothetical) uniform tax on the uncovered group and lower-than-average taxation on those whose equilibrium quantities would rise most sharply.

The analysis underlying the above statements is straightforward, and one can even call economic intuition into play to explain the conclusion. If the tax authorities are denied the possibility of taxing certain goods or activities, then it can to some degree ‘get around’ the ban by putting higher taxes on those items within the taxable subset which are complements of those that cannot be taxed. In a similar vein, since one way of thinking of the resource misallocation that occurs when only a subset of activities is allowed to be taxed is that resources are ‘artificially’ shunted from the taxed to the untaxed subset, it seems quite plausible that the optimal patterning of tax rates within the taxed subset should entail taxing at somewhat lower-than-average rates those particular activities in which a percentage point increment of tax would lead to notably greater-than-average ‘shunting’ of resources to untaxed activities.

The line of reasoning just presented is persuasive – sufficiently so that some economists have been tempted to write off uniformity altogether as a plausible objective of tax policy. There remain many, however, who adhere to uniformity as a goal. Given the ease with which propositions (a) through (c) above can be derived, one should hope that most of those who hold to uniformity base their adherence on considerations extraneous to the derivation, say, of the Ramsey rule and other similar propositions in the literature on optimal taxation. The discussion that follows assumes so.

To build a case for uniformity in taxation in the face of the foregoing logic, one should (appropriately, I think) postulate that one is not dealing with two quite arbitrary categories of goods and/or activities, viz., the taxed subset and the untaxed subset. Instead, one should assume that the taxed subset, rather than being ‘any arbitrary bundle’, is so selected as to contain all the goods and activities that can plausibly and without unusual administrative or regulatory effort be brought into the tax net. One then proceeds to view the problem not as a simple analytical puzzle but as one of guiding or governing the interaction between the society's fiscal authorities and its members.

With this objective in mind, an advocate of uniform taxation might set up a quite different problem from that posed earlier. He might consider the 'disturbance' with which he is dealing to be a consumer changing his mind about how to spend his money or a worker changing his preference about where or for whom to work. A uniform-tax advocate would likely place a considerable value on the authorities' simply not caring about these various changes of mind.

When one solves the Ramsey problem one takes as given the tastes and preferences of economic agents and maximizes government revenue for a given aggregate level of the agents' welfare. Under the differentiated set of tax rates that emerges from this exercise, the maximizer is not indifferent to changes in tastes of the agents. The maximizer likes it when agents shift their tastes from low-taxed to high-taxed activities, and is disappointed by shifts in the other direction.

Something of the same thing occurs when uniform taxation is implemented. Here the 'good' event would be a shift in tastes that caused untaxed activities to contract and taxed activities to expand; the 'bad' event would be the opposite. But there would be a wide range of changes of tastes that would be neutral–these would cover shifts among commodities or activities within the sector subject to the uniform tax, and also shifts among activities in the untaxed sector. To the degree that the authorities are successful in extending the tax net over quite a wide range, it may turn out to be true that most changes in tastes simply lead to shifts in the composition of goods within the taxed group. This is the sort of scenario that would best fit the vision of an advocate of broad-based, uniform taxation and at the same time would (at least if changes in tastes within the taxed sector were frequent and important) create problems for proponents of Ramsey rule taxation.

Subtle overtones of a less technical nature also arise when Ramsey-rule taxation is compared to a broad-based, uniform levy. In Ramsey-rule taxation individuals are genuinely presented with incentives to shift their demand from high-taxed to low- taxed products, and workers are likewise motivated to shift their labour efforts from high-taxed to low-taxed activities. Both these incentives are counterproductive from the social point of view. Subtly hidden in the way the problem is framed is the assumption that people's tastes are given. The reality of the world is that tax laws change only rarely; once enacted, they stay in effect for long periods of time, over which economists can be certain that there will be important changes in the parameters of tastes and technology. The goal of having a tax system that is *robust* against these unknown future shifts in demand and supply is not capricious; it deserves to be taken seriously.

In a quite different vein, there arises the question of to what degree we want our choice of tax patterns to depend on parameters like elasticities of supply and demand about which our knowledge is very spotty and imperfect. Proponents of uniform taxation can fairly argue that their choice of such a form does not depend seriously on knowledge about the parameters of demand and supply. Economic theory assures us that the dominant force is substitution (in the sense that a tax on an activity will, other things equal, cause that activity to contract). There is thus a very strong presumption that broadening the coverage and lowering the rate of a uniform tax will reduce the deadweight loss associated with it (for given revenue yield). One can build policy on this basis without having any detailed knowledge of the parameters of supply and demand, without any particular hope of gaining anything more than a very patchy knowledge about them in the future, and indeed *with* an almost absolute assurance that whatever the relevant parameters might be now, they will undergo substantial changes in the future. If one believes that these conditions come close to describing our present and likely future state of knowledge about the relevant parameters, he will likely be predisposed toward uniform as against Ramsey-rule taxation.

The last line of argument favouring uniform taxation has to do with the interplay between equity and efficiency considerations in governing tax policy. The motivations that fall under the umbrella of 'equity' are too numerous and too varied to try to recount here. But nowhere among them can one find that it is fairer to tax more heavily factors of production that cannot flee to other activities or that it is more just to tax heavily those items whose demand happens to be less elastic. To tax salt more heavily than sugar simply and solely because it has a lower elasticity of demand is at least as capricious (from the standpoint of equity) as taxing people differently according to the colour of their eyes.

Ultimately, I believe, the issue of uniform versus Ramsey-rule taxation may turn out to be just one facet of much broader philosophical differences. Consider the philosophy of government that assigns to government the role of creating a framework of laws and regulations within which the private sector then is encouraged to operate freely. Under this philosophy a positive value is placed on the authorities' not caring about what private agents do (so long as they abide by the rules). It is a position desideratum to create a tax system that is robust against changes in tastes and technology.

On the other side of the coin we have a philosophy of social engineering, in which the detailed tastes and technology of the society enter as data into a process by which the policy makers choose parameters such as tax rates and coverages so as to maximize some measure of social net benefit.

Each of these philosophies has had its own long trajectory within the profession of economics. Each has its representatives today. Each will surely be reflected in the literature of future decades. In my opinion, the future debate as to how the concept of neutrality in taxation should be reflected in real-world policy decisions will swirl around the subtle differences between the ways in which holders of these two philosophies view the world, between the roles they envision for government, and between the ways they see the science of economics interacting with government in the formation of policy.

See Also

- ▶ Optimal Taxation
- ▶ Public Finance
- ▶ Ramsey Pricing

Bibliography

Atkinson, A.B.. 1977. Optimal taxation and the direct versus indirect tax controversy. *Canadian Journal of Economics* 10: 590–606.

Atkinson, A.B.., and J.E. Stiglitz. 1980. *Lectures on public economics*. New York/Maidenhead: McGraw-Hill. (esp. Lectures 12–14).

Corlett, W.J., and D.C. Hague. 1953. Complementarity and the excess burden of taxation. *Review of Economic Studies* 21: 21–30.

Diamond, P.A., and J.A. Mirrlees 1971. Optimal taxation and public production. I: Production efficiency; II: Tax rules. *American Economic Review* 61: 8–27, 261–278.

Dixit, A.K. 1970. On the optimum structure of commodity taxes. *American Economic Review* 60: 295–301.

Harberger, A.C. 1964. Taxation, resource allocation and welfare. In *The role of direct and indirect taxes in the Federal Revenue System*, ed. J.F. Due. Princeton: Princeton University Press.

Hotelling, H. 1938. The general welfare in relation to problems of taxation and of railway and utility rates. *Econometrica* 6: 242–269.

Lipsey, R.G., and K. Lancaster. 1956–7. The general theory of second best. *Review of Economic Studies* 24: 11–32.

Meade, J.E. 1955. *Trade and welfare*, *vol. 2: Mathematical supplement.* Oxford: Oxford University Press.

Mirrlees, J.A. 1976. Optimal tax theory: A synthesis. *Journal of Public Economics* 6: 327–358.

Mirrlees, J.A. 1979. The theory of optimal taxation. In *Handbook of mathematical economics*, ed. K.J. Arrow and M.D. Intriligator. Amsterdam: North-Holland.

Neutrality of Money

Don Patinkin

Keywords

Homogeneity postulate; Homogeneity property; Hume, D.; Inflation; Inside and outside money; Intermediate products; IS-LM model; Liquidity preference; Money illusion; Money supply; Neutrality of money; Nominal interest rate; Perfect foresight; Phillips curve; Quantity theory of money; Rational expectations; Real interest rate; Superneutrality; Time preference

JEL Classifications

E4

‘Neutrality of money’ is a shorthand expression for the basic quantity-theory proposition that it is only the level of prices in an economy, and not the level of its real outputs, that is affected by the quantity of money which circulates in it. Thus the notion – though not the term – goes back to early statements of the quantity theory, such as the classic one by David Hume in his 1752 essays ‘Of Money’, ‘Of Interest’ and ‘Of the Balance of Trade’. At that time the notion also served as one of the arguments against the mercantilist doctrine that the wealth of a nation was to be measured by the quantity of gold (which in 18th-century England constituted a – if not the – major form of metallic money: Feaveryear 1963, p. 158) that it possessed. The term itself is much more recent. Though attributed by Hayek (1935, pp. 129–31) to Wicksell, it is actually due to continental economists in the late 1920s and early 1930s to whom Hayek also refers (see 1935, pp. 129–31; see also Patinkin and Steiger 1988).

1. The rigorous demonstration of, the neutrality of money is based on the critical assumption that individuals are free of ‘money illusion’. An individual is said to suffer from such an illusion if he changes his economic behaviour when a currency conversion takes place: when, for example (as in Israel in 1985), a new monetary unit – the ‘new shekel’ – is introduced in circulation and declared to be equivalent to 1,000 old shekels.

It can be shown (Patinkin 1965) that an illusion-free individual in an economy with borrowing who maximizes utility subject to his budget constraint will have demand functions which depend on relative prices, the rate of interest, and the real value of his initial wealth – which consists of physical capital, bond holdings, and money balances. That is, the demand of this representative individual for the jth good, d_j, is described by the function

$$d_j = f_j(p_1/=p,\ldots,p_{n-2}/p,r,K_0+B_0/P+M_0/p)(j=1,\ldots,n-2),$$

where the p_j are the respective money (or absolute) prices of the $n - 2$ goods; p is the average price level as defined by $p = \sum_j w_j p_j$ where the w_j are fixed weights; r is the rate of interest; K_0 is physical capital, B_0 is the initial nominal value of bond holdings (which, for a debtor, is negative), and M_0 is the initial quantity of money. Thus when the new shekel is introduced in circulation, the price of each good in terms of this shekel (and hence the general price level), the terms of indebtedness, and the nominal quantity of initial money holdings are respectively reduced to 1/1,000th of what they were before; hence relative prices and the real value of initial wealth are unaffected; hence so are the amounts demanded of each good.

Mathematically, the foregoing property of the demand functions is described by the statement that these functions are homogeneous of degree zero in the money prices *and* in the initial quantity of financial assets, including money. Accordingly, the absence of money illusion is sometimes referred to as the homogeneity property of the demand functions. (For the necessary and sufficient conditions that must be satisfied by the utility function in order to generate such illusion-free demand functions, see Howitt and Patinkin 1980.) This homogeneity property is to be sharply distinguished from what the earlier literature denoted as the ‘homogeneity postulate’, by which it meant the invariance of demand functions with respect to an equiproportionate change in money prices alone, and which invariance it erroneously regarded as the condition for the absence of money illusion and hence for the neutrality of money (Leontief 1936, p. 192; Modigliani 1944, pp. 214–15): for even in the case of an individual who is neither debtor nor creditor, such a change affects the real value of his initial money balances, hence is not analogous to a change in the monetary unit, and hence – by virtue of the real-balance effect – will generally lead him to change the amounts he demands of the various goods.

For a closed economy, the aggregate value of B_0 is obviously zero, for to each creditor there corresponds a debtor. For simplicity, we can also consider the amount of physical capital, K_0, to remain constant. Disregarding distribution effects, the demand functions of the economy as a whole for the $n - 2$ goods can then be represented by

$$D_j = F_j(p_1/p,\ldots,p_{n-2}/p, r, M_0/p)(j = 1,\ .\ .\ .\ , n - 2)$$

and the corresponding supply functions by

$$S_j = G_j(p_1/p,\ldots,p_{n-2}/p, r).$$

The general-equilibrium system of the economy is then

$$\left|\begin{array}{l} F_1(p_1/p,\ldots,p_{n-2}/p,r,M_0/p) = G_1(p_1/p,\ldots,p_{n-2}/p,r). \\ \qquad\cdot\quad\cdot\quad\cdot \\ \qquad\cdot\quad\cdot\quad\cdot \\ \qquad\cdot\quad\cdot\quad\cdot \\ F_{n-2}(p_1/p,\ldots,p_{n-2}/p,r,M_0/p) = G_{n-2}(p_1/p,\ldots,p_{n-2}/p,r) \\ F_{n-1}(p_1/p,\ldots,p_{n-2}/p,r,M_0/p) = 0 \\ F_n(p_1/p,\ldots,p_{n-2}/p,r,M_0/p) = M_0/p. \end{array}\right|$$

The $(n - 1)$st equation is for real bond holdings, whose aggregate net value is (as already noted) zero; and the nth equation is for real money balances. Assume that this system has a unique equilibrium solution with money prices $p_1^0,\ldots,p_{n-2}^0, p^0$ and the rate of interest r^0, and that the economy is initially at this position. Let the quantity of money now be changed to kM_0, where k is some positive constant. From the preceding system of equations we can immediately see that (on the further assumption that the system is stable) the economy will reach a new equilibrium position with money prices $kp_1^0,\ldots,kp_{n-2}^0, kp^0$ and an unchanged rate of interest r^0. (Clearly, this conclusion would continue to hold if the supply functions $G_j(\,)$ were also dependent on M_0/p.) Thus the increased quantity of money does not affect any of the real variables of the system, namely, relative prices, the rate of interest, the real value of money balances, and hence the respective outputs of the $n - 2$ goods. In brief, money is neutral: or in the picturesque phrase which Robertson (1922, p. 1) apparently coined, money is a veil. (For empirical studies, see Lucas 1980, and Lothian 1985.)

Furthermore, Archibald and Lipsey (1958) have shown that if the initial equilibrium exists not only with respect to the economy as a whole, but also with respect to each and every individual in it (which, inter alia, means that each individual was initially holding his optimum quantity of money), then this neutrality will obtain in the long run even if one does take account of distribution effects. That is, even if one takes account of differences in tastes, endowments, and hence individual demand functions, an increase in the quantity of money, no matter how distributed among individuals, will in the long run cause an equiproportionate increase in prices and leave the rate of interest invariant. This conclusion in turn follows from the fact that the sequence of short-run equilibria generated by the increase in the quantity of money will in the long run redistribute this quantity in a way that results in an equiproportionate increase in the money holdings of each individual, relative to his holdings in the initial equilibrium position (see also Patinkin 1965, pp. 50–9).

It should also be noted that the preceding analysis has implicitly assumed a unitary elasticity of expectations with respect to future prices, so that neutrality is not disturbed by substitution between present and future commodities.

2. The conclusions of the foregoing analysis are clearly those of long-run comparative-statics analysis. It was this fact that led Keynes – even in his quantity-theory period as represented by his *Tract on Monetary Reform* (1923) – to disparage their policy implications with the famous remark that '*in the long run* we are all dead' (1923, p. 80, italics in original). It should therefore be emphasized that at the same time they demonstrated the long-run neutrality of money, quantity theorists (including Keynes of the *Tract*) also emphasized its non-neutrality in the short run (Patinkin 1972a). Thus Hume emphasized that prices do not immediately rise proportionately to the increased quantity of money and that in the intervening period this stimulates production. In Hume's words:

> it is of no manner of consequence, with regard to the domestic happiness of a state, whether money be in a greater or less quantity. The good policy of the magistrate consists only in keeping it, if possible, still increasing; because, by that means, he keeps alive a spirit of industry in the nation . . . (1752, pp. 39–40)

Hume's emphasis on the irrelevance of the absolute level of the money supply (and hence of money prices) in contrast with the significance of the rate of change of this level was also made by later quantity-theorists. Some of them stressed the stimulating effects of rising prices on 'business confidence' and hence economic activity. A more frequent explanation of the short-run non--neutrality of money was in terms of the shift in the distribution of real income as between creditors and debtors generated by a changing price level. Of particular importance was the danger that a sharply declining price level would increase the number of bankruptcies among debtors, with all its adverse repercussions on the economy. Another source of non-neutrality was the fact that individual prices do not change at the same rate in response to a monetary change. Thus if after a monetary decrease, wage rigidities cause the decline in wages to lag behind that of product prices, the resulting increase in the real wage rate would generate unemployment; conversely, the lag of wages in the case of an inflation would increase profits and hence stimulate production. This consideration led some quantity-theorists to deny even the long-run neutrality of money on the grounds that profit-recipients had a higher tendency to save than wage-earners, so that the shift in income in favour of profits would increase savings, and that these would lead to an increase in the real stock of physical capital in the economy, and hence to a decline in the long-run rate of interest.

For Irving Fisher, the important lag was that of the nominal rate of interest behind the rate of (say) inflation generated by a monetary increase. In particular, because of the lack of perfect foresight on the part of savers (who are the lenders), the nominal rate does not rise sufficiently to offset this inflation; and the resulting decline in the real rate of interest causes entrepreneurs to increase their borrowings, hence investments and economic activity in general. Conversely, when prices decline, corresponding misperceptions cause an increase in the real rate of interest and hence a decline in economic activity. Indeed, Fisher (1913, ch. 4) based his whole theory of the business cycle on this process: the cycle was for him 'the dance of the dollar' (Fisher 1923).

The greatly increased importance of income and capital-gains taxation since Fisher's time is the background of the present-day view – much stressed by Feldstein (1982, and references there cited) – that inflation would have real effects on the economy even if there were perfect foresight, so that the nominal rate fully adjusted itself to the rate of inflation, leaving the real rate of interest unchanged. This is particularly true for the taxation of income from capital, with the simplest example being the increased tax burden on corporations generated by the calculation of depreciation expenses on the basis of historical (as distinct from replacement) costs in an inflationary economy (see also Birati and Cukierman 1979). This is a specific instance of the short-run non-neutrality of money generated by the existence of a tax structure formulated in nominal terms (as is the case with, for example, specific taxes and income-tax brackets) which are generally adjusted to the rate of inflation only after a lag.

Short-run non-neutrality is a basic feature of Keynesian monetary theory and stems from the contention that in a situation of unemployment, prices will not rise proportionately to the increased quantity of money, and that the resulting increase in the real quantity of money will cause a decline in the rate of interest and hence an increase in the volume of investment and the level of national income. The short-run non-neutrality of money is, however, also a basic tenet of today's monetarists, who contend that though the long-run effect of a change in the quantity of money is primarily on prices, its short-run effect is primarily on output. In Friedman's words: 'In the short run, which may be as much as five or ten years, monetary changes affect primarily output. Over decades, on the other hand, the rate of monetary growth affects primarily prices' (Friedman 1970, pp. 23–4).

This non-neutrality has been rationalized by Lucas (1972) in terms of the individual's inability to determine whether a change in the price of a good with which he is particularly concerned (for example labour, in the case of a wage-earner) is a change only in the price of that good (in which case it represents a change in its relative price, which calls for a quantity adjustment) or is part of a general change in prices which does not affect relative prices. In accordance with this approach, and under the assumption that markets always clear, it has also been claimed that only an unanticipated change in the quantity of money will have real effects; for an anticipated one will be expected by the individual to affect all prices proportionately (Lucas 1975; Barro 1976). A far-reaching corollary of this claim is that if, in accordance with the assumption of rational expectations, the public anticipates the actions that government will carry out within the framework of its proclaimed monetary policy, then this policy too will be neutral: that is, the systematic component of monetary policy will not affect any of the real variables of the system (cf. McCallum 1980 and references there cited). Thus under these circumstances even the short-run Phillips curve is – from the viewpoint of systematic monetary policy – vertical.

Empirical support for the claim that only unanticipated monetary changes will have real effects was at first provided by Sargent (1976) and Barro (1978). Contrary conclusions were, however, reached in subsequent empirical studies by Fischer (1980), Boschen and Grossman (1982), Gordon (1982), Mishkin (1982, 1983) and Cecchetti (1986). These differing conclusions stem from different views about the respective ways to estimate (1) that part of a monetary change that is anticipated and/ or (2) the extent of the time lags that must be taken account of in measuring the effects of a monetary change on output. In any event, the weight of opinion today is that both anticipated and unanticipated changes in the money supply have short-term real effects. To the extent that anticipated changes have such effects, this can be interpreted either as reflecting the influence of nominally formulated elements (for example the aforementioned tax structure, or long-term wage contracts – Fischer 1977) in an economy functioning in accordance with the hypothesis of rational expectations cum market-clearing; or, alternatively, it can be interpreted as a refutation of this hypothesis in part or in whole. Thus once again we are confronted with *la condition scientifique* of our discipline: its inability in all too many cases to reach definitive conclusions about theoretical questions on the basis of empirical studies, an inability which increases directly with the political significance of the question at issue.

3. Neoclassical quantity-theorists contended that a shift in the demand curve for money would also have a long-run neutral effect on the economy. Thus consider the Cambridge cash-balance equation, $M = KPY$, where Y is the real volume of expenditures and K is that proportion of his planned money expenditures, PY, which the individual wishes to hold in the form of money. Assume that the economy is in equilibrium with a fixed quantity of money M_0 and price level P_0. Let there now take place a positive shift in the demand for money – that is, an increase in K. Because of the budget constraint, this must be accompanied by a negative shift in the demand for goods. Consequently, the price level P will decline until equilibrium is reestablished with the same nominal quantity of money, M_0, but at a lower price level, $P_1 < P_0$. Thus the automatic functioning of the market will in the long run generate the additional quantity of real balances that individuals wish to hold, without affecting the output of goods.

This neutrality can also be demonstrated in terms of the general-equilibrium system presented above. In particular, if we assume that the increased demand for money is accompanied by a symmetric decrease in the demand for all other goods and for bonds, then a new equilibrium will be established with all money prices reduced in the same proportion, and with an unchanged rate of interest; correspondingly, the respective outputs of goods are also unchanged. In Keynesian monetary theory, however, the increased demand for money is assumed to be solely at the expense of bond holdings: this, after all, is an implication of Keynes's theory of liquidity preference. Such a shift in liquidity preference will accordingly not be neutral in its effects; instead, it will cause an increase in the rate of interest with consequent effects on investment and other real variables of the system (Patinkin 1965, chs VIII:5 and X:4).

In an analogous manner, a change in the proportions between inside and outside money generated by a change in the currency/deposit ratio and/or the bank-reserve/deposit ratio will not be neutral in its effects (Gurley and Shaw 1960, pp. 231–6). It should, however, be emphasized that if the demand and supply functions of the financial sector are also characterized by absence of money illusion, then an increase in outside money will leave these ratios unchanged and hence be neutral (Patinkin 1965, ch. XII: 5–6).

So far, our concern has implicitly been an increase in the quantity of money generated by a one-time government deficit, after which the government returns to a balanced budget. This results in an initial net increase in the total of financial assets in the economy and is thus the real-world analytical counterpart of an increase in the quantity of money generated by the proverbial helicopter dropping down money from the skies. If, however, the monetary increase is generated by an open-market purchase of government bonds (so that initially there is no change in total financial assets), and if there is a real-balance effect in the commodity market, then, as Metzler (1951) showed in a classic article, the equilibrium rate of interest will decline, so that money will not be neutral in its effects. If, however, individuals fully anticipate and discount the future stream of tax payments needed to service the government bonds (in which case these bonds are not part of net wealth), neutrality will obtain in this case too (Patinkin 1965, ch. XII:4).

4. The discussion until this point has dealt almost entirely with the neutrality of a once-and-for-all increase in the quantity of money in a stationary economy. An analogous question arises with reference to the long-run neutrality of a change in the rate of growth of the money supply in a growing economy – in which context the notion is referred to as ‘superneutrality’. Thus consider an economy in steady-state equilibrium whose population is growing at the rate n. Assume that the nominal quantity of money is growing at a faster rate, $\mu = \dot{M}/M$ so that (in order to maintain the constant level of per-capita real money balances that is one of the characteristics of such a steady state) prices rise at the constant rate $\pi = \mu - n.$ Money is said to be superneutral if (say) an increase in the steady-state rate of its expansion, and hence in the corresponding rate of inflation, will not affect any of the steady-state real variables in the system, with the exception of per-capita real-balances: that is, per- capita capital, k; per-capita output, y; and the real rate of interest, r, equal to the marginal productivity of capital. On the other hand, because of the higher costs of holding real balances – in terms of loss of purchasing power, or, alternatively, in terms of the forgone higher nominal rate of interest, i, generated by the increased rate of inflation – the steady-state per capita real value of these balances, m, should generally be expected to decrease.

As already indicated, for Irving Fisher (1907, ch. 5; 1913, pp. 59–60; 1930, pp. 43–4) it was only the absence of perfect foresight which prevented such superneutrality from obtaining: for were such foresight to exist, the nominal rate of interest would simply increase so as to compensate for the inflation and thus leave the real rate of interest (which, under the assumption of continuous compounding, equals $i - \pi$) unchanged. Fisher, however, did not take account of the possible effects of the way the increased amount of money is injected into the economy and/ or the possible effects of the resulting decrease in real balances on other markets. Thus by assuming that the government increases the quantity of money in the economy by distributing it to households and thereby increasing their disposable income, Tobin (1965, 1967) – in a generalization of the Solow (1956) growth model to a money economy – showed that a higher rate of inflation will generally cause individuals to change the composition of their asset portfolios by shifting out of real money balances and into physical capital, thus increasing the steady-state values of k and y – and hence (by the law of diminishing returns) decreasing that of r – so that superneutrality does not obtain.

Tobin’s analysis assumes a constant savings ratio. In a critique of this analysis, Levhari and Patinkin (1968) showed inter alia that if instead this ratio is assumed to depend positively on the respective rates of return on capital and on real money balances – that is, on the real rate of interest and on the rate of deflation – then an increase in the rate of inflation might decrease steady-state savings and hence k, thus causing an increase in the real rate of interest. Similarly, if real money balances were explicitly introduced into the production function, an increase in the rate of inflation might so decrease these balances as to decrease steady-state per-capita output and hence savings sufficiently to offset the positive substitution effect on *k*, thus generating a decrease in the latter.

Patinkin (1972b) analysed superneutrality by means of an IS–LM model generalized to a full employment economy with a real-balance effect in the commodity market (the following largely reproduces the relevant material in this reference). As in Solow (1956), the economy is assumed to have a linearly homogeneous production function, Y = *F*(*K*, *L*), where *Y* is output, *K* capital, and *L* labour, with the labour force assumed to be growing at the exogenous rate *n*. The intensive form of this function is then $y = f(k)$ and its derivative, $f'(k)$ is accordingly the marginal productivity of capital, so that the equilibrium real rate of interest is $r = f'(k)$ Following Mundell (1963, 1965), the crucial assumption of this model is that whereas investment and saving (and hence consumption) decisions depend upon the real rate of interest, $r = i - \pi$, the decision with respect to the amount of real money balances to hold depends on the nominal rate of interest, *i*– for the alternative cost of holding money instead of a bond is precisely this rate. The same is true if we measure this cost in terms of the alternative of holding physical capital: for the total yield on this capital is its marginal product (equal in equilibrium to the real rate of interest) *plus* the capital gain generated by the price change (π): that is, it is $r + \pi = i$. Alternatively, if we measure rates of return in real terms, the rate of return on money balances is $-\pi$ and that on physical capital *r*; hence the alternative cost of holding money is the difference between these two rates, or $r - (-\pi) = i$.

Consider now the commodity market. Let *E* represent the aggregate real demand for consumption and investment commodities combined. For simplicity, assume that this demand is a certain proportion, α, of total real income, *Y*. Assume further that this proportion depends inversely on the real rate of interest and directly on the ratio of real money balances, *M/p,* to physical capital, *K*. The second dependence is a type of real-balance effect, reflecting the assumption that the greater the ratio of real money balances to physical capital in the portfolios of individuals, the more they will tend (for any given level of income) to shift out of money and into commodities. The equilibrium condition in the commodity market is then represented by

$$\alpha\left(i - \pi, (M/p)/K\right) \ . \ Y = Y. \tag{1}$$

By assumption, $\alpha_1(.)$ is negative and $\alpha_2(.)$ positive, where $\alpha_1(\alpha_2)$ is the partial derivative of $\alpha(.)$ with respect to its first (second) argument.

Consider now the money market. Following Tobin (1965, p. 679), assume that the demand in this market depends on the volume of physical capital and the nominal rate of interest. More specifically, assume that the demand for money is a certain proportion, λ of physical capital. Thus the larger *K,* the greater (other things equal) the total portfolio of the individuals, hence the greater the demand for money: this can be designated as the scale or wealth effect of the portfolio. Assume further that the proportion λ depends inversely on the nominal rate of interest. That is, the higher this rate, the smaller the proportion of money relative to physical capital which individuals wish to hold in their portfolios: this can be designated as the composition or substitution effect. The equilibrium condition in the money market is then

$$\lambda(i).K = M/p \tag{2}$$

where by assumption the derivative $\lambda^{'}(.)$ is negative.

Dividing Eqs. (1) and (2) through by *Y* and *K*, respectively – and transforming them into per capita form – we then obtain the equations

$$\alpha(i - \pi, m/k) = 1 \tag{3}$$

$$\lambda(i) = m/k \tag{4}$$

In the steady state,

$$\mu = \pi + n. \tag{5}$$

Since μ and n are both assumed to be exogenously determined, the same can be said for the steady-state value of π. Thus in steady states, Eqs. (3) and (4) can be considered as a system of two equations in the two endogenous variables i and m/k, and in the exogenous variable π. On the assumption of the solubility of these equations, the specific value of k (and hence m) can then be determined by making use of the additional equilibrium condition that the marginal productivity of capital equals the real rate of interest, or,

$$f'(k) = i - \pi. \tag{6}$$

In accordance with the usual assumption of diminishing marginal productivity, we also have

$$f''(k) < 0. \tag{7}$$

The solution of system (3)–(4) can be presented diagrammatically in terms of Fig. 1. The curve *CC* represents the locus of points of equilibrium in the commodity market for a given value of π. Its positive slope reflects the assumption made above about the respective influences of the real rate of interest $(i - \pi)$ and of the real-balance effect (as represented by m/k) on a. Namely, a (say) increase in i increases the real rate of interest and thus tends to decrease a: hence the ratio m/k must increase in order to generate a compensating increase in α and thus restore equilibrium to the commodity market. On the other hand, *LL* – the locus of points of equilibriums in the money market – must be negatively sloped: an increase in the supply of money and hence in m/k must be offset by a corresponding increase in the demand for money, which means that i must decline. The intersection of the two curves at *W* thus determines the steady-state position of the economy.

Assume for simplicity that the given value of π for which *CC* and *LL* are drawn is $\pi = \pi_2 > 0$, corresponding to the rate of monetary expansion μ_2. Assume now that this rate is exogenously increased to $\mu = \mu_{3,}$ so that (by (5)) the steady-state value of π is increased accordingly to $\pi_3 = \mu_3 - n > \pi_2$. From the fact that π does not appear in (4), it is clear that *LL* remains invariant under this change. On the other hand, the curve *CC* must shift upwards in a parallel fashion by the distance $\pi_3 - \pi_2$: for at (say) the point Z' on the curve $C'C'$ so constructed, the money/capital ratio m/k and the real rate of interest $i - n$ are the same as they were at point *Z* on the original curve *CC*; hence Z' too must be a position of equilibrium in the commodity market.

We can therefore conclude from Fig. 1 that the increase in the rate of monetary expansion (and hence rate of inflation) shifts the steady-state position of the economy from *W* to *Y'*. From the construction of $C'C'$ it is also clear that the real rate of interest at *Y'* is $r_3 = i_3 - \pi_3$ which is less than the real rate at *W*, namely, $r_0 = i_0 - \pi_2$. Thus the policy of increasing the rate of inflation decreases the steady-state value of the real rate of interest, and also the money/capital ratio.

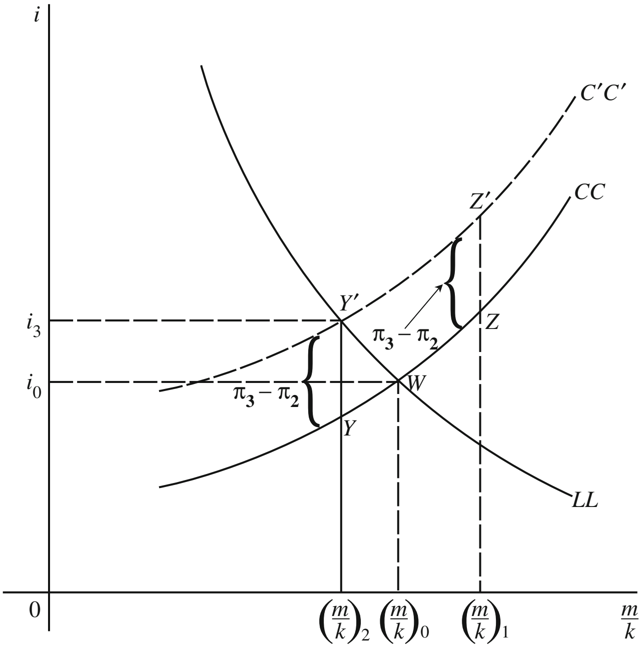
Because of the diminishing marginal productivity of capital, the decline in r implies that k has increased. Thus the fact that m/k has declined does not necessarily imply that m has declined. This indeterminacy reflects the two opposing influences operating on m reflected in Eq. (2), rewritten here in the per capita form as

$$\lambda(i)\ .\ k = m. \tag{8}$$

To use the terminology indicated above, the increased inflation increases the steady- state stock of physical capital, and thus exerts a positive wealth effect on the quantity of real-money balances demanded. At the same time, the increased inflation means that the alternative cost of holding money balances (for a given level of k and hence r) has increased, and this exerts a negative substitution effect on the demand for these balances; that is, individuals will tend to shift out of money and into capital. Thus the final effect on m depends on the relative strength of these two forces. As is, however, generally assumed in economic theory, we shall assume that the substitution effect dominates, so that an increase in π decreases m.

We now note that the only exogenous variable which appears in system (3)–(5) is the rate of change of the money supply, as represented by its steady-state surrogate, $\pi = \mu - n$. In contrast, the absolute quantity of money, *M*, does not appear. It follows that once-and-for-all changes in *M* (after

Neutrality of Money, Fig. 1



which the money supply continues to grow at the same rate) will not affect the steady-state values of *m*, *k*, and *i* as determined by the foregoing system for a given value of π. In brief, system (3)–(5) continues to reflect the neutrality of money. On the other hand, because of the Keynesian-like interdependence between the commodity and money markets, the system is not superneutral.

Note that in the absence of this interdependence, the system would also be superneutral. This would be the case either if the demand for commodities depended only on the real rate of interest, and not on *m*/*k* (that is if there were no real-balance effect); or if the demand for money depended only on *k*, and not on the nominal rate of interest – an unrealistic assumption, particularly in inflationary situations which cause this rate to increase greatly.

The first of these cases is analogous to the dichotimized case of stationary macroeconomic models (cf. Patinkin 1965, pp. 242, 251 (n.19), and 297–8). It would be represented in Fig. 1 by a *CC* curve which was horizontal to the abscissa. Correspondingly, the upward shift generated by the rate of inflation would cause the new *CC* curve to intersect the unchanged *LL* curve at a money rate of interest which was $\pi_3 - \pi_2$ greater than the original one, and hence at a real rate of interest (and hence value of *k*) which was unchanged; the value of m, however, would unequivocally decline. The second of these cases would be represented by a vertical *LL* curve. Hence the upward parallel shift in the *CC* curve generated by inflation would once again shift the intersection point to one which represented an unchanged real rate of interest. In this case (which, as already noted, is an unrealistic one) the value of *m* also remains unchanged.

5. A common characteristic of the foregoing money-and-growth models is that their respective savings functions are postulated and not derived from utility maximization. An analysis which does derive consumption (and hence savings) behaviour from such maximization was presented by Sidrauski (1967) in an influential article. As before, consider an economy growing at the constant rate n with a linearly homogeneous production function having the intensive form $y = f(k)$. Assume now that the representative individual of this economy is infinitely lived with a utility function which depends on consumption and real balances, and that he maximizes the discounted value of this function over

infinite time, using the constant subjective rate of time preference, q. Under these assumptions, Sidrauski shows that money is superneutral.

As Sidrauski is fully aware, this conclusion follows from the form of his production function together with his assumption of a constant rate of time preference; for this fixes the steady-state real rate of interest at $r = q + n = f'(k)$, which determines the steady-state value of k and hence of r. If, however, the production function depends also on real balances – say, $y = g(k, m)$ – then this superneutrality no longer obtains. For the necessary equality between the marginal productivity of capital and $q + n$ in this case is expressed by the equation $g_k(k, \mathrm{m}) = q + n$ (where $g_k(k, m)$ is the partial derivative with respect to k), which no longer fixes the value of k (Levhari and Patinkin 1968, p. 234). In an analogous argument, Brock (1974) showed that if the individual's utility function depends also on leisure, then an increase in the rate of inflation will affect his demand for leisure, which means that it will affect his supply of labour (that is, labour *per capita)*. Hence even though (in accordance with Sidrauski's argument) the increased rate of inflation will not affect the steady-state values of r, k (that is, capital per *labour-input*), and y (that is output per *labour-input*), it will affect the respective amounts of labour and capital *per capita* and hence output *per capita* – so that it will not be superneutral. Needless to say, Sidrauski's results will also not obtain if the rate of time preference is not constant.

6. The conclusion that can be drawn from this discussion is that whereas there is a firm theoretical basis for attributing long-run neutrality to money (but see Gale 1982, pp. 7–58, and Grandmont 1983, pp. 38–45, 91–5), there is no such basis for long-run superneutrality: for changes in the rate of growth of the nominal money supply and hence in the rate of inflation generally cause changes in the long-run equilibrium level of real balances; and if there are enough avenues of substitution between these balances and other real variables in the system (viz., commodities, physical capital, leisure), then the long-run equilibrium levels of these variables will also be affected. An exception to this generalization would obtain if money were to earn a rate of interest which varied one-to-one with the rate of inflation, so that the alternative cost of holding money balances would not be affected by changes in the latter rate; but though it is generally true that interest (though not necessarily at the foregoing rate) will eventually be paid on the inside money (that is bank deposits) of economies characterized by significant long-run inflation, this is not the case for the outside money which is a necessary (though in modern times quantitatively relatively small) component of any monetary system.

The discussion to this point has treated the economy's output as a single homogeneous quantity. A more detailed analysis which considers the sectoral composition of this output yields another manifestation of the absence of superneutrality. In particular, it is a commonplace that the higher the rate of inflation, the higher the so-called 'shoe-leather costs' of running to and from the banks and other financial institutions in order to carry out economic activity with smaller real money balances. In the case of households, the resulting loss of leisure is denoted as the 'welfare costs of inflation' as measured by the loss of consumers' surplus: that is, by the reduction in the triangular area under the demand curve for real money balances (cf. Bailey 1956). In the case of businesses, the costs of inflation take the concrete form of the costs of the additional time and efforts devoted to managing the cash flow. What must now be emphasized is that the obverse side of the additional efforts of both households and businesses is the additional resources that must be diverted to the financial sector of the economy in order to enable it to meet the increased demand for its services. Thus the higher the rate of inflation, the higher (say) the proportion of the labour force of an economy employed in its financial sector as opposed to its 'real' sectors, and hence the smaller its 'real' output. This is a phenomenon that has been observed in economies with two- and especially threedigit inflation (cf. Kleiman 1984 on the Israeli experience). Viewing the phenomenon in this way implicitly assumes that the services of the financial sector are not final products (which are a component of net national product) but 'intermediate products', whose function it is 'to eliminate friction in the productive system' and which accordingly are 'not net contributions to ultimate consumption' (Kuznets 1951, p. 162; see also Kuznets 1941, pp. 34–45).

See Also

- ▶ General Equilibrium
- ▶ Money Illusion
- ▶ Quantity Theory of Money
- ▶ Real Balances

Bibliography

Archibald, G.C., and R.G. Lipsey. 1958. Monetary and value theory: A critique of Lange and Patinkin. *Review of Economic Studies* 28: 50–56.

Bailey, M.J. 1956. The welfare cost of inflationary finance. *Journal of Political Economy* 64: 93–110.

Barro, R.J. 1976. Rational expectations and the role of monetary policy. *Journal of Monetary Economics* 2: 1–32.

Barro, R.J. 1978. Unanticipated money, output, and the price level in the United States. *Journal of Political Economy* 86: 549–580.

Birati, A., and A. Cukierman. 1979. The redistributive effects of inflation and of the introduction of a real tax system in the US bond market. *Journal of Public Economics* 12: 125–139.

Boschen, J.F., and H.I. Grossman. 1982. Tests of equilibrium macroeconomics using contemporaneous monetary data. *Journal of Monetary Economics* 10: 309–333.

Brock, W.A. 1974. Money and growth: The case of long run perfect foresight. *International Economic Review* 15: 750–777.

Cecchetti, S.G. 1986. Testing short-run neutrality. *Journal of Monetary Economics* 17: 409–423.

Feaveryear, A. 1963. *The pound sterling: A history of english money*. 2nd ed., revised by E.V. Morgan. Oxford: Clarendon Press.

Feldstein, M. 1982. Inflation, capital taxation, and monetary policy. In *Inflation: Causes and effects*, ed. R.E. Hall. Chicago: University of Chicago Press.

Fischer, S. 1977. Long-term contracts, rational expectations, and the optimal money supply rule. *Journal of Political Economy* 85: 191–205.

Fischer, S. 1980. On activist monetary policy with rational expectations. In *Rational expectations and economic policy*, ed. S. Fischer. Chicago: University of Chicago Press.

Fisher, I. 1907. *The rate of interest*. New York: Macmillan.

Fisher, I. 1913. *The purchasing power of money: Its determination and relation to credit interest and crises.* Rev. ed. New York: Macmillan. Reprinted, New York: Augustus M. Kelley, 1963.

Fisher, I. 1923. The business cycle largely a 'Dance of the Dollar'. *Journal of the American Statistical Association* 18: 1024–1028.

Fisher, I. 1930. *The theory of interest*. New York: Macmillan. Reprinted, New York: Kelley and Millman, 1954.

Friedman, M. 1970. *The counter-revolution in monetary theory*. London: Institute of Economic Affairs.

Gale, D. 1982. *Money: In equilibrium*. Cambridge: Cambridge University Press.

Gordon, R.J. 1982. Price inertia and policy ineffectiveness in the United States, 1890–1980. *Journal of Political Economy* 90: 1087–1117.

Grandmont, J.-M. 1983. *Money and value: A reconsideration of classical and neoclassical monetary theories*. New York: Cambridge University Press.

Gurley, J.G., and E.S. Shaw. 1960. *Money in a theory of finance*. Washington, DC: Brookings Institution.

Hayek, F.A. 1935. *Prices and production*. 2nd ed. London: Routledge and Kegan Paul.

Howitt, P., and D. Patinkin. 1980. Utility function transformations and money illusion: Comments. *American Economic Review* 70 (819–22): 826–828.

Hume, D. 1752. 'Of money', 'Of interest' and 'Of the balance of trade'. As reprinted in D. Hume, *Writings on economics*, ed. E. Rotwein, Wisconsin: University of Wisconsin Press, 1970.

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

Kleiman, E. 1984. Alut ha-inflatzya [The costs of inflation]. *Rivon Le-kalkalah [Economic Quarterly]* 30: 859–864.

Kuznets, S. 1941. *National income and its composition, 1919–1938*. New York: National Bureau of Economic Research.

Kuznets, S. 1951. National income and industrial structure. *Proceedings of the International Statistical Conferences 1947* 5: 205–239. As reprinted in S. Kuznets, *Economic Change,* London: William Heinemann, 1954.

Leontief, W. 1936. The fundamental assumption of Mr Keynes' monetary theory of unemployment. *Quarterly Journal of Economics* 51: 192–197.

Levhari, D., and D. Patinkin 1968. The role of money in a simple growth model. *American Economic Review* 58: 713–753. As reprinted in Patinkin (1972c), 205–242.

Lothian, J.R. 1985. Equilibrium relationships between money and other economic variables. *American Economic Review* 75: 828–835.

Lucas, R.E., Jr. 1972. Expectations and the neutrality of money. *Journal of Economic Theory* 4: 103–124. As reprinted in Lucas (1981), 66–89.

Lucas, R.E., Jr. 1975. An equilibrium model of the business cycle. *Journal of Political Economy* 83: 1113–1144. As reprinted in Lucas (1981), 179–214.

Lucas, R.E. Jr. 1980. Two illustrations of the quantity theory of money. *American Economic Review* 70: 1005–1014.

Lucas, R.E. Jr. 1981. *Studies in business cycle theory.* Cambridge, MA: MIT Press.

McCallum, B.T. 1980. Rational expectations and macroeconomic stabilization policy: An overview. *Journal of Money, Credit, and Banking* 12: 716–746.

Metzler, L.A. 1951. Wealth, saving and the rate of interest. *Journal of Political Economy* 59: 93–116.

Mishkin, F.S. 1982. Does anticipated monetary policy matter? An econometric investigation. *Journal of Political Economy* 90: 22–51.
Mishkin, F.S. 1983. *A rational expectations approach to macroeconometrics*. Chicago: University of Chicago Press.
Modigliani, F. 1944. Liquidity preference and the theory of interest and money. *Econometrica* 12: 45–88. As reprinted in American Economic Association. *Readings in Monetary Theory*. Philadelphia: Blakiston for the American Economic Association, 1951.
Mundell, R.A. 1963. Inflation and real interest. *Journal of Political Economy* 71: 280–283.
Mundell, R.A. 1965. A fallacy in the interpretation of macroeconomic equlibrium. *Journal of Political Economy* 73: 61–66.
Patinkin, D. 1965. *Money, interest, and prices*. 2nd ed. New York: Harper & Row.
Patinkin, D. 1972a. On the short-run non-neutrality of money in the quantity theory. *Banca Nazionale del Lavoro Quarterly Review* 100: 3–22.
Patinkin, D. 1972b. Money and growth in a Keynesian full-employment model. In Patinkin (1972c).
Patinkin, D. 1972c. *Studies in monetary economics*. New York: Harper & Row.
Patinkin, D., and O. Steiger. 1988. On the terms ‘neutrality of money’ and ‘veil of money’. *Scandinavian Journal of Economics* 90.
Robertson, D.H. 1922. *Money*. Cambridge: Cambridge University Press.
Sargent, T.J. 1976. A classical macroeconometric model for the United States. *Journal of Political Economy* 84: 207–237.
Sidrauski, M. 1967. Rational choice and patterns of growth in a monetary economy. *American Economic Review* 57: 534–544.
Solow, R.M. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70: 65–94.
Tobin, J. 1965. Money and economic growth. *Econometrica* 33: 671–684.
Tobin, J. 1967. The neutrality of money in growth models: A comment. *Economica* 34: 69–72.

New Classical Macroeconomics

Stanley Fischer

JEL Classifications
E1

The new classical macroeconomics (NCM) attempts to build macroeconomics entirely on the foundations of market clearing and optimization by economic agents. It is also known as the rational expectations–equilibrium approach to macroeconomics. The leading figures are Robert Lucas of the University of Chicago and Thomas Sargent of the University of Minnesota, whose 1981 volume contains many of the formative contributions. Lucas (1977) and Sargent (1982) provide nontechnical accounts of the approach. Other leading figures include Edward Prescott and Neil Wallace of the University of Minnesota and Robert Barro of the University of Rochester.

The Monetary Approach: The Lucas Supply Function

The new classical macroeconomics can be dated from work by Robert Lucas in the early 1970s. The article with greatest popular impact is Lucas’s (1973) ‘Some International Evidence on Output–Inflation Tradeoffs’. This is a market-clearing model from which the Phillips curve emerges as a result of imperfect information about the aggregate price level. (Lucas (1972) is a more difficult article that produces a similar result.) The nature of the approach is clarified by outlining the Lucas model and by contrasting it with other models of the Phillips curve.

Markets are physically separated. There are two types of disturbance in the economy, aggregate disturbances that move the aggregate price level and relative disturbances that affect price in each market, but by definition average zero across all markets. Knowledge about past events and the probability distributions of disturbances is complete, but suppliers and demanders within each market observe only the nominal price in that market in the current period in which they have to make their output and purchase decisions.

In a full information set-up, supply and demand in an individual market would depend on relative price. Participants in the market know the price in that market, but cannot calculate relative price without an estimate of the aggregate price level. The optimal estimate of the aggregate price level, conditioned on the observed price in the market, is a weighted average of the expected aggregate price level and the absolute price observed in the market.

Estimated relative price in each market thus increases with the absolute price in that market *relative to the expected aggregate price level.* Aggregating across all markets, aggregate output is an increasing function of the absolute price level relative to the expected price level. This is the famous Lucas supply function

$$Y_t = \alpha(p_t - {}_{t-1}P_t)$$

where Y is aggregate output or its logarithm, P is the logarithm of the aggregate price level, and ${}_{t-1}P_t$ is the expectation of P_t based on information available at the end of period ($t-1$). The model is closed by assuming that aggregate demand is determined by the quantity equation.

The Lucas model contains a Phillips curve in the sense that output and the price level (relative to the expected price level) are positively correlated. If the price level followed a random walk, the standard Phillips curve relationship between output and the inflation rate would be observed in the data.

What NCM Is Not

The Lucas supply function illustrates the difference between NCM and alternative approaches. The original Phillips–Lipsey approach views the Phillips curve as a reflection of *dis*equilibrium in the labour market, with the wage adjusting to the excess demand for labour according to 'the law of supply and demand'. Such an assumption is regarded as unsatisfactory by NCM because the existence of labour market disequilibrium (or disequilibrium anywhere) implies a failure to exploit mutually beneficial trades. NCM would rule out models with that feature – such as Keynesian models with unemployment – unless the failure to trade is explained within the model.

Despite many shared policy positions, the new approach also differs radically from monetarism. While the Lucas supply function is closely related to the Phillips curve model in Friedman's Presidential Address (1968), Friedman assumed that expectations were adaptive and that the monetary authority by accelerating inflation could keep the unemployment rate below the natural rate. The rational expectations assumption distinguishes NCM from monetarism. It is clear from a reading of Friedman and Schwartz (1963) that monetarists are more willing than the NCM to entertain the possibility of disequilibrium and slow adjustment of expectations. Indeed, from the perspective of NCM, monetarism and Keynesianism are of a piece – and equally unsatisfactory – in their willingness to use rules of thumb and crude empirical relationships to model economic behaviour, and in their willingness to proceed on macroeconomic issues in models without firm microfoundations.

Rational expectations is necessary but not sufficient for NCM. Many economists who do not assume that markets clear do assume that expectations are rational.

Policy Ineffectiveness

The Lucas supply function has two important implications that are central to the new classical macroeconomics: the *policy ineffectiveness result,* to be taken up now, and the *econometric policy evaluation critique,* examined later.

The policy ineffectiveness result is that any *anticipated* monetary policy action will not affect output. Rather, such actions are reflected in both the expected and the actual price levels, leading to no effect on output. The result, contained in Lucas (1973) but made most explicit in Sargent and Wallace (1975), is that monetary policy actions affect output only if they are unanticipated –meaning not reflected in pricing decisions. The result has been misinterpreted as applying to all macroeconomic policy, but would not apply to any *real* policy action: for instance an anticipated increase in the investment tax credit would certainly affect investment and typically also aggregate output. The ineffectiveness result relates only to monetary policy, in a model in which money is neutral except for its Phillips curve effects. That is, the Lucas supply curve produces a tradeoff between inflation and unemployment that is not systematically exploitable by policy makers.

The monetary policy ineffectiveness result has been the subject of much controversy. Models in which the monetary policy makers can respond to events after prices have been set leave open the possibility that systematic monetary policy can have real effects. Long-term labour contracts (as in Fischer 1977, or Taylor 1980) may be a source of effective monetary policy. Barro (1977) pointed out that the assumed form of contracts in Fischer was not optimal in that output decisions were left to the firm rather than being set as part of the contract. In practice, output decisions are made by firms; subsequent microeconomic research has shown that asymmetric information may generate that feature of contracts (Hart and Holmstrom 1987) though it remains difficult to account for the failure of contracts to index for nominal disturbances.

Much of the controversy over the effectiveness of monetary policy derives from an implicit view that the aims of the government and the private sector differ. Stabilizing monetary policy may have a useful role to play if contracts cannot fully describe future contingencies, and if there are costs of frequent renegotiation. By creating a stable macro-economic environment, active monetary policy can encourage long-term contracting even when not all states of nature can be described –but it thereby also increases the damage that can be done by inappropriate policy (Fischer 1980).

Early Success

The NCM derived early success from empirical work by Barro (1978) that appeared to support the implication of the Lucas supply function that only unanticipated changes in the money stock had real effects. However, this implication of the NCM approach is shared by sticky wage theories, such as Fischer (1977), and turns out not to distinguish the NCM from other approaches. Further, empirical work by Mishkin (1983) shows that the result that only unanticipated money matters is not robust to lag length.

Within the NCM school, three sets of empirical results led to a loss of confidence in the Lucas supply function approach and the view that monetary shocks affect output. First, Barro (1978) found that although output was closely related to unanticipated changes in the money stock, the aggregate price level was not. This raised doubts about the Lucas supply function, in which prices are the transmission mechanism through which unanticipated money induces suppliers to increase output. Second, Barro and Hercowitz (1980) and Boschen and Grossman (1982) find that currently perceived changes in the money stock, as reflected in preliminary money stock data, do affect output. Since the theory is built on the assumption that money has real effects only because it is not known, this result was a serious blow to the view that the Phillips curve is a result of imperfect information about current nominal variables. Third, Sims (1980) found in a vector autoregressive system including output, money and interest rates that interest rate shocks accounted for a far larger share of variations in output than money shocks.

Econometric Implications

The rational expectations assumption used by NCM has led to the development of major new econometric methods for the treatment of expectations. Much of the econometric development is contained in Lucas and Sargent (1981). One focus has been on methods of testing the typical rational expectations *cross equation constraints.* These are restrictions on relations between parameters in different equations that follow from the assumption that expectations are optimal predictors of variables accounted for elsewhere in the model. A second focus is the econometric policy evaluation critique.

Econometric Policy Evaluation

In deriving the supply function, Lucas shows that the parameter α, the slope of the Phillips curve, is a decreasing function of the variance of the absolute price level. That is because it is a mixture of the structural supply elasticity in an individual market and the signal extraction problem solved by the supplier in deciding how much to respond to any observed nominal price in her market.

The implication is that parameters of macroeconomic models that appear structural, such as α, the slope of the Phillips curve, may not be invariant to changes in policy. In this case a reduction in the variance of the money supply, which is a policy parameter, will make the Phillips curve steeper.

The implication that parameters may not be invariant to changes in policy is the central point of Lucas's influential *econometric policy evaluation critique,* which has had a profound effect on both policy modelling and econometric practice in general (Lucas 1976). On policy modelling, the argument is that existing econometric models, almost all of which are large-scale versions of textbook IS–LM models with an aggregate supply sector appended, cannot be used for analysing changes in policy, since the parameters in those models would likely change as policy changes. Lucas (1976) concedes that existing econometric models, some of which are commercially successful, may do a good job of forecasting. Nor does he argue that econometric models cannot ever be used for policy evaluation, since the true structural parameters (in the Phillips curve example the micro supply elasticity in an individual market) could in principle sometimes be identified. However in practice identification would be almost impossible for many parameters unless there had been frequent changes in policy 'regimes', or policy rules, that would produce variation in parameters such as the variance of the aggregate price level that affect responses to price signals.

The effect of the Lucas critique on econometric practice arises from a pervasive fear that parameters that had previously been thought structural and that were routinely estimated in empirical macroeconomics, such as the propensity to consume out of wealth, or the interest elasticity of money demand, are not invariant to economic policy. Few practising macroeconomists estimate a demand function for money or consumption function without making a pro forma bow in the direction of the Lucas critique – and those who do not are reminded of the protocol by their discussants.

The influence of the Lucas critique is remarkable in that parameter instability induced by policy changes has not been shown to have been empirically important in whatever failures macroeconometric models have suffered. Nonetheless, the critique has led to a new empirical research agenda in macroeconomics.

Deep Structural Parameters

The argument is that the only truly structural parameters in the economy are tastes and technology, utility and production functions. Technology is to be widely interpreted as including the transactions technology and mechanisms for intertemporal trade. Once these primitives are known, it becomes possible to deduce how consumers and producers will respond to policy actions, whose only significance is in how they modify the constraints facing economic agents. Sargent (1982) presents an eloquent account of the research agenda.

The new approach has been to estimate parameters of utility and production functions from first order conditions rather than to attempt to estimate structural relations. In intertemporal optimization first order conditions are Euler equations. For instance in the life cycle consumption model with one consumption good and intertemporally and contemporaneously separable utility function, the discrete time Euler equation is:

$$U'(C_1) = \beta E_t[(1 + r_{t+1})U'(C_{t+1})]$$

where $\beta < 1$ is the discount factor, r is the (perhaps stochastic) rate of return on any asset, and E_t is the expectation conditional on information available in period t.

Aggregate and cross section data can be used to estimate such equations. Hall and Mishkin (1982) on panel data and Hansen and Singleton (1983) are examples. The purpose may be both to estimate utility function parameters and to test restrictions imposed by the underlying model of consumer optimization. Hall and Mishkin for instance conclude that 20 per cent of consumption is accounted for by consumers who are not satisfying the first order condition with equality, and that such consumers may be liquidity constrained. Mankiw et al. (1985) attempt using aggregate time series data to estimate parameters of utility functions defined over consumption and leisure. Examples of estimates of technological relations include Sargent (1978) on the demand for labour and Blanchard (1983) on inventory demand. Garber and King (1983) have severely criticized the Euler equation approach on the grounds that the identification problem has not been faced squarely.

Real Business Cycles

The apparent failure of the Lucas supply function to account for the correlation between inflation and output as a result of imperfect information has led to the alternative real business cycle approach. In this view, business cycles are equilibrium real phenomena, driven largely by productivity shocks. Endogeneity of the money stock accounts for the inflation- or money-output link.

The most fully worked out real business cycle model is that of Kydland and Prescott (1982). There is a representative agent, an infinite horizon intertemporal maximizer. Production inputs are labour, capital and inventories. The economy is hit by imperfectly observed productivity shocks, which are a mixture of permanent and transitory components. Slow acquisition of information about past shocks is one source of lags in the economy; another is lags in the process by which investment turns into capital. Kydland and Prescott can find parameter values, including the variance of the productivity shocks, that enable them to broadly match the stochastic processes that characterize United States business cycles.

The Kydland–Prescott paper has to deal with a basic problem in the NCM approach, that of the cyclical patterns of wages and leisure.

Intertemporal Substitution of Leisure

All theories of the business cycle have to account for relatively large movements in labour input accompanied by only small changes in real wages. If disequilibrium is disallowed, then the problem is to explain labour's willingness to supply, say, five per cent more labour in booms than in slumps for real wages that may be only one per cent higher. The obvious explanation, if the real wage is in fact procyclical, is that labour supply is very responsive to the wage. If this hypothesis explains business cycle correlations, it remains to reconcile short- and long-run labour supply behaviour, for in the long run labour supply curves may be backward bending.

The theoretical explanation comes from the distinction between responses to transitory and permanent increases in the real wage (Lucas 1977). Workers may respond significantly to a transitory increase in the real wage, choosing to work harder now and substitute future for current leisure when the cost of leisure returns to normal. The intertemporal substitution of leisure mechanism plays an extremely significant role in NCM, for at a deeper level it is the rationale for the Lucas supply function.

Direct evidence in support of this hypothesis has been difficult to find (Altonji 1982). Indeed there is some evidence that the real wage follows a random walk, which means that real wage changes are permanent. Unless transitory wage changes are identifiable at a local level, this result rules out the intertemporal substitution of leisure explanation of large movements of labour input over the cycle. Alternative explanations may be available in which the observed wage does not measure the marginal utility of leisure because long-term arrangements between firms produce efficient allocations of resources without using the wage for short-term allocative purposes. Hart and Holmstrom (1987) present several models of contracts in which the wage is not equal to the marginal utility of leisure.

Leisure and Consumption Over the Cycle

It is well known that an intertemporally separable utility function in which both consumption and leisure are normal goods implies that consumption and leisure should be positively correlated unless their relative price (the real wage) changes. In fact, measured consumption and leisure move in opposite directions over the cycle. The correlation cannot be explained in the typical model without significant movements in the real wage, which do not occur. Mankiw et al. (1985) empirical work documents this difficulty.

Kydland and Prescott account for cyclical patterns of leisure and goods consumption by, first, making productivity shocks the driving force in the cycle, and second, by assuming that past levels of leisure affect the current marginal utility of leisure.

Endogenous Money

The real business cycle approach accounts for the Phillips curve by assuming that the money stock accommodates itself to the level of economic activity (King and Plosser 1984). This view derives some support from the fact that the correlation with output is closer for inside than for outside money.

Ironically the real business cycle and early Keynesian views of the unimportance of money are close, despite the dissimilarities of the analytic approaches.

Policy Analysis

The game-theoretic view of the operation of economic policy implicit in the policy ineffectiveness result has become extremely influential in the wake of the important paper on dynamic inconsistency by Kydland and Prescott (1977). Dynamic inconsistency occurs when a future policy decision that forms part of an optimal plan formulated at an initial date is no longer optimal from the viewpoint of a later date, even though no new information has appeared in the meantime.

The problem is likely to arise when expectations of future policy affect current decisions. For instance, to produce low rates of wage change, policymakers would like it believed that future policy will not accommodate wage increases. However, if wage increases occur, policy may well accommodate them rather than cause unemployment.

Kydland and Prescott view dynamic inconsistency as a major argument for the use of policy rules rather than discretion. Dynamic inconsistency will not occur if policy rules are set out and adhered to. Subsequent developments have analysed the tradeoff between the gains from flexibility produced by discretion and the losses due to dynamic inconsistency (e.g. Rogoff 1985). It is also possible that a rational concern for reputation by policy makers will produce consistent behaviour (Barro and Gordon 1983).

The game theory approach implies a stress on the credibility of policy makers, leading for instance to the view that a credible change in monetary policy could lead to a costless disinflation. This view was expressed in the United States before the disinflation of the early Eighties; the subsequent recessionary disinflation helped reduce support for the NCM. Although the game theory approach is not inherently related to NCM, in that expectations of future policy may matter in models without market clearing, it has in practice been pursued largely in an NCM context.

Summary

The promise of the original Lucas NCM model that an imperfect information market clearing approach to macroeconomics could satisfactorily account for most business cycle phenomena including the Phillips curve has not been fulfilled. Beyond its difficulty in accounting for the apparent real effects of monetary policy, the theory is not good at explaining unemployment in a market-clearing context.

The NCM approach builds on the joint assumptions of market-clearing and optimizing behaviour. The market-clearing hypothesis is unlikely to persist as an analytic axiom, unless it is redefined to the point of being meaningless. But the assumption of maximizing behaviour within a specified environment is the microeconomic ideal to which economists aspire. That component of NCM will surely remain as a major impulse in macroeconomics. So too will the rational expectations assumption and the econometrics associated with that approach.

See Also

- ▶ Business Cycle Measurement
- ▶ IS–LM
- ▶ Natural Rate of Unemployment
- ▶ Neoclassical Synthesis
- ▶ Rational Expectations

Bibliography

Altonji, J.G. 1982. The intertemporal substitution model of labour market fluctuations: An empirical analysis. *Review of Economic Studies* 49 (5): 783–824.

Barro, R.J. 1977. Long-term contracting, sticky prices, and monetary policy. *Journal of Monetary Economics* 3 (3): 305–316.

Barro, R.J. 1978. Unanticipated money, output, and the price level in the United States. *Journal of Political Economy* 86 (4): 549–580.

Barro, R.J., and D. Gordon. 1983. Rules, discretion, and reputation. *Journal of Monetary Economics* 12 (1): 101–121.

Barro, R.J., and Z. Hercowitz. 1980. Money stock revisions and unanticipated money growth. *Journal of Monetary Economics* 6 (2): 257–267.

Blanchard, O.J. 1983. The production and inventory behaviour of the American automobile industry. *Journal of Political Economy* 91 (3): 365–400.

Boschen, J.F., and H.I. Grossman. 1982. Tests of equilibrium macroeconomics using contemporaneous monetary data. *Journal of Monetary Economics* 10 (3): 309–333.

Fischer, S. 1977. Long-term contracts, rational expectations, and the optimal money supply rule. *Journal of Political Economy* 85 (1): 191–205.

Fischer, S. 1980. On activist monetary policy with rational expectations. In *Rational expectations and economic policy*, ed. S. Fischer. Chicago: University of Chicago Press.

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

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

Garber, P.M. and King, R.G. 1983. *Deep structural excavation*. University of Rochester, Department of Economics, Working Paper 83–14, September.

Hall, R.E., and F.S. Mishkin. 1982. The sensitivity of consumption to transitory income –estimates from panel data on households. *Econometrica* 50 (2): 461–481.

Hansen, L.P., and K. Singleton. 1983. Stochastic consumption, risk aversion, and the temporal behaviour of asset returns. *Journal of Political Economy* 91 (2): 249–265.

Hart, O., and B. Holmstrom. 1987. The theory of contracts. In *Advances in economic theory, 5th world congress*, ed. T. Bewley. Cambridge: Cambridge University Press.

King, R.G., and C.I. Plosser. 1984. Money, credit, and prices in a real business cycle model. *American Economic Review* 74 (3): 363–380.

Kydland, F.E., and E.C. Prescott. 1977. Rules rather than discretion: The inconsistency of optimal plans. *Journal of Political Economy* 85 (3): 473–493.

Kydland, F.E., and E.C. Prescott. 1982. Time to build and aggregate fluctuations. *Econometrica* 50 (6): 1345–1370.

Lucas, R.E. 1972. Expectations and the neutrality of money. *Journal of Economic Theory* 4 (2): 103–124.

Lucas, R.E. 1973. Some international evidence on output–inflation tradeoffs. *American Economic Review* 63 (3): 326–334.

Lucas, R.E. 1976. Econometric policy evaluation: A critique. In *The phillips curve and labor markets,* ed. K. Brunner and A.H. Meltzer, Carnegie-Rochester conference series on public policy, vol. 1. Amsterdam: North-Holland.

Lucas, R.E. 1977. Understanding business cycles. In *Stabilization of the domestic and international economy,* ed. K. Brunner and A.H. Meltzer, Carnegie-Rochester conference series on public policy, vol. 5. Amsterdam: North-Holland.

Lucas, R.E., and T.J. Sargent. 1981. *Rational expectations and econometric practice*. Minnesota: University of Minnesota Press.

Mankiw, N.G., J.J. Rotemberg, and J.H. Summers. 1985. Intertemporal substitution in macroeconomics. *Quarterly Journal of Economics* 100 (1): 225–251.

Mishkin, F.S. 1983. *A rational expectations approach to macroeconometrics*. Chicago: University of Chicago Press.

Rogoff, K. 1985. The optimal degree of commitment to an intermediate monetary target. *Quarterly Journal of Economics* 100 (4): 1169–1190.

Sargent, T.J. 1978. Estimation of dynamic labour demand schedules under rational expectations. *Journal of Political Economy* 86 (6): 1009–1044.

Sargent, T.J. 1982. Beyond demand and supply curves in macroeconomics. *American Economic Review, Papers and Proceedings* 72 (2): 382–389.

Sargent, T.J., and N. Wallace. 1975. 'Rational' expectations, the optimal monetary instrument, and the optimal money supply rule. *Journal of Political Economy* 83 (2): 241–254.

Sims, C.A. 1980. Macroeconomics and reality. *Econometrica* 48 (1): 1–48.

Taylor, J.B. 1980. Aggregate dynamics and staggered contracts. *Journal of Political Economy* 88 (1): 1–23.

New Deal

Price V. Fishback

Abstract

US President Franklin Roosevelt's New Deal created the most dramatic peacetime expansion of government in American economic history. It established the basic structures for modern federal/state social welfare programmes, farm programmes, labour policies, regulations of many industries, and government insurance of deposits and mortgages. Roosevelt experimented with a cartel-like industrial policy that was declared unconstitutional by the Supreme Court. The emergency public works and relief programmes built a large number of roads, dams, and other public works, and employed millions of labourers. Recent studies suggest that the impact of the New Deal varied greatly by programme.

Keywords

Banking crises; Cartels; Child labour; Civilian Conservation Corps (USA); Farm programmes; Federal Reserve System; Friedman, M.; Gold standard; Great Depression; Hoover, H.; Internal migration; Keynes, J. M.; Minimum wages; Monetary policy; National Labor Relations Boards (USA); National Recovery Administration (USA); New Deal; Pensions; Protection; Public works; Real business cycles; Reconstruction Finance Corporation (USA); Roosevelt, F.D.; Schwartz, A.; Smoot–Hawley Tariff Act of 1930; Social Security in the United States; Trade unions; Unemployment insurance; Works Progress Administration (WPA)

JEL Classifications

N3

Franklin Roosevelt's New Deal created the most dramatic peacetime expansion of government in American economic history.

When Franklin D. Roosevelt became president in March 1933, real output had fallen 30% from its 1929 peak and the unemployment rate exceeded 25%. Within his first hundred days in office Roosevelt and the Democratic Congress established an incredible array of programmes, a virtual 'alphabet soup' of acronyms. More programmes were added under the First New Deal until 1935, when the Supreme Court declared the National Recovery Administration's (NRA) codes of 'fair' competition for industry and the Agricultural Adjustment Administration (AAA) farm programme unconstitutional. A Second New Deal re-established the farm programme in the name of soil conservation, strengthened the role of unions in collective bargaining, and established the basic structure of most of America's current social insurance and public assistance programmes.

After Roosevelt took office, the federal government, often in conjunction with state and local governments, built a huge number of roads, dams, sanitation facilities, schools, public housing projects, and other public works. The federal government expanded regulation of banking, finance, labour, and a host of other markets, insured and refinanced housing loans, and made extensive loans to numerous private and public entities. In the decades following the 1930s, several waves of historians have provided narratives and interpretations of the New Deal and introductions to their work can be found in collections edited by Dubofksy (1992), Braeman et al. (1975), and Hamby (1969). The recent trends in New Deal studies include a series of studies by economists and economic historians (Fishback et al. 2007a; Bordo et al. 1998).

Searching for an overarching theme for the programmes is a daunting task. The doubling of annual federal spending between the Hoover (1929–32) and Roosevelt years tempts many to describe the New Deal as Keynesian expansionary policy. But the Roosevelt administration ran relatively small budget deficits, as federal tax collections also more than doubled. In a brief meeting and a letter to the *New York Times* Keynes had encouraged Roosevelt to follow an expansionary policy, but the levels of government spending and the small budget deficits pale in comparison with the fall in output to be counteracted (Barber 1996; Brown 1956; Peppers 1973; Romer 1992).

One goal appeared to have been to raise prices and wages, as the establishment of the NRA allowed each industry to establish cartel-like codes that stifled price and quality competition, labour policies promoted unionization and high wages, and farm policies offered price guarantees while cutting output. Ultimately, Roosevelt and his advisors were pragmatists faced with terrible economic problems of nearly every kind. They established agencies and programmes meant to try to solve nearly each and every one. At times the programmes operated at cross-purposes. Higher farm and industry prices worsened the plight of the unemployed and other consumers. The pressure to raise wages exacerbated the unemployment problem, and the NRA codes limited output growth. The administration made constant adjustments in policies, creating a climate of uncertainty about the regulatory environment that left businesses wary of making new investments (Higgs 1997).

New Deal Monetary, Banking, and International Policy

Building on the seminal work by Friedman and Schwartz (1963), many economists argue that monetary policy contributed significantly to the harsh decline in the economy between 1929 and 1933. The Federal Reserve took seriously its international responsibilities in maintaining the gold standard and thus failed to respond sufficiently to three major waves of bank failures in a timely fashion. Many states had begun declaring ‘holidays’ that closed state banks to stave off bank runs. Roosevelt took office in the midst of the third wave of failures and declared a Bank Holiday that closed all national banks. Two-thirds of the banks were declared sound and reopened within the week. The troubled banks were reorganized and the Reconstruction Finance Corporation (RFC) subscribed to their new stock issues, reassuring the public about the solvency of the banking system (Smiley 2002; Mason 2001).

In 1933 Roosevelt also announced that the United States was leaving the gold standard, prohibited gold exports, and devalued the dollar to $35 per ounce of gold. In response, the United States received a substantial flow of gold that stimulated the money supply, and economic growth resumed. Japan, Britain, France and several other leading nations experienced similar resumptions of economic growth when they broke free of their ‘golden fetters’ (Eichengreen 1992; Temin 1989; Temin and Wigmore 1990). Gold inflows continued for the rest of the 1930s as Europe moved towards war. By choosing not to offset the gold inflows, Roosevelt and the Federal Reserve allowed the money supply to expand (Romer 1992). The Federal Reserve took a misstep, however, when it used its newly awarded control over reserve requirements to double them in three steps between 1935 and 1937. The goal was to prevent a potentially inflationary rise in lending by soaking up the substantial excess reserves that banks were holding at the time. The banks responded by increasing their reserves and keeping the same cushion because they did not trust the Federal Reserve to provide adequate liquidity if a bank run occurred. The money supply fell and contributed to a sharp rise in unemployment and drop in real GDP in 1937–8 (Friedman and Schwartz 1963; Romer 1992). There is some disagreement about the impact of the monetary policies. Real business cycle economists argue that monetary and investment changes played much smaller roles than productivity shocks and high-wage labour policies in accounting for the fluctuations during the 1930s (Chari et al. 2005).

The decision to leave the gold standard was accompanied by efforts to expand world trade beginning in 1934 with the Reciprocal Trade Agreement Act (RTA). The Smoot–Hawley Tariff Act of 1930 had helped touch off a series of protectionist responses by other countries that had caused total imports for a group of 75 countries to fall to one-third of their 1929 level. The RTA freed the Roosevelt administration to sign a series of tariff reduction agreements with Canada, several South American countries, Britain and key European trading partners. Consequently, American imports rose from a 20-year low in 1932–3 to an all-time high by 1940 (Irwin 1998; Kindleberger 1986).

Meanwhile, the Banking (Glass–Steagall) Act of June 1933 enacted an additional set of banking policies. Despite the checkered history experienced by state deposit insurance programmes (Calomiris and White 2000), the act created the Federal Deposit Insurance Corporation (FDIC) to insure commercial bank deposits of up to $10,000. Insurance for savings and loans followed within the year. The Banking Act also established regulations, eliminated in the late 1970s, that prevented commercial banks from investing more than 10% of their assets in stocks and paying interest on deposits (Regulation Q). To increase the capital available for housing loans, the Home Owners' Loan Corporation (HOLC) provided funds to refinance troubled mortgages between 1933 and 1936, and the Federal Housing Administration (FHA) began offering insurance of mortgages and home improvement loans. Both agencies aided in the spread of the modern long-term, amortized mortgage loan that replaced short-term loans in which repayment of only interest over the course of the loan was followed by a balloon payment of the principal when it fell due.

The Reconstruction Finance Corporation (RFC): New Deal Lender

Established by President Herbert Hoover in 1932, the RFC was an off-budget government corporation that maintained control of the funds repaid on its earlier loans. The RFC offered the Roosevelt administration flexibility because they could start funding programmes without constantly seeking new appropriations from Congress. In consequence, the RFC became the lender during the starting phase of nearly every major New Deal grant and lending programme. In addition, the RFC provided loans to large numbers of financial institutions of all types, railroads, farmers and local governments (Olson 1998). The RFC loans to private business met with mixed success. The liquidity loans to failing banks in 1932 had not prevented many bankruptcies because the RFC loans were given first priority over depositors and other lenders in case of failure; therefore, banks were prevented from selling their most liquid assets to meet depositor demands for cash. The RFC's purchases of preferred stock in banks reorganized after the Bank Holiday of 1933 exposed the RFC funds to more risk but led to more success at preventing failures (Mason 2001). RFC lending to railroads succeeded in preventing several railroad bankruptcies. However, the spared railroads continued to underinvest in maintenance and capital improvements. In contrast, railroads forced into bankruptcy had to make such investments to attract enough capital to reopen for business (Mason and Schiffman 2004).

Emergency Relief and Public Works Programmes

Unprecedented unemployment rates ranging from 10 to 25% through the 1930s were the New Deal's greatest challenge. Prior to the New Deal, aid to the poor and labour policies had been the purview of state and local governments. Claiming unemployment to be a national emergency, Roosevelt and Congress raised the federal share of relief spending as high as 79% while nearly quadrupling relief spending even as unemployment rates fell by the mid-1930s. The Federal Emergency Relief Administration (FERA, 1933–5), the Civil Works Administration (CWA, winter of 1933–4), and the Works Progress Administration (WPA, 1935–42) offered work relief jobs to households whose incomes fell below a target budget for necessities. The Civilian Conservation Corps (CCC) offered conservation jobs in the nation's hinterlands to youths whose earnings were shared with their parents. The FERA also handed out direct relief until 1935, when the responsibility for 'unemployables' was returned to state and local governments, and the federal government began offering matching grants for public assistance for children, the blind, and the elderly.

Harry Hopkins, who headed the FERA, CWA and the WPA, preferred work relief because it 'provided a man with something to do, put money in his pocket, and kept his self-respect' (Adams 1977, p. 53). To give people incentive to leave work relief for private jobs, WPA monthly earnings averaged 40–50% of full-time private earnings, and the WPA assured people that they would be reaccepted should the private job end. Even so, a significant percentage of workers stayed on work relief jobs for periods as long as a year and in some cases several years (Margo 1993).

Roughly one-fourth of New Deal grant spending went to the Public Works Administration (PWA), Public Buildings Administration (PBA), the Public Roads Administration (PRA), and the Tennessee Valley Authority (TVA). The planning stages on these large-scale projects were longer, the wages were higher, and there was more freedom to hire already employed workers. The relief and public works programmes grants were designed to provide employment, build public projects, and stimulate the economy.

At one level the relief and public works programmes were very successful. Millions of Americans obtained work relief jobs to tide them over, and most of the original public works, many renovated since, are still in place today. To understand the true impact of the New Deal, areas with different amounts of spending need to be compared to get a sense of how their economies would have performed without the New Deal. Since the mid-1990s economists have been using the substantial variation in spending across local areas to make such comparisons while working to control for the feedbacks caused by administrators using New Deal programmes to respond to economic problems. At the local level the benefits of the projects were likely to be stronger when the general share of goods produced in the area for local consumption was higher, the projects hired the unemployed without crowding out private or state and local government employment, and expansions did not raise incomes enough to generate federal income tax payments.

Although cross-sectional studies show little effect of relief jobs on private employment, analysis of panel data can control for unmeasured factors using the information across time for a cross section of areas. The panel studies suggest that an additional relief job reduced private employment by up to half a job (Wallis and Benjamin 1981, 1989; Fleck 1999a). A new relief job also raised 'measured' unemployment by one person because many discouraged workers, who had been out of the labour force and thus not counted as unemployed, were defined as re-entering the labour force as unemployed workers when they accepted relief jobs (Darby 1976; Fleck 1999a).

The impact of public works and relief programmes had more clearly beneficial effects on other measures of socio-economic welfare. Cross-sectional studies of US counties suggest that an added dollar of public works and relief spending per person raised per capita income by roughly 85 cents and stimulated in-migration (Fishback et al. 2005, 2006). Panel studies of more than 100 major cities between 1929 and 1940 show that increased relief spending stimulated birth rates, reduced property crime, and reduced infant deaths and deaths from suicide and several diseases. The relief costs per death prevented in today's dollars are within the range of modern market values of life, and the costs are lower than the costs per death prevented of many modern safety programmes (Fishback et al. 2007b; Johnson et al. 2006).

Farm Programmes

To raise the incomes of farmers, who had struggled through over a decade of hard times, the New Deal established the structure of the modern US farm programmes. The Agricultural Adjustment Administration (AAA) paid farmers to take land out of production. In 1935 in *United States v. Butler* the Supreme Court struck down the output processing tax that had originally funded the payments. The AAA payments were quickly reinstituted (minus the processing tax) under the Soil Conservation and Domestic Allotment Act (1935). The Commodity Credit Corporation (CCC) insured that farmers were paid higher prices by making loans that could be repaid with the crop itself if market prices fell below a target price. The Farm Credit Administration (FCA) reorganized and expanded farm lending, ultimately becoming involved in more than half of all farm mortgages and a large share of production loans. Meanwhile, the Rural Electrification Administration (REA) provided subsidized loans to give farmers access to electricity, while the Farm Security Administration (FSA) developed programmes to aid low-income farmers.

Efforts to determine the AAA's impact on limiting farm output have been confounded because a series of major climatic disasters in the 1930s served to cut output anyway. There is evidence that farmers stopped planting their least productive land and raised the inputs used on the remaining land. The AAA clearly aided large farmers but possibly at the expense of farm workers and tenants (Alston and Ferrie 1999; Whatley 1983). Cross-county studies show that increases in AAA payments in counties led to no increases in retail sales, were associated with higher infant mortality in the South, and stimulated net outmigration (Fishback et al. 2001, 2005, 2006; Alston and Ferrie 1999; Whatley 1983). On the positive side, the AAA soil conservation programmes encouraged a move to larger farms and practices that cut soil erosion, so that the Great Plains avoided a recurrence of the Dust Bowl when the same drought and wind conditions arose later (Hansen and Libecap 2004).

The Political Economic Geography of New Deal Spending

New Deal grant spending across states and counties varied enormously, as some western states received several times more per head than some southern states. Roosevelt in a radio 'fireside chat' vowed that the New Deal would promote 'Relief, Recovery, and Reform'. Critics argued that Roosevelt used the monies primarily to aid his re-election efforts. The distribution process for many programmes was opaque, so New Deal scholars have turned to econometric analysis that simultaneously tests the importance of the stated motives and presidential politicking. Politicking was clearly part of the process in the distribution of total funds and at the programme level. Nearly every study finds that more grants went to swing states and areas with higher political turnout, while some find rewards for loyal Democratic areas as well as districts represented by powerful congressmen. The Roosevelt administration was innovative in targeting radio owners in their push to win elections (Wright 1974; Wallis 1998; Fleck 1999b; Stromberg 2004; Couch and Shughart 1998).

Winning elections required more than just manipulation of spending to hit specific political targets. The Roosevelt administration also enhanced its future re-election prospects by following its stated aims. Many studies find evidence that the Roosevelt administration promoted recovery and relief by spending more in areas with higher unemployment and larger declines in income from 1929 to 1933. Few find signs that the total spending was reform-oriented, but specific relief programmes did target areas with long-term poverty. State governments influenced the distribution by the intensity of their lobbying and their spending in matching grant programmes, while the presence of federal land in a state also drew substantial public works grants. Specific programmes typically followed stated goals. There were so many programmes that nearly everybody could find one that benefited them, ranging from relief for the unemployed and poor to loans and AAA grants for large farmers. The HOLC and FHA housing programmes benefited carefully vetted home owners who were perceived as having lower risk of default (Fishback et al. 2003). There were constant charges of corruption, but the WPA actively battled corruption at the state and local levels by establishing an internal investigative agency. When the federal government increased its control of the distribution of funds within states in the switch from the FERA to the WPA, the distribution of funds within states more closely mirrored the relief, recovery and reform goals (Wallis et al. 2006).

Industrial and Labour Policies

To combat 'destructive competition', low prices and low wages, the National Recovery Administration (NRA) was created to allow industries to establish their own codes for minimum prices, quality standards, trade practices, and labour relations (Bellush 1975). The NRA appeared to be sponsoring a series of industry cartels, as large firms tended to dominate the code-writing process in most industries. Wholesale prices jumped 23% in 2 years, although consumer prices were much slower to rise. Simulations of the economy with and without the NRA imply that it served to slow economic recovery (Cole and Ohanian 2004). The internal problems of cartels were also present, as industries with diverse firms had trouble coming to agreement and a number of firms routinely violated the codes (Alexander and Libecap 2000). The

NRA ended in 1935 when the Supreme Court declared it unconstitutional in the Schechter Poultry case, and few mourned its passing.

The National Labor Relations (Wagner) Act of 1935 expanded the right of workers to collective bargaining through their own representatives beyond the protections originally offered in the 1933 act that created the NRA. Employers were required to bargain with unions when a majority of workers voted for union representation, and employer-sponsored unions were banned. The National Labor Relations Board (NLRB) was established to oversee union elections and the collective bargaining process. As a result, unionization expanded rapidly through a mixture of strikes and elections. In the long run the NLRB policies regularized the union recognition and bargaining process, and the incidence of violent strikes has diminished sharply since (Freeman 1998).

The emphasis on raising wages continued when the Fair Labor Standards Act (FSLA) of 1938 set a national minimum wage, overtime requirements, and child labour restrictions. Workers in agriculture or not employed in interstate commerce were exempted. Congressional support for the act was centred in states outside the South with high-wage industries, more unionization, and more advocates for teenage workers. As a result, the first minimum wage was binding only for low-wage industries in the South, where employers in some southern industries responded by reducing employment, and others switched to labour-saving technologies or limited their business to intra-state commerce to avoid federal regulation (Seltzer 1995, 1997; Fleck 2004).

The Social Security Act of 1935

The legislative centerpiece of the Second New Deal was the Social Security Act (SSA) of 1935, which established the modern structure of public assistance and social insurance programmes. The public assistance grants set some federal guidelines and offered matching grants that gave the states latitude in setting benefits. The new Aid to Dependent Children (ADC), Aid to the Blind (AB), and Old-Age Assistance (OAA) programmes replaced similar state programmes in more than half of the states, and provided coverage for the first time in the remaining states.

State unemployment insurance programmes funded by employer contributions with administrative costs paid by the federal government were established as a long-term alternative to providing emergency work relief. The states retained control over benefits offered. Each designed its own experience-rating system that required employers who laid off more workers to pay higher premiums, a feature not commonly found in other countries' unemployment insurance systems. The experience rating helped reduce seasonal unemployment fluctuations (Baicker et al. 1998).

Social security is most associated with the federal old-age retirement system. In the debates over social security, Roosevelt pressed for an actuarially sound system where the individual's retirement benefits were based purely on his and his employer's own contributions. He was not convinced the old-age pensions were necessary and sought to ensure that future generations would not be saddled with the costs. Others pressed for a subsidized system that provided adequate payments to all who contributed. The plan adopted in 1935 was a hybrid, but the inadequacies of the hybrid system had become apparent by 1939, and the current pay-as-you-go structure was created. A worker and his employer pay taxes into an administrative trust fund that pays benefits to current retirees and serves as a commitment by the federal government to collect enough taxes to pay the worker his own social security pension when he reaches retirement age. The initial taxes were1% of wages each for workers and employers, and the initial benefits paid in 1940 were roughly 25% of the average earnings of workers contributing to the system. Average pension payments are now roughly 40% of the contributing workers' average earnings, and the increase in average lifespans has caused rapid increases in the ratio of retirees to workers. In consequence, the tax rates had risen to over 5.3% each for worker and employer by 2000, with expectations that relative benefits will have to be cut or taxes raised in the future to sustain the system (Schieber and Shoven 1999).

Conclusion

The New Deal was a response to the Great Depression, a major peacetime crisis sandwiched between two world wars. All three crises contributed to short-run rapid expansions of the federal government. When each ended, the government's role retracted somewhat but never to the level that would likely have occurred without the crisis (Higgs 1987). In the span of 6 years the Roosevelt administration built an incredible array of public works and established a series of regulations, government insurance, and public assistance programmes that are still in place today. The New Deal arguably did more to expand the role of government in the United States than the more evolutionary changes that have occurred since the end of the Second World War.

See Also

► Great Depression

Bibliography

Adams, H.H. 1977. *Harry Hopkins: A biography*. New York: G.P. Putnam Sons.

Alexander, B., and G. Libecap. 2000. The effect of cost heterogeneity in the success and failure of the New Deal's agricultural and industrial programs. *Explorations in Economic History* 37: 370–400.

Alston, L., and J. Ferrie. 1999. *Southern paternalism and the American welfare state*. New York: Cambridge University Press.

Baicker, K., C. Goldin, and L. Katz. 1998. A distinctive system: Origins and impact of U.S. unemployment compensation. In *The defining moment: The Great Depression and the American economy in the twentieth century*, ed. M. Bordo, C. Goldin, and E.N. White. Chicago: Chicago University Press.

Barber, W.J. 1996. *Designs within disorder: Franklin D. Roosevelt, the economists, and the shaping of American economic policy, 1933–1945*. New York: Cambridge University Press.

Bellush, B. 1975. *The failure of the NRA*. New York: Norton.

Bordo, M., C. Goldin, and E.N. White, eds. 1998. *The defining moment: The Great Depression and the American economy in the twentieth century*. Chicago: University of Chicago Press.

Braeman, J., R.H. Bremner, and D. Brody, eds. 1975. *The New Deal*. Columbus: Ohio State University Press.

Brown, E.C. 1956. Fiscal policy in the 'thirties: A reappraisal. *American Economic Review* 46: 857–879.

Calomiris, C., and E.N. White. 2000. The origins of federal deposit insurance. In *U.S. bank deregulation in historical perspective*, ed. C. Calomiris. New York: Cambridge University Press.

Chari, V.V., P. Kehoe, and E.R. McGrattan. 2005. Spectral methods in business cycle accounting. *Revista de Economia* 12: 5–18.

Cole, H., and L. Ohanian. 2004. New Deal policies and the persistence of the Great Depression: A general equilibrium analysis. *Journal of Political Economy* 112: 779–816.

Couch, J., and W. Shughart III. 1998. *The political economy of the New Deal*. New York: Edward Elgar.

Darby, M.R. 1976. Three and a half million US employees have been mislaid: Or, an explanation of unemployment, 1934–1941. *Journal of Political Economy* 84: 1–16.

Dubofksy, M., ed. 1992. *The New Deal: Conflicting interpretations and shifting perspectives*. New York: Garland Publishers.

Eichengreen, B. 1992. *Golden fetters: The gold standard and the depression, 1919–1939*. New York: Oxford University Press.

Fishback, P., H. Haines, and S. Kantor. 2001. The impact of the New Deal on black and white infant mortality in the South. *Explorations in Economic History* 38: 93–122.

Fishback, P., J.J. Wallis, and S. Kantor. 2003. Can the New Deal's three R's be rehabilitated? A program-by-program, county-by-county analysis. *Explorations in Economic History* 40: 278–307.

Fishback, P., W. Horrace, and S. Kantor. 2005. The impact of New Deal expenditures on local economic activity: An examination of retail sales, 1929–1939. *Journal of Economic History* 65: 36–71.

Fishback, P., W. Horrace, and S. Kantor. 2006. Do federal programs affect internal migration? The impact of New Deal expenditures on mobility during the Great Depression. *Explorations in Economic History* 43: 179–222.

Fishback, P., R. Higgs, G. Libecap, J.W. Wallis, S. Engerman, J. Hummel, S. LaCroix, R. Margo, R. McGuire, R. Sylla, L. Alston, J. Ferrie, M. Guglielmo, E.C. Pasour, R. Rucker, and W. Troesken. 2007a. *Government and the American economy: A new history*. Chicago: University of Chicago Press.

Fishback, P., H. Haines, and S. Kantor. 2007b. Births, deaths, and New Deal relief during the Great Depression. *Review of Economics and Statistics* 89: 1–14.

Fleck, R. 1999a. The marginal effect of New Deal relief work on county-level unemployment statistics. *Journal of Economic History* 59: 659–687.

Fleck, R. 1999b. The value of the vote: A model and test of the effects of turnout on distributive policy. *Economic Inquiry* 37: 609–623.

Fleck, R. 2004. Democratic opposition to the Fair Labor Standards Act of 1938: Reply to Seltzer. *Journal of Economic History* 62: 231–235.

Freeman, R. 1998. Spurts in union growth: Defining moments and social processes. In *The defining moment: The Great Depression and the American economy in the twentieth century*, ed. M. Bordo, C. Goldin, and E.N. White. Chicago: University of Chicago Press.

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

Hamby, A., ed. 1969. *The New Deal: analysis and interpretation*. New York: Weybright and Talley.

Hansen, Z., and G. Libecap. 2004. Small farms, externalities, and the dust bowl of the 1930s. *Journal of Political Economy* 112: 665–694.

Higgs, R. 1987. *Crisis and leviathan: Critical episodes in the growth of American government*. New York: Oxford University Press.

Higgs, R. 1997. Regime uncertainty: Why the Great Depression lasted so long and why prosperity resumed after the war. *Independent Review* 1: 561–590.

Irwin, D. 1998. Changes in U.S. tariffs: The role of import prices and commercial policies. *American Economic Review* 88: 1015–1026.

Johnson, R., S. Kantor, and P. Fishback. 2006. Striking at the roots of crime: The impact of social welfare spending on crime during the Great Depression. Working paper no. 12825. Cambridge, MA: NBER.

Kindleberger, C. 1986. *The world in depression, 1929–1939*, Rev. ed. Berkeley: University of California Press.

Margo, R. 1993. Employment and unemployment in the 1930s. *Journal of Economic Perspectives* 7(2): 41–59.

Mason, J. 2001. Do lenders of last resort policies matter? The effects of the reconstruction finance corporation assistance to banks during the Great Depression. *Journal of Financial Services Research* 20: 77–95.

Mason, J., and D. Schiffman. 2004. Too-big-to-fail, government bailouts, and managerial incentives: The case of Reconstruction Finance Corporation assistance to the railroad industry during the Great Depression. In *Too-big-to fail: Policies and practices in government bailouts*, ed. B.E. Gup. Westport: Greenwood Press.

Olson, J.S. 1998. *Saving capitalism: The Reconstruction Finance Corporation and the New Deal*. Princeton: Princeton University Press.

Peppers, L. 1973. Full employment surplus analysis and structural change: The 1930s. *Explorations in economic history* 10: 197–210.

Romer, C.D. 1992. What ended the Great Depression? *Journal of Economic History* 52: 757–784.

Schieber, S.J., and J.B. Shoven. 1999. *The real deal: The history and future of social security.* New Haven: Yale University Press.

Seltzer, A.J. 1995. The political economy of the Fair Labor Standards Act. *Journal of Political Economy* 103: 1302–1342.

Seltzer, A.J. 1997. The effects of the Fair Labor Standards Act of 1938 on the southern seamless hosiery and lumber industries. *Journal of Economic History* 57: 396–415.

Smiley, G. 2002. *Rethinking the Great Depression: A new view of its causes and consequences*. Chicago: Ivan R. Dee.

Stromberg, D. 2004. Radio's impact on public spending. *Quarterly Journal of Economics* 119: 189–221.

Temin, P. 1989. *Lessons from the Great Depression*. Cambridge, MA: MIT Press.

Temin, P., and B. Wigmore. 1990. The end of one big deflation. *Explorations in Economic History* 27: 483–502.

Wallis, J.J. 1998. The political economy of New Deal spending revisited, again: With and without Nevada. *Explorations in Economic History* 35: 140–170.

Wallis, J.J., and D.K. Benjamin. 1981. Public relief and private employment in the Great Depression. *Journal of Economic History* 41: 97–102.

Wallis, J.J., and D.K. Benjamin. 1989. Private employment and public relief during the Great Depression. Working paper, Department of Economics, University of Maryland.

Wallis, J.J., P. Fishback, and S. Kantor. 2006. Politics, relief, and reform: Roosevelt's efforts to control corruption and manipulation during the New Deal. In *Corruption and reform*, ed. E. Glaeser and C. Goldin. Chicago: University of Chicago Press.

Whatley, W.C. 1983. Labor for the picking: The New Deal in the South. *Journal of Economic History* 43: 905–929.

Wright, G. 1974. The political economy of New Deal spending: An econometric analysis. *Review of Economics and Statistics* 56: 30–38.

New Economic Geography

Anthony J. Venables

Abstract

New economic geography provides an integrated and micro-founded approach to spatial economics. It emphasizes the role of clustering forces in generating an uneven distribution of economic activity and income across space. The approach has been applied to the economics of cities, the emergence of regional disparities, and the origins of international inequalities.

Keywords

Clustering; Comparative advantage; Congestion; Core–periphery mode; Dispersion; Factor price equalization; Foreign direct investment; Gravity modelling; Imperfect markets; Increasing returns to scale; Industrial organization; International portfolio investment; International trade (theory); Knowledge spillovers; Labour mobility; Linkages; Location of economic activity; Market access; Marshall, A.; Monopolistic competition; New economic geography; Productivity; Regional development; Spatial economics; Urban agglomeration; Urban economics

JEL Classifications
R10

Why is economic activity distributed unevenly across space, with centres of concentrated activity surrounded by 'peripheral' regions of lower density? What economic interactions are there between different geographical areas, and how do these shape income levels in the areas? How does the spatial organization of economic activity respond to exogenous shocks, such as technological change or policy measures? The contribution of 'new economic geography' (NEG) is to address these questions in a manner that is based on rigorous microeconomic foundations. It shows how the spatial structure of an economy is determined by the interplay between costs of transactions across space and various types of increasing returns to scale. The questions posed above can be addressed at different spatial levels – international, regional and urban. NEG provides a unified framework for analysis at these different levels.

Clustering Versus Dispersion

The NEG approach has several key analytical ingredients. The first is the recognition that spatial interactions are costly. These costs are shaped by geography and depend on the nature of the interaction. Thus, trade in goods incurs shipping costs and costs of time in transit, depending on distance shipped, on transport infrastructure and on geography. Communications and coordination costs mean that workers may be less effective if they are not in close proximity with co-workers. Factor mobility may be impeded by distance and geography. This approach contrasts with that of international trade theory, in which spatial units are identified solely with countries – jurisdictions rather than geography – and where goods and factors are typically assumed to either be traded freely or to be completely non-tradable. The NEG approach shows how outcomes depend on the extent to which different goods and activities are mobile between locations.

The second key ingredient is the possibility that there are clustering forces, inducing activity to concentrate in space. Clustering arises because of spatially concentrated increasing returns to scale which can derive from a number of different underlying forces. (The classic discussion is Marshall 1890; for a recent survey see Duranton and Puga 2004.) One possibility is that there are public goods, the enjoyment of which depends on geographical access, such as a town centre. Another possibility is that there are positive technological externalities such as knowledge spillovers; firms produce ideas that can be observed and copied by other firms, depending on their proximity. These approaches have been prominent in much of the urban economics literature (for example, Henderson 1988), but writers in the NEG literature have generally sought to derive clustering forces from spatial interactions in imperfect markets rather than to simply assume them through public goods or technological externalities.

One way to derive clustering forces is through thick market effects, particularly in the labour market. Dense labour markets may allow for better matching of the skills of workers and the requirements of firms (Helsley and Strange 1990). Incentives to acquire skills may be greater where workers face more prospective employers (Matouschek and Robert-Nicoud 2005). Another way in which to derive clustering is to use industrial organization models of imperfect competition. The route followed in much of the NEG literature is to suppose that an industry (we will call it 'manufacturing') contains a number of firms, each of which has increasing returns to scale. The presence of internal economies of scale means that firms are faced with a location choice (if they had constant or diminishing returns then, given transport costs and dispersed consumers, they would choose to produce a very small amount in all locations – 'backyard capitalism', Starrett 1978). The questions are, then, where do firms choose to locate, and under what circumstances will they cluster together? The model often used to analyse the choice is the Dixit and Stiglitz (1977) model of monopolistic competition and its international trade extensions (Krugman 1980). In this model each firm has a distinct variety of product which it produces in a single location and exports to other locations, and entry and exit occur until profits are bid down to zero. It turns out that, as firms take location decisions in order to maximize profits, so their location pattern tends to amplify any underlying differences between locations, and from this it is possible to generate an outcome in which clustering occurs.

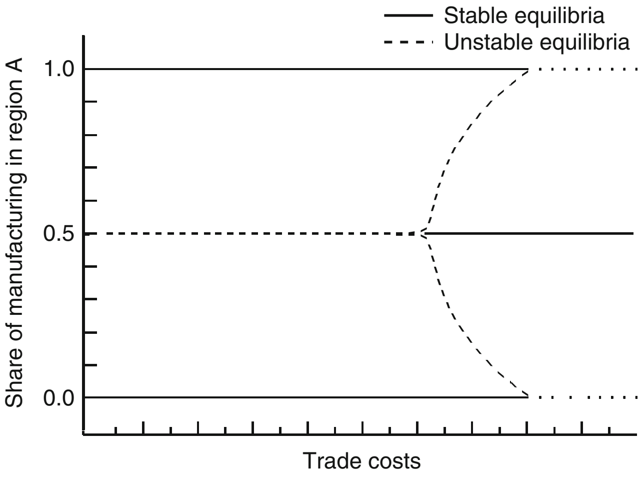
To understand the argument, suppose that there are two regions A and B, and that A has demand $k > 1$ times larger than B (we ignore factor supply considerations for the moment). Could there be an equilibrium in which firms are located in proportion to the size of the regions, so A has k times more manufacturing firms than B? If trade costs are prohibitively high the answer is 'yes'; only local firms supply each market, and the number of firms is proportional to the size of the market. (Notice that this argument uses the Dixit–Stiglitz property that all firms are the same size in equilibrium.) But as trade costs are reduced and firms start to export, two things happen. First, the region B market comes to be supplied by k times as many importing firms as does the country A market, thus reducing the profitability of producers in B. Second, each firm in B will pay transport costs on a large part of their output (sales to the large country A market) while firms in A will pay transport costs only on a smaller fraction of their output (sales to the smaller region B market). Both arguments suggest that firms in A become relatively more profitable, implying that in equilibrium with free entry the number of firms in A must exceed the number in B by a factor greater than k. The large region therefore has a disproportionately large share of manufacturing production, and is a net exporter of manufactures and importer of agriculture. More generally, a region with good 'market access' will attract a high share of firms.

This argument holds only if transport costs lie strictly between zero and a prohibitive level. If transport costs are prohibitive no firms ship any exports; each region is self-sufficient, and the location of industry is in proportion to the size of the regions. Conversely, if transport costs are zero, then the argument collapses, as firms in all regions have equally good access to all markets. The argument shows that it is at intermediate levels of transport costs that market access matters, and manufacturing is pulled disproportionately into the large region.

While this argument creates an incentive for clustering of firms, it is balanced by dispersion forces. These could be due to negative externalities, such as congestion, or arise as a consequence of immobility of some factors of production. Which factors are immobile depend on context, but typically include land (as in the tradition of urban economic modelling) and some or all types of labour. Thus, if labour were immobile, any benefit that firms derived from locating in one region rather than another would create a regional wage differential, until profits (more generally, the return to mobile activities) were equalized across regions.

Labour mobility is central to the Krugman (1991) 'core–periphery' model. This analyses two regions and two sectors, a constant returns to scale agriculture and manufacturing modelled as outlined above. Each sector uses a sector-specific type of labour ('peasants' and manufacturing workers respectively), and the regions' endowments of these factors are, *ex ante*, identical. Crucially, manufacturing workers are mobile between the locations, whereas peasants are immobile. What is the division of manufacturing workers and firms between the two locations? Outcomes, as a function of trade costs, are illustrated on Fig. 1. When trade costs are high manufacturing is equally divided between regions. However, when trade costs are

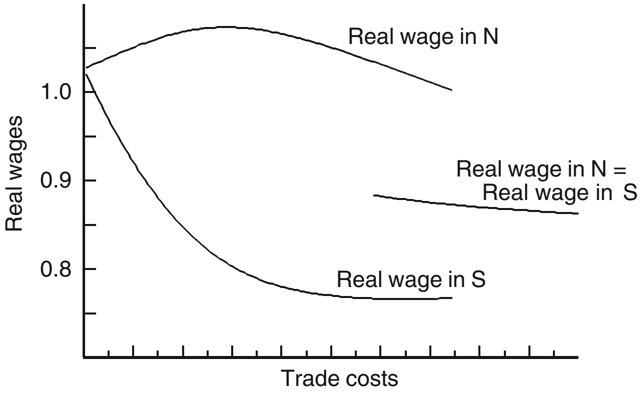
New Economic Geography, Fig. 1 Location of manufacturing in two regions



low enough, manufacturing (and all manufacturing workers) concentrate entirely in one region or the other. There are two mutually reinforcing arguments supporting this clustering. The concentration of manufacturing workers creates a large market, so making the location profitable for firms. And the entry of firms bids up wages, so making the location attractive for workers (this effect reinforced by the fact that workers also benefit from not having to pay trade costs on their consumption of manufactures). It is not profitable for any single firm to leave the cluster, because the benefit of lower wages is outweighed by the loss of market access. As Fig. 1 makes clear, the switch from dispersed manufacturing to agglomeration arises discontinuously. There is a critical value of trade costs, t^*, above which dispersed production is the stable equilibrium, and below which dispersed activity is unstable, while clustering of activity, in either of the regions, is a stable equilibrium.

Krugman's 'core–periphery' model is perhaps the seminal paper, and brings the insight that agglomeration forces can be derived from a standard model of trade and monopolistic competition (see Fujita et al. 1999, for further development these ideas). These micro-foundations mean that outcomes (clustering or dispersion) can be linked to parameters such as trade costs, as in Fig. 1. The model also makes it clear that *ex ante* identical locations can be different *ex post,* and that there are multiple equilibria – we have to look outside the model, or rely on chance, to determine which of the regions has the manufacturing cluster.

The model was constructed with just two locations. How do these insights extend when there are many locations? With many locations the number of equilibria increases dramatically, and there is a danger that little can be said about outcomes. There are several ways through this problem. One is to investigate how the size and number of manufacturing centres on a given geographical space depends on underlying parameters such as trade costs and population levels. The approach of Fujita et al. (1999) is to hypothesize a circular economy (with population on the circumference) and to show that an initial random allocation of manufacturing grows into a determinate number of centres, the size of which is greater (and number of which is smaller) the lower trade costs are. Given some number of centres, reducing trade costs will have no effect until some critical point is reached, at which the economy will reorganize itself to a new economic geography with fewer and larger centres. The approach of Fujita and Mori (1997) is to suppose that initially there is a small populated region. Population growth causes this to expand, at first with the spread of agricultural production into the hinterland. However, these agriculture workers demand manufactures, and this will cause new manufacturing centres to develop. The expanding economy therefore grows its urban structure, and cities will tend to be larger (and further apart) the greater



New Economic Geography, Fig. 2 Real wages in a two-country model

increasing returns to scale are and the lower trade costs are. Both of these approaches work with underlying geographies that are undifferentiated. Adding structure to these underlying geographies simplifies the problem in fairly natural ways. A transport node – such as a port or river crossing – will attract manufacturing, as firms in such a location have better access to a larger number of consumers.

Intermediate Goods and Industrial Clusters

The clustering mechanisms described in the preceding section turn on the mobility of labour. Clustering occurs because, as firms and workers move, so do both supply *and* demand for manufactures. What if labour is immobile? An analogous mechanism can work between firms when we take into account intermediate goods, that is, goods that are both supplied and demanded by the manufacturing sector. This mechanism is similar to the idea of 'linkages' common in the development economics literature of the 1950s and 1960s. This studied the roles of backward linkages (demands from downstream firms to their suppliers) and of forward linkages (supply from intermediate producers to downstream activities) in developing industrial activity. However, as we saw above, rigorous treatment requires that the concepts are placed in an environment with increasing returns to scale, in order to force firms to make a location choice. This can be done in a model isomorphic to that outlined above, but in which firms in the manufacturing sector produce and use intermediate as well as final goods. Clustering can occur as it is profitable for firms producing intermediate and final goods to co-locate. Depending on the strength of linkages within and between industrial sectors, clustering might occur through a wide part of the economy or within narrowly defined sectors.

In this model clustering arises purely from the mobility of firms, even if there is little or no labour mobility. It is applicable to a number of different situations. For example, within a country there might be inelastic supply of land or housing in each city which places a limit on labour mobility. Clustering of particular sectors can nevertheless occur, and might be associated with different levels of employment and different house prices across cities.

The model has also been applied in the international context, with labour immobile across national boundaries. Manufacturing may then concentrate in a single country or group of countries, and this clustering may lead to international wage differences. This idea is developed by Krugman and Venables (1995) in a model with two countries, N and S, assumed to be *ex ante* identical. Firms produce final and intermediate goods, and use labour and intermediates as inputs. Equilibrium outcomes are summarized in Fig. 2, which has trade costs on the horizontal axis and real wages on the vertical axis. At very high trade costs there is no clustering, so the two economies are identical; this is because firms operate in each country to supply local consumers. As trade costs fall (moving left on the figure) so the possibility of supplying consumers through trade rather than local production develops, and clustering forces become relatively more important. Below some level of trade costs, *t**, clustering forces come to dominate, and one of the countries (N) gains most of manufacturing, and consequently has a high real wage. This clustering 'deindustrializes' the other country (S), which experiences a fall in its real wage. For the case illustrated in Fig. 2, there is a range of trade costs in which the world necessarily has a dichotomous structure. Wages are lower in S than in N, but it does not pay any firm to move to S as to do so would be to forgo the clustering benefits of large markets and proximity to suppliers that are found in N. However, as trade costs fall it becomes cheaper to ship intermediate goods, so the location of manufacturing becomes more sensitive to factor price differences. This is the era of globalization, in which manufacturing starts to move to S and the equilibrium wage gap narrows. In this model factor price equalization is attained when trade is perfectly free – the 'death of distance'.

This model offers quite a general theory of location, in which four forces are at work, two of which are dispersion forces, and two favour clustering. The dispersion forces are *factor supply* and *product market competition:* moving a firm from S to N reduces the profitability of firms in N both by bidding up wages and by driving down product prices. Against this there are two agglomeration forces, *demand linkages* and *cost linkages:* moving a firm from S to N raises the profitability of firms in N by increasing the size of the market and by increasing the supply of intermediate goods. The balance between these four forces depends on parameters, including trade costs, giving the outcomes illustrated on Fig. 2. It is worth comparing the four forces present in this model with the conventional model of free international trade, in which factor supply alone determines the location of economic activities.

Extensions of this approach provide a number of further insights concerning international inequalities. It suggests that the world may tend to organize into a rich club of countries and a poor club. Economic development takes the form of countries growing from the poor club to the rich club in sequence rather than in parallel. Parallel growth is unstable because of the tendency of developing manufacturing sectors to cluster in a few countries.

Empirical Findings

The new economic geography literature offers explanations of a number of phenomena that are empirically well documented – even obvious – such as the existence of cities and the presence of regional and international inequalities. Its insights range across different spatial scales, from the urban to the international. Empirical work is correspondingly diverse, and we refer to just four elements of it.

First, there is strong evidence of the importance of geography in shaping economic interactions. Trade costs are high (Anderson and van Wincoop 2004), and 'gravity modelling' points to the fact that bilateral trade flows approximately halve with each doubling of distance between country pairs. Similar results hold for other cross-border interactions such as foreign direct investment flows, telephone calls, and international portfolio investments.

To turn to outcomes, a number of researchers have investigated the extent to which individual sectors are prone to clustering. There is a long business school tradition of work in this area, for example Porter (1990), who studies a number of industrial clusters. Econometric work has established that sectors are more prone to cluster than would be explained by chance or by comparative advantage (Ellison and Glaeser 1997). A further prediction of NEG is that prices of immobile factors will be high in locations with good market access. As we have seen, in the national context this will show up in the price of land and housing and hence nominal wages differences, a prediction confirmed for US counties by Hanson (2005). In the international context this may show up as real wage differences. Gallup and Sachs (1999) find that 70 per cent of cross-country variation in per capita income can be accounted for by just four measures of physical and economic geography (malaria, hydrocarbon endowment, coastal access and transport costs). A structural approach to identifying the importance of market access in explaining cross-country income differentials is adopted by Redding and Venables (2004), who use gravity modelling to calculate measures of market access for each country. With other factors (such as institutional quality) controlled for, these measures of market access are important determinants of international wage gaps.

Finally, there is considerable evidence of the productivity benefits derived from being located in dense centres of economic activity. A recent survey of the literature on cities (Rosenthal and Strange 2004) reports a consensus view that doubling city size is associated with a productivity increase of some three to eight per cent. However, a good deal of uncertainty surrounds the extent to which this is driven by the different clustering mechanisms – knowledge spillovers, thick labour markets, market access benefits, or inter-firm linkages – that we described above. Identifying the importance of each of these underlying mechanisms remains an active area of current research.

See Also

- ▶ City and Economic Development
- ▶ Growth and International Trade
- ▶ International Trade Theory
- ▶ Regional Development, Geography Of
- ▶ Spatial Economics
- ▶ Symmetry Breaking
- ▶ Systems Of Cities
- ▶ Urban Agglomeration
- ▶ Urban Economics
- ▶ Urban Production Externalities

Bibliography

Anderson, J., and E. van Wincoop. 2004. Trade costs. *Journal of Economic Literature* 42: 691–751.

Dixit, A., and J. Stiglitz. 1977. Monopolistic competition and optimum product diversity. *American Economic Review* 67: 297–308.

Duranton, G., and D. Puga. 2004. Micro-foundations of urban agglomeration economies. In *Handbook of urban and regional economics*, vol. 4, ed. J. Henderson and J.-F. Thisse. Amsterdam: North-Holland.

Ellison, G., and E. Glaeser. 1997. Geographic concentration in US manufacturing industries: A dartboard approach. *Journal of Political Economy* 105: 889–927.

Fujita, M., and T. Mori. 1997. Structural stability and the evolution of urban systems. *Regional Science and Urban Economics* 27: 399–442.

Fujita, M., P. Krugman, and A. Venables. 1999. *The spatial economy: Cities, regions and international trade*. Cambridge, MA: MIT Press.

Gallup, J., and J. Sachs. 1999. Geography and economic development. In *Annual world bank conference on development economics 1998*, ed. B. Pleskovic and J. Stiglitz. Washington, DC: World Bank.

Hanson, G. 2005. Market potential, increasing returns and geographic concentration. *Journal of International Economics* 67: 1–24.

Helsley, R., and W. Strange. 1990. Matching and agglomeration economies in a system of cities. *Regional Science and Urban Economics* 20: 189–212.

Henderson, J. 1988. *Urban development: Theory, fact and illusion*. New York: Oxford University Press.

Krugman, P. 1980. Scale economies, product differentiation and the pattern of trade. *American Economic Review* 70: 950–959.

Krugman, P. 1991. Increasing returns and economic geography. *Journal of Political Economy* 49: 137–150.

Krugman, P., and A. Venables. 1995. Globalization and the inequality of nations. *Quarterly Journal of Economics* 110: 857–880.

Marshall, A. 1890. *Principles of economics*. 8th ed, 1920. London: Macmillan.

Matouschek, N., and F. Robert-Nicoud. 2005. The role of human capital investments in the location decisions of firms. *Regional Science and Urban Economics* 35: 570–583.

Porter, M. 1990. *The competitive advantage of nations*. New York: Macmillan.

Redding, S., and A. Venables. 2004. Economic geography and international inequality. *Journal of International Economics* 62: 53–82.

Rosenthal, S. and Strange, W. 2004. Evidence on the nature and sources of agglomeration economies. In *Handbook of urban and regional economics*, vol. 4, ed. J. Henderson and J.-F. Thisse. Amsterdam: North-Holland.

Starrett, D. 1978. Market allocations of location choice in a model with free mobility. *Journal of Economic Theory* 17: 21–37.

New Institutional Economics

L. J. Alston

Abstract

The new institutional economics (NIE) consists of a set of analytical tools or concepts from a variety of disciplines in the social sciences, business and law. The NIE addresses two overarching issues: what are the determinants of institutions – the formal and informal rules shaping social, economic and political behaviour? And what impact do institutions have on economic performance? It is the impact of institutions via property rights and transaction costs that ultimately affect the ability of individuals and societies (at a macro level) to extract the gains from trade which in turn can lead to enhanced economic well-being.

Keywords

Agenda control; Collective action; Contract enforcement; Corruption; Creative destruction; Credible commitment; Firm, theory of the; Free-rider problem; Institutional path dependence; Monitoring costs; New Deal; New institutional economics; Political entrepreneurship; Principal and agent; Prisoner's Dilemma; Property rights; Rational ignorance; Rule of law; Side payments; Social capital; Social norms; Technology; Transaction costs of exchange; Transaction costs of production; Transformation costs

JEL Classifications

E12

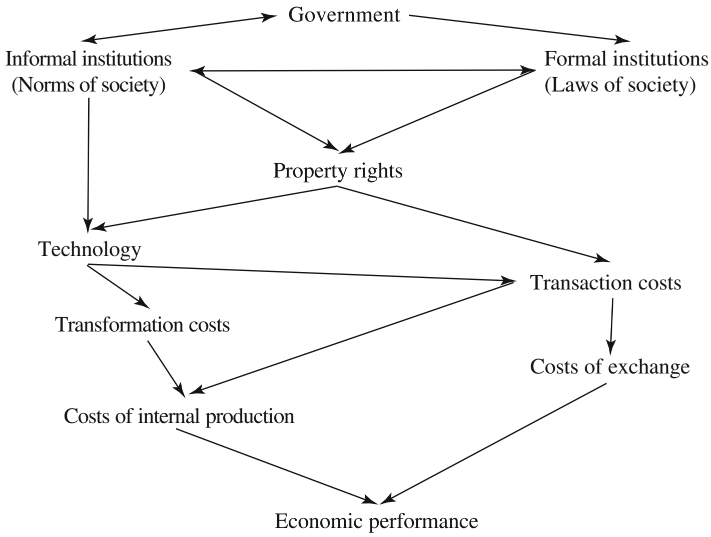
What is the new institutional economics (NIE)? The NIE adds to the neoclassical framework insights and concepts from a variety of social sciences as well as business organization, history and law. Unlike past interdisciplinary forays by economists into other disciplines, proponents of the NIE have been less imperialist and instead have been importers of various concepts. This does not mean that the NIE is internally inconsistent. Indeed, the NIE is a set of analytical spokes that when put together properly form a wheel of analysis capable of addressing a broad variety of issues. The NIE consists of analytical spokes from a variety of disciplines: anthropology, business organization, economics, history, law, political science, psychology, and sociology. My purpose in this article is to identify the spokes and try to form the wheel in order to give a better understanding of the NIE.

A Framework for Understanding the New Institutional Economics

The alpha and the omega of the NIE are institutions and economic performance (Alston and Ferrie 1999; Eggertsson 1996; North 1990). Institutions determine economic performance and economic performance determines institutions. This is nothing new. What is new are the conceptual spokes such as transaction costs, property rights, credible commitment, and agenda control that determine the simultaneous causal links between institutions and economic performance. It is important to emphasize that the NIE does not abandon neoclassical theory. As Fig. 1 illustrates, the conceptual arrows beginning with technology to transformation costs (production isoquants, along with relative prices) are still the backbone of the theory of the firm that determine the costs of production and in the neoclassical world led to discussions of how far inside and/or where on the production possibilities frontier a country would be. Because of the limited ability of this stark depiction of the theory of the firm to explain many of the 'big' questions facing economists – for example, the lack of convergence in standards of living across countries – many economists added various concepts. Let us begin with the role of institutions.

Institutions are the informal norms and formal laws of societies that constrain and shape decision-making or, as North (1990) defined them, 'the rules of the game'. For a good treatment of the interaction of norms and laws see Greif

New Institutional Economics, Fig. 1 Institutions and economic performance



(2006). For the importance of social capital or norms see Keefer and Knack (2005). Informal norms do not rely on the coercive power of the state for enforcement whereas formal laws do, in part. The enforcement of formal laws does not rely entirely on the coercive power of the state because some of their force is derived from the beliefs of its citizens For example, if more people believe that littering is morally wrong, the costs that governments incur to police littering are lower. Similarly, if more people believe that recycling is morally right then they will incur their own costs to recycle even though to do so would not be in their self-interest strictly speaking. The existence of certain laws may simply be the codification of the norms of the majority. But, at times, and particularly during crises, some political leaders can influence the norms of citizens (Higgs 1987). To the extent that political leaders can sway public opinion, the passage of laws may affect the beliefs of the constituents.

As Fig. 1 shows, the norms and laws of society determine the property rights that individuals possess. Here I am concerned with rights that individuals have in regard to goods and services: (1) the right to sell an asset; (2) the right to use and derive income from an asset; and (3) the right to bequeath an asset. Property rights are enforced in three ways. Individuals themselves enforce their assigned rights; for example, we put locks on our doors to protect our property. Societal sanctions such as ostracism can deter individuals from violating the assigned rights of others. And the coercive power of the state can be used to enforce property rights; for example, the police will evict trespassers.

Technology, which the standard neoclassical model took as exogenous, is shaped by the property rights, and the norms and endowments of citizens. Property rights along with technology determine the transaction costs and transformation costs associated with exchange and production. Robertson and Alston (1992) present a schematic framework for analysing the impact of technology on the transaction costs of production. Transformation costs are the physical costs (in an engineering sense but based also on relative prices) of combining inputs to produce output. The transformation costs of production depend on the technology in society. The transaction costs of production are the invisible costs of production and initially discussed by Coase (1937) in his seminal article for the NIE, ‘The Nature of the Firm’. Transaction costs include: (1) search and negotiation costs; (2) monitoring labour effort; (3) coordinating the physical factors of production; (4) monitoring the use of the physical and financial capital employed in the production process; and (5) enforcing the terms of the contract. It is the transaction costs within a firm – along with transformation costs – relative to the transaction costs of using the market that Coase first identified as being decisive in determining the firm/market boundary. Others within the tradition of the NIE have extended this considerably, most notably Yoram Barzel (1989) and Oliver Williamson (1985). The extensions have provided answers to issues associated with long-term contracting, for example, Goldberg and Erickson (1987); Joskow (1985); hybrid contracts of various sorts (Menard 2005) and various forms of business organization, for example, franchises (Lafontaine 1992).

Both technology and property rights can affect the transaction costs of production in a variety of ways. Technology generally reduces both the direct costs of monitoring, through better surveillance, and reduces the need to monitor, that is, capital standardizes the marginal productivity of labour, holding constant monitoring. As an historical example, in agriculture, when workers cut down weeds by hand, monitoring costs were higher than when workers drove through the fields with a mechanical cultivator that cut down the weeds. Whether on the farm or in the factory, machines by their very nature reduce the discretion of labour. They standardize the production process and thereby reduce the variation in the marginal product of labour. In addition, technology influences the transaction costs of coordinating production; for example the computer is partially responsible for the observed increase in horizontal integration in commercial banking in the United States in the 1990s. The huge merger wave in the banking industry in the 1990s was partially the result of legal changes that in turn could have been prompted by the lobbying efforts from the financial industry in recognition of the cost savings associated with the advent of computer technology.

Norms and property rights can also affect the transaction costs of production. For example, if people believe in working hard in some cultures (perhaps because of past incentives), providing ‘an honest day’s work for an honest day’s pay’, then the monitoring costs borne by the residual claimant are lower. Similarly, if the property rights in a society make it easy to dismiss workers for shirking, then monitoring costs would also decrease.

The transaction costs of exchange include the costs associated with negotiating and enforcing contracts. For some exchanges, the transaction costs of exchange are low because informal norms suffice to uphold bargains. Most local communities have well-established customs that limit opportunistic behaviour. Similarly, repeat transactions often give a sufficient incentive to deal fairly. Though local or repeat exchanges may have low transaction costs, the gains from such trade are limited because the extent of the market limits the number of individuals with whom one can deal locally or repeatedly. Formal institutions are necessary if the full gains from specialization in an extended market are to be captured. I use the term ‘full gains’ because some trade can be accomplished through self-generated reputation and the prospect of repeat business without relying on outside formal government institutions (Telser 1981). This is particularly evident in the case of international transactions where the participants do not share a common body of law. For example, the extension of the market may require that more trades occur among anonymous parties or that more trades occur where payment and delivery are not simultaneous. Institutions can reduce the potential for unscrupulous behaviour inherent in such arrangements.

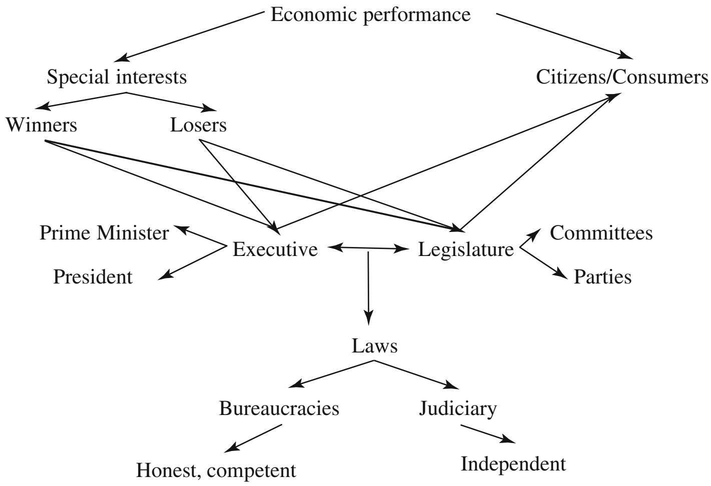
The presence of ‘honest’ courts and a body of law that upholds contracts and safeguards exchanges is a formal institution that determines the property rights of individuals which in turn affect the transaction costs of exchange. The shorthand concept used to describe this system is ‘the rule of law’ (Arrunada and Adonova 2005; Beck and Levine 2005; Hadfield 2005). This does not imply that the courts are used frequently, only that they form a backdrop for exchange. The availability of recourse to law and the courts provides a safeguard for market participants engaged in anonymous or non-simultaneous exchanges. In the absence of honest courts, negotiation and enforcement costs will be higher. As a consequence, contracts will be written in ways that will safeguard the exchange should one party desire to act opportunistically. Williamson (1985) describes how contractors shield themselves from the potential opportunistic behaviour of others. Levy and Spiller (1994) illustrate the role of institutions in providing commitment in the context of safeguarding investments in the regulation of telecommunications. Firms (and legislative and executive bodies) also use the courts strategically but here I treat firms as responding exogenously to their expectation of decisions by courts.

At times there may be insufficient safeguards so that the result is not an exchange. For example, large investments are generally required to reap economies of scale. A part of that investment may not be readily transferable to other uses (that is the investments are asset specific – see Williamson 1985, for an expansive treatment of specific assets). Before the investment is made, if there is a fear that some of the value of the investment will be expropriated, either through nationalization, taxes, regulations, or opportunistic behaviour by one of the contractors, firms will not invest as much as they would in the absence of such fears (Spiller and Tommasi 2005). Expropriation could occur either through actions taken by the state (such as regulation or nationalization) or through actions taken by one of the parties (such as refusing to execute the exchange without a renegotiation of terms).

Given the set of institutions in a society, residual claimants will construct contracts with the suppliers of inputs to minimize the sum of transformation and transaction costs within a firm, and between firms and firms and consumers. The results are a variety of contracts with differing transaction cost and production cost components, and different total costs of production. The varying contracts in turn influence economic performance. As an example there is a voluminous literature associated with principal agent problems ranging from tenancy in agriculture (Alston 2003) to corporate governance (Fama 1980).

The conceptual framework presented in Fig. 1 and discussed thus far is basically static; it illustrates the ultimate importance of institutions for economic performance but it does not address the determinants of institutions and institutional change (Alston 1996; North 2005). To understand the process of institutional change, it is useful to think about economic performance or economic growth as a process of creative destruction (Schumpeter 1942). Creative destruction means that there are winners and losers associated with economic performance (see Fig. 2). The losers have an incentive to lobby government for institutional change to protect them from the ravages of the market, while the winners have an incentive to lobby for the status quo or an even better outcome. Consumers have an interest in the outcome, but given the existence of rational ignorance and free-rider problems consumers tend not to be as effective as special interests in the political marketplace. By rational ignorance, we mean that it does not pay the consumer to be as informed about legislation as special interest groups (Olson 1965; Buchanan and Tullock 1962). The free-rider problem arises because of the large numbers of consumers have difficulties in organizing collectively to prevent policy changes. Political entrepreneurs may attenuate both these problems because the interests of consumers are represented somewhat through competition amongst politicians who bring issues to the attention of consumers, and thus limit the power of special interests (Denzau and Munger 1986).

We can think of those who lobby for changes in institutions or for the status quo as the demand side of legislation. But special interest groups do not enact legislation. Their demands get filtered through a political process of government institutions – what I call the supply side of legislation. By using the terms ‘demand’ and ‘supply’ I do not mean that there is necessarily a unique outcome; the term ‘bargaining’ may be more appropriate. Curiously, until recently, economists have paid little attention to the supply side of government, leaving the modelling of the political process to political scientists; ‘curiously’ because the concepts of demand and supply are the two most important components of neoclassical economics. The supply of legislation can be initially decomposed into the executive, legislative and judicial branches. In parliamentary systems, the executive, prime minister, and the legislature are more interconnected than in presidential systems, so that the same demands may end up with a different result depending on whether a country



New Institutional Economics, Fig. 2 The determinants of formal institutions

has a presidential or parliamentarian system (Carey 2005). Within legislatures there are a myriad of coordinating devices; historically, in the United States, political parties and the committee structure in legislatures have played major roles in shaping political outcomes (Cox and McCubbins 1993, 2005; Shepsle 1978; Shepsle and Weingast 1984). Political parties and committees have a certain amount of agenda control. For example, the party leadership makes appointments to committees, and committees in turn have the power to veto bills simply by refusing to report the bill out of committee. In addition they can amend bills to better suit their preferences. In parliamentary systems, particularly two-party dominant parliamentary systems, the majority power has significant agenda control. In other countries, most notably those with strong executive powers, such as in Brazil or Chile, the demand for legislation is filtered through the preferences of the president who negotiates with members of Congress using his powers to sway votes (Alston and Mueller 2006). Changes in either demand or supply side forces will result in institutional change. Legislation can be either specific or vague in content (Spiller 1996). In either case the law is administered through bureaucracies, giving rise to another set of principal–agent problems between the legislature and the agency to which the law is delegated (Ferejohn and Shipan 1990; Weingast and Moran 1983; McCubbins et al. 1987). In the United States, the Environmental Protection Agency (EPA) is frequently cited as an example of a bureaucracy with large discretion because of the vagueness of its mandate from Congress.

The outcomes of this demand and supply side bargaining are the formal laws and regulations of a society, subject to the explicit or implicit sanction of the courts. It matters a great deal whether the courts that interpret the constitutionality of legislation are independent of the executive and legislative branches. If the courts are truly independent the executive and legislative branches will enact legislation ‘in the shadow of the court’, knowing that the court could overturn legislation. The dismal political and economic history of Argentina since 1945 is a good example of the impact on economic performance from a Supreme Court that has not been independent (Alston and Gallo 2007; Iaryczower et al. 2002).

Where Do We Go from Here?

Before discussing institutional lock-in, the topic to which I believe we should devote more of our intellectual resources, it is worth considering which parts of the framework of the new institutional economics we know best. The hands-down winner is the area of contracting. We have much empirical evidence on how contracts change in response to different transaction costs, which in turn result from the formal laws and informal norms in societies. We also know a good deal about why governments pass the laws and regulations that they do. Here there has been an outpouring by both economists and political scientists, with economists tending to specialize in demand-side explanations – for example, the role of special interests – and political scientists specializing in supply-side explanations – for example, the role of committees and the importance of agenda control. So if we know why we get the laws, and we know how laws affect contracting, what is missing? What is missing is a better understanding of the transaction costs associated with getting laws and regulations that are more conducive to better economic performance, especially when it becomes obvious that the existing laws and regulations are not fostering economic growth (Shirley 2005). In many scenarios special interests are in a position to either enact legislation or block legislation so that they reap the gains. Yet society is worse off by such activity. The question is: why cannot 'we', the citizens or consumers, buy out the special interests? For many societies, poor economic performance is explained by corrupt governments, who are more or less stealing from their own citizens. Here we focus on issues beyond corruption, though corruption is clearly in the domain of the NIE. There are several possible explanations for institutional lock-in:

1. Informational problems abound such that citizens are unaware of possible policy moves that would improve on the status quo (North 2005 and citations therein).
2. Though citizens do not like the outcome, they approve of the process that produced the outcome.
3. Even when aware, there are serious collective action problems.
4. Insecurity in political property rights prevents transactions from occurring, that is, you cannot buy what someone else does not own.

Let us explore each of these in turn.

Given rational ignorance it may be that many citizens are simply unaware of property rights arrangements that would improve societal welfare. For example, under the Homestead Act in the United States settlers could acquire property rights to 160 acres of unoccupied federal land by residing and 'improving' the land. These homestead plots turned out to be economically too small and promoted externalities associated with wind erosion. Even after the great dust bowl of the 1930s, plots remained small because subsidies by the federal government enabled farmers to remain on the land. Why did the federal government not move to reallocate land or at least not interfere with consolidation through markets? It appears that the answer rests with the information available to citizens and their beliefs in the virtues of small landholdings. This is coupled with the efforts of local politicians to maintain a population base (Hansen and Libecap 2004a, b). Ironically, in the latter part of the 19th century Major John Wesley Powell recognized the potential problems of settlements in the arid or sub-humid regions of the country, but his reports to Congress were ignored in favour of boosterism (Stegner 1954).

Another example concerns consumers who may simply be unaware of policy moves that would improve their welfare. For example, the United States and many other countries have allowed their ocean shippers to participate in cartels that set prices on ocean routes around the globe (Sicotte 1997). When I mention this to scholars, most are unaware of this price fixing. What needs to be done is to determine how many redistributive programmes exist where a policy move would be wealth-enhancing yet does not occur either because of insufficient information or an inability of citizens to process the cause and effects of potential policy moves in the face of risk aversion, that is, we know the effect of the status and do not fully comprehend a counterfactual policy world. Many institutions are bundled in ways that makes decoupling difficult. It is partially a coordination problem and it is partially a case of risk aversion – once you open Pandora's Box you are uncertain as to the final outcome.

Another reason for institutional path dependence is circumstances where citizens have a deep belief in the process that produces laws and regulations even though they may disapprove of some legislative outcomes. The majority may opt to support the status quo legislation because changing the law would entail changing a higher order institution concerning overall institutional development. From US history an example of public disapproval of changing the system of checks and balances was the attempt by President Roosevelt to add Justices to the Supreme Court. Roosevelt wanted to stack the Court because the Court was ruling that some major legislative acts were unconstitutional, for example, the Agricultural Adjustment Act and the National Industrial Recovery Act. By adding Justices, Roosevelt believed that his New Deal legislation would pass the constitutional test. Even though most people supported the New Deal legislation, there was a public outcry against Roosevelt's attempt to change the rules affecting checks and balances so as to achieve his legislative goals.

Alternatively, people may be aware of the dissipation associated with the status quo arrangement of property rights, but it is in no one's self-interest to mount an organizational campaign to change the existing regulations. This is the classic collective action problem developed independently but almost simultaneously by Buchanan and Tullock (1962) and Olson (1965) – one could also model this as a multi-player Prisoner's Dilemma game. The collective action problems are particularly acute in situations entailing multiple governments across international boundaries, for example, overfishing in international waters or global warming. The difficulties for international property rights are twofold: specification and enforcement. Specification is difficult because of knowledge or beliefs about the state of the world differ (for example, global warming) but even if beliefs are the same, preferences can vary across countries because of incomes (for example, the United States versus Mexico) or simply preferences (for example, the United States versus Germany on green issues). Collective action problems occur in representative democracies as well as dictatorial regimes. We have instances of both types of regimes not specifying and enforcing property rights at what would appear to be optimal times. For example, the United States squandered considerable oil reserves in the early 20th century and Indonesia mowed through a large stock of their tropical hardwoods in the latter part of the 20th century.

A fourth possibility for the lack of policy reform is insecure political property rights. It may be that individuals are aware and willing to organize but there is no 'market' for the emergence of property rights. Suppose that the winners from a status quo policy have the political power to veto or allow policy changes. Given their power, they would be foolish to acquiesce to policy moves that made them worse off, even if they were wealth enhancing. But, they would allow such a policy move if they were compensated. The actions of the landless peasants' movement (MST) in Brazil are consistent with this argument. The MST is very effective at swaying public opinion and thereby prompting politicians to expropriate land and transfer it to peasants; but they do not support deeding the land to peasants. The MST prefers to keep the peasants dependent on the MST as a collective because it is easier for them to extract payments from the group than individual farmers (Alston et al. 2005).

Why is it that we generally do not allow such side payments? One answer is that transparent side payments would undermine the legitimacy of the organization, whether the organization is the MST, a union or a government. If the current property rights arrangement is viewed as inferior to an alternative, people 'believe' that they should not pay to move to a better property rights arrangement. The result is institutional lock-in. Yet there have been examples of improving the status quo for all parties involved. A case involving the sale of water in the 1990s illustrates the difficulties in changing the status quo. The Imperial Valley Irrigation District, a governmental unit that has jurisdiction over water, entered into a contract to sell some of its water to the city of San Diego. The Imperial Valley Water District has property rights to water that are subsidized by US taxpayers. As such it can sell water at prices higher than it pays. Interestingly, members of the Imperial Water District decided that they would only sell water that they have conserved through better irrigation technologies. The interesting question is: why didn't they fallow all their land and sell their entire water allocation? I speculate that they were concerned about the political fallout that could have resulted in the district losing its current subsidy. In short, it appears as if they have secure property rights to the rental stream of water but not the clear ‘political’ property right to the stock. The establishment of ‘water banks’ throughout the West – whereby farmers could sell their flow of water to urban users or resort users – have failed primarily because farmers are afraid of losing their property right when it becomes transparent that farming is not the highest-valued use of water in the West.

Another factor promoting the insecurity of political property rights falls under the rubric of credible commitment (North and Weingast 1989). In representative democracies politicians face the demands of constituents who may be harmed or obtain benefits from a rearrangement of property rights. The demands of the majority of voters may not coincide with the optimal arrangements of property rights, and politicians cannot commit to making side payments over time to compensate the losers. Authoritarian regimes are subject to similar problems associated with catering to populist demands. A good example of this was the infringement in property rights by Peron in Argentina in the late 1940s. Peron imposed rent and price controls in the Pampas, the most fertile and productive agricultural producing area in Argentina. The punitive arrangement in property rights lead to a decline in investment which, along with political instability, affected growth in the long run (Alston and Gallo 2007; Spiller and Tommasi 2003, 2007).

A more cynical view of political behaviour suggests that we do not want to encourage paying for changes in property rights because to do so would promote the creation and maintenance of non-optimal property rights in order to be paid to move to a more optimal situation. Campaign finance and corruption around the globe may be testimony to special interests trying to ‘bribe’ politicians to maintain or change property rights. In some instances politicians may use part of the contributions to make side payments (Norlin 2003).

Explaining institutional rigidities in the face of poor economic performance is a difficult research agenda. To understand the lock-in requires insights from the disciplines that comprise the NIE – anthropology, business organization, economics, history, law, political science, psychology and sociology. Yet the potential reward from an understanding of the forces that account for poor economic performance is huge. The research agenda includes both international cross-sectional studies and case studies of successful and unsuccessful institutional change. The international cross-sections allow us to quickly determine the correlates of successful economic performance, for example secure property rights, while the case studies allow us to stack the building blocks that will ultimately allow us to produce a more general framework for the determinants of institutional change (Alston 2007).

See Also

- ▶ Firm, Theory of the
- ▶ Growth and Institutions
- ▶ ‘Political Economy’ and ‘Economics’
- ▶ Political Institutions, Economic Approaches to
- ▶ Property Rights
- ▶ Transaction Costs, History of

Bibliography

Alchian, A., and H. Demsetz. 1972. Production, information costs, and economic organization. *American Economic Review* 62: 777–795.

Alston, L.J. 1996. Empirical work in institutional economics. In *Empirical studies in institutional change*, ed. L.J. Alston, T. Eggertsson, and D. North. New York: Cambridge University Press.

Alston, L.J. 2003. Tenant farming. In *The Oxford encyclopedia of economic history*, ed. J. Mokyr, Vol. 5. New York: Oxford University Press.

Alston, L.J. 2007. The ‘case’ for case studies zin the new institutional economics. In *New institutional economics: A guidebook*, ed. J. Glachant and E. Brousseau. Cambridge: Cambridge University Press.

Alston, L.J., T. Eggertsson, and D.C. North. 1996a. *Empirical studies in institutional change*. New York: Cambridge University Press.

Alston, L.J., and J.P. Ferrie. 1993. Paternalism in agricultural labor contracts in the U.S. South: Implications for the growth of the welfare state. *American Economic Review* 83: 852–875.

Alston, L.J., and J.P. Ferrie. 1999. *Southern paternalism and the rise of the welfare state: Economics, politics, and institutions in the U.S. South, 1865–1965*. New York: Cambridge University Press.

Alston, L.J. and Gallo, A. 2007. *Electoral fraud, the rise of Peron, and demise of checks and balances in Argentina*. Working Paper No. ES2007-0001, Institute of Behavioral Science, University of Colorado.

Alston, L.J., G.D. Libecap, and B. Mueller. 1999. *Titles, conflict and land use: The development of property rights and land reform in the Brazilian Amazon Frontier*. Ann Arbor: University of Michigan Press.

Alston, L.J., G.D. Libecap, and B. Mueller. 2000. Land reform policies: The sources of violent conflict, and implications for deforestation in the Brazilian Amazon. *Journal of Environmental and Economics and Management* 39: 162–188.

Alston, L.J., Libecap, G. and Mueller, B. 2005. *How interest groups can influence political outcomes indirectly through information manipulation: The landless peasants movement in Brazil*. Working Paper No. EB2005-0005, Institute of Behavioral Science, University of Colorado.

Alston, L.J., G.D. Libecap, and R. Schneider. 1996b. The determinants and impact of property rights: Land titles on the Brazilian frontier. *Journal of Law, Economics and Organization* 12: 25–61.

Alston, L.J., and B. Mueller. 2005. Property rights and the state. In Menard and Shirley (2005).

Alston, L.J., and B. Mueller. 2006. Pork for policy: Executive and legislative exchange in Brazil. *Journal of Law Economics and Organization* 22: 87–114.

Anderson, T.L., and P.J. Hill. 2004. *The not so wild, wild west: Property rights on the frontier*. Stanford: Stanford University Press.

Anderson, T.L., and F.S. McChesney. 2003. *Property rights: Cooperation, conflict and law*. Princeton: Princeton University Press.

Arrunada, B., and V. Andonova. 2005. Market institutions and judicial rule-making. In Menard and Shirley (2005).

Barzel, Y. 1989. *Economic analysis of property rights*. New York: Cambridge University Press.

Barzel, Y. 2002. *A theory of the state: Economic rights, legal rights and the scope of the state*. New York: Cambridge University Press.

Beck, T., and R. Levine. 2005. Legal institutions and financial development. In Menard and Shirley (2005).

Buchanan, J.M., and G. Tullock. 1962. *The calculus of consent: Logical foundations of constitutional democracy*. Ann Arbor: University of Michigan Press.

Carey, J.M. 2005. Presidential versus parliamentary government. In Menard and Shirley (2005).

Coase, R.H. 1937. The nature of the firm. *Economica* 4: 386–405.

Coase, R.H. 1960. The problem of social cost. *Journal of Law and Economics* 3: 1–44.

Coase, R.H. 1992. The institutional structure of production. *American Economic Review* 82: 713–719.

Cox, G.W., and M.D. McCubbins. 1993. *Legislative leviathan: Party government in the house*. Berkeley: University of California Press.

Cox, G.W., and M.D. McCubbins. *2005*. *Setting the agenda: Responsible government in the U.S. house of representatives*. Cambridge: Cambridge University Press.

Crocker, K.J., and S.E. Masten. 1991. Pretia ex machina? Prices and process in long-term contracts. *Journal of Law and Economics* 34: 69–99.

Denzau, A.T., and M. Munger. 1986. Legislators and interest groups: How unorganized interests get represented. *American Political Science Review* 80: 84–106.

Eggertsson, T. 1990. *Economic behavior of institutions*. New York: Cambridge University Press.

Eggertsson, T. 1996. A note on the economics of institutions. In *Empirical studies in institutional change*, ed. L.J. Alston, T. Eggertsson, and D. North. New York: Cambridge University Press.

Fama, E.F. 1980. Agency problems and the theory of the firm. *Journal of Political Economy* 88: 283–307.

Ferejohn, J., and C. Shipan. 1990. Congressional influence on bureaucracy. *Journal of Law, Economics, and Organization* 6: 1–27.

Furubotn, E.G., and R. Richter. 2005. *Institutions and economic theory: The contribution of the new institutional economics*. Ann Arbor: University of Michigan Press.

Goldberg, V.P., and J.R. Erickson. 1987. Quantity and price adjustment in long-term contracts: A case study in petroleum coke. *Journal of Law and Economics* 31: 369–398.

Greif, A. 2006. *Institutions and the path to the modern economy*. New York: Cambridge University Press.

Hadfield, G.K. 2005. The many legal institutions that support contractual commitments. In Menard and Shirley (2005).

Hansen, Z.K., and G.D. Libecap. 2004a. The allocation of property rights to land: US land policy and farm failure in the northern great plains. *Explorations in Economic History* 41: 103–129.

Hansen, Z.K., and G.D. Libecap. 2004b. Small farms, externalities, and the dust bowl of the 1930s. *Journal of Political Economy* 112: 665–694.

Higgs, R. *1987*. *Crisis and leviathan: Critical episodes in the growth of American government*. New York: Oxford University Press.

Iaryczower, M., P. Spiller, and M. Tommasi. 2002. Judicial independence in unstable environments, Argentina 1935–1998. *American Journal of Political Science* 46: 699–716.

Joskow, P.L. 1985. Vertical integration and long-term contracts: The case of coal- burning electric generating plants. *Journal of Law, Economics and Organization* 1: 33–80.

Keefer, P., and S. Knack 2005. Social capital, social norms and the new institutional economics. In Menard and Shirley (2005).

Lafontaine, F. 1992. Agency theory and franchising: Some empirical results. *Rand Journal of Economics* 23: 263–283.

Levy, B., and P.T. Spiller. 1994. The institutional foundations of regulatory commitment: A comparative analysis of telecommunications regulation. *Journal of Law, Economics and Organization* 10: 201–246.

Libecap, G.D. 1989. *Contracting for property rights*. New York: Cambridge University Press.

Libecap, G.D., and Z.K. Hansen. 2003. *Small farms, externalities, and the dust bowl of the 1930s*. Working Paper No. 04-01, Department of Economics, University of Arizona.

McCubbins, M., R. Noll, and B. Weingast. 1987. Administrative procedures as instruments of political control. *Journal of Law, Economics, and Organization* 3: 243–277.

Menard, C. 2005. A new institutional approach to organization. In Menard and Shirley (2005).

Menard, C., and M.M. Shirley. 2005. *Handbook of new institutional economics*. Dordrecht: Springer.

Norlin, K. 2003. *Political corruption: Theory and evidence from the Brazilian experience*. Ph.D. thesis, University of Illinois.

North, D.C. 1990. *Institutions, institutional change, and economic performance*. New York: Cambridge University Press.

North, D.C. 2005. *Understanding the process of economic development*. Princeton: Princeton Economic Press.

North, D., and B. Weingast. 1989. Constitutions and commitment: The evolution of institutions governing public choice in seventeenth-century England. *Journal of Economic History* 49: 803–832.

Olson, M. 1965. *The logic of collective action*. Cambridge, MA: Harvard University Press.

Ostrom, E. 1990. *Governing the commons: The evolution of institutions for collective action*. New York: Cambridge University Press.

Robertson, P.L., and L.J. Alston. 1992. Technological change and the organization of work in capitalist firms. *Economic History Review* 45: 330–350.

Schumpeter, J. 1942. *Capitalism, socialism, and democracy*. New York: Harper.

Shepsle, K. 1978. *The giant jigsaw puzzle: Democratic committee assignments in the modern house*. Chicago: University of Chicago Press.

Shepsle, K.A., and B.R. Weingast. 1984. Legislative politics and budget outcomes. In *Federal budget policy in the 1980s*, ed. G.B. Mills and J.L. Palmer. Washington, DC: Urban Institute Press.

Shirley, M.M. 2005. Institutions and development. In Menard and Shirley (2005).

Sicotte, R. 1997. *The organization and regulation of international shipping cartels*. Unpublished Ph.D. thesis, University of Illinois.

Smith, H.E. 2000. Semicommon property rights and scattering in the open fields. *Journal of Legal Studies* 29: 131–170.

Spiller, P.T. 1996. A positive theory of regulatory instruments: Contracts, administrative law or regulatory specificity. *USC Law Review* 69: 477–515.

Spiller, P.T., and M. Tommasi. 2003. The institutional foundations of public policy: A transactions approach with application to Argentina. *Journal of Law Economics and Organization* 19: 281–306.

Spiller, P.T., and M. Tommasi 2005. The institutions for regulation: An application to public utilities. In Menard and Shirley (2005).

Spiller, P.T., and M. Tommasi. 2007. *The institutional foundation of public policy in Argentina*. New York: Cambridge University Press.

Stegner, W.E. 1954. *Beyond the hundredth meridian: John Wesley Powell and the second opening of the west*. New York: Penguin.

Telser, L. 1981. A theory of self-enforcing agreements. *Journal of Business* 53: 27–44.

Weingast, B.R., and W. Marshall. 1988. The industrial organization of congress: Or why legislatures like firms are not organized as markets. *Journal of Political Economy* 96: 132–163.

Weingast, B.R., and M.J. Moran. 1983. Bureaucratic discretion or congressional control? Regulatory policymaking by the FTC. *Journal of Political Economy* 91: 765–800.

Williamson, O. 1985. *The economic institutions of capitalism*. New York: Free Press.

Williamson, O. 2000. The new institutional economics: Taking stock, looking ahead. *Journal of Economic Literature* 38: 595–613.

New Keynesian Macroeconomics

Huw David Dixon

Abstract

The term ‘new Keynesian economics’ refers to a body of work done by macroeconomists in the late 1970s and 1980s in which the notion of imperfect competition was introduced into macroeconomics in order to provide a microfoundation for nominal rigidities and also to provide an alternative to supply-equals-demand equilibrium. This led in the 1990s to the new-neoclassical-synthesis approach to monetary economics in which dynamic pricing models have become central to our understanding how monetary policy influences output and inflation. Other themes in the new Keynesian approach include the effect of imperfect competition on the fiscal multiplier, and coordination failures.

Keywords

Adaptive expectations; Aggregate supply and demand models; Balanced budget multiplier; Competitive equilibrium; Constant hazard rate models; Consumer surplus; Coordination failure; Efficiency wages; Envelope theorem; Extrinisic and intrinsic uncertainty; Hybrid Phillips curve; Imperfect competition; Inflation; Invisible auctioneer; IS–LM model; Keynesian Revolution; Menu costs; Microfoundations; Monetary transmission mechanism; Monopolistic competition; Monopoly; Neoclassical synthesis; Neutrality of money; New classical macroeconomics; New Keynesian macroeconomics; New Keynesian Phillips curve; New neoclassical synthesis; Nominal rigidities; Perfect competition; Phillips curve; Price indexation; Price making; Price rigidity; Price taking; Rational expectations; Real business cycles; Real rigidities; Representative agent; s-S models; Staggered price-setting; Staggered wage-setting; State-dependent pricing; Strategic complementarity; Taylor principle; Time-dependent models; Wage indexation; Wage rigidity

JEL Classifications

E12

The term ‘new Keynesian economics’ came into popular usage in the 1980s. The origins of the term are fairly easy to understand in broad historical terms. In the classical approach of the pre-Keynes world (prior to 1936), wages and prices were seen as perfectly flexible and markets competitive (or at least ideally so). The Keynesian Revolution argued that prices, and more importantly wages, were rigid, and in order to understand phenomena like prolonged mass unemployment it was necessary to see how the economy operated when not in competitive equilibrium. In the post-Second World War period there emerged the neoclassical synthesis model that dominated macroeconomics from the 1950s to the mid-1970s. The essence was that in the *long run* all prices are perfectly flexible and the competitive or ‘Walrasian’ equilibrium will hold. However, in the *short run* prices and/or wages were treated as given. Thus there were the IS–LM and aggregate supply and demand (AS–AD) models, which were the workhorses of macroeconomic research until the mid-1970s and have remained established in many textbooks to the present day.

This approach was in the process of being overtaken at the level of research by the ‘new classical’ or rational expectations revolution of the 1970s. One aspect of the neoclassical synthesis was that not only prices but also expectations were treated as fixed in the short run, or subject to ad hoc adjustment, as under the adaptive expectations hypothesis. The new classical approach was based on the idea that wages and prices are perfectly flexible, but that agents did not have full information: even though agents used the information they had optimally (rational expectations), markets could deviate from the full information equilibrium. For example, agents might not know about the values of certain current variables such as aggregate price or the money supply when deciding how much output to produce or labour to supply.

The new Keynesian economics was to incorporate the rational expectations framework. However, it was to focus on the key issue of nominal rigidity: how do we understand the short-term rigidity of wages and/or prices in terms of providing a microfoundation that will explain why prices might not be perfectly flexible? Now, this required a ‘revolution’ of the order of magnitude of the rational expectations revolution. That revolution consisted in one idea: in order to understand nominal rigidity, it was necessary to abandon the approach of perfect competition with price- taking agents, and replace it with an approach where there are wage and price-setting agents. This is self-evident in hindsight: if you want to understand why wages and prices are rigid in the short run, you have to have agents who set the price, so that you can understand the microeconomics of price adjustment. If all agents (firms, households) are price-takers, prices can only be explained by some notion of ‘demand equals supply’ and a shadowy Walrasian auctioneer acting like an invisible puppet master-cum-market maker, adjusting prices gradually in response to excess demand or supply. This is hardly the basis for a rigorous theory of why prices and wages are not always at their market clearing levels: maybe the auctioneer called in sick or went on holiday!

Just to complete the historical setting, alongside the new Keynesian ideas there was the real business cycle (RBC) research programme which put forward the radical idea that nominal wage and price behaviour were irrelevant for understanding macroeconomic dynamics. Changes in output and employment were seen to be driven by real things such as productivity shocks, and the savings and investment decisions of agents as inherently dynamic. This was a radical agenda, which also pushed macroeconomics into trying to provide a quantitative explanation of economic fluctuations based on a competitive equilibrium model. However, despite many successes, the methodological idea of ignoring nominal things was an unsustainable self-limitation. For one thing, governments and central bankers are interested in the nominal side of the economy – inflation, the transmission mechanism of monetary policy, to name a few.

So, in the mid-1990s there emerged the 'new' neoclassical synthesis (NNS). This combined the dynamic framework of the RBC approach with dynamic pricing models developed by the new Keynesian approach. The key idea is that in the long run money is neutral, but in the short run there is some nominal rigidity resulting from the price-setting behaviour of firms (and wage-setting behaviour of unions). This approach to modelling has certainly become the dominant school of thought, at least in central banks of Europe and the United States. It differs from the old neoclassical synthesis in that the model is fully dynamic and microfounded and the equilibrium imperfectly competitive.

The Microfoundations of Wage and Price Rigidity

So the problem in the late 1970s and early 1980s was clear. Most of economics was based on models of perfect competition, where all agents are price-takers. An agent is a 'price-taker' if it believes that it can trade any quantity at the market price which it treats as given, or exogenous. Price-taking makes sense only when markets clear, and supply equals demand. If supply does not equal demand, then something has got to give because the chosen trades do not add up to zero. An alternative was needed. Up until then, various ad hoc assumptions had been made: the simplest was that wages and/or prices were simply assumed to be fixed (this was justified by the notion that the model was a short-run model). Another ad hoc fix was that the market was competitive but that the price cleared the market *ex ante*: the invisible auctioneer sets the price which he or she expects to clear the market before it opens. The basic and fundamental new Keynesian insight was that the assumption of price- taking behaviour had to be abandoned. Real agents such as firms, households or unions needed to be price-makers. But this meant that the notion of perfectly competitive equilibrium had to be abandoned: the alternative was going to be an imperfectly competitive equilibrium where (some) agents have market power. The classic imperfectly competitive equilibrium is pure monopoly: a monopolist can set any price he pleases, and will maximize profits. The monopolist equates marginal revenue with marginal cost: if he faces a downward sloping demand curve, this means that the monopolist will set a price above the competitive price and output will be lower than in the competitive equilibrium. While the firm increases its profits there is also a decline in consumer surplus and the total surplus (consumer plus producer) declines.

In the absence of market failure, the perfectly competitive equilibrium is Pareto optimal. If we are adopting a representative agent framework (as has most often been the case in macroeconomics since the neoclassical synthesis), Pareto optimality means that the equilibrium outcome maximizes the utility of the representative agent. Hence, if we look at small deviations from equilibrium (in terms of output, employment and so on), they will not have a first-order effect on welfare. This is an envelope theorem: the first-order conditions for optimality state that the first-order effect is zero at the optimum. With imperfect competition, by contrast, we start away from the optimum. Hence there are first-order effects of changes in output and employment: since the monopolist restricts output, an increase is good and a decrease bad. To many macroeconomists, this seems more plausible and common sense than the implication of the first welfare theorem that holds that, if one starts from the competitive equilibrium, increases and decreases in output and employment are both (slightly) bad.

The introduction of imperfect competition into a tractable general equilibrium framework (albeit a static one) was achieved by Oliver Hart (1982), who stressed the 'Keynesian features' of the model. However, Hart's was a real model without money: what was needed was to link this idea to nominal rigidity. It was a few years later that the concept was taken up simultaneously in three papers: Akerlof and Yellen (1985), Mankiw (1985) and Parkin (1986). The new idea was that of 'menu costs', whereby there might be 'costs' to changing a price, which might be interpreted broadly as decision and implementation costs (the line taken by Akerlof and Yellen and interpreted as a sort of bounded rationality) or as literally the cost of implementing a price change (having new menus printed). This idea was not new: it was used by the (S,s) models of pricing with inflation developed in the 1970s by Sheshinski and Weiss (1977), and in some other papers in the non-macroeconomic literature.

The insight is that if a monopolist sets its price optimally, a small deviation from the optimum will have no first-order effect on *profits*. If there is a small but lump-sum cost of changing a price, then the effect of a price-setting monopolist to an increase in demand (or cost) might be to leave the price where it is, not to change it. Thus, even small menu costs can give rise to some nominal rigidity: because at the optimum there is no first-order effect on profits, the menu costs only have to overcome the smaller higher order effects. Thus began a theory of nominal rigidity based on monopolistic competition and menu costs. The nice feature of the model was that, although the menu costs could be small, the nominal rigidity they created would give rise to first-order welfare effects (since we start from a level of output and employment below equilibrium). Whilst the idea is very simple and powerful, it did alas run into a problem. In static models it is easy to use the menu-cost approach. However, macroeconomists in the 1980s were interested in dynamic models, and menu-cost models have proven very difficult to solve except under very special cases. For example, Caplin and Spulber (1987) looked at steady-state inflation and found that although the menu costs caused individual firms to have prices that remained fixed for a time, in aggregate prices they drifted up, with the aggregate money supply yielding the same aggregate output and inflation as with flexible prices. It has only been much later, since the late 1990s, that these models are beginning to be solved for interesting dynamic cases (under the new name 'state- dependent pricing' models).

However, the menu-cost idea spawned a large literature that looked into how certain features of the economy might allow even smaller menu costs to give rise to nominal rigidity. For example, Ball and Romer (1990) argued that if there were some real rigidity in the economy, it would interact with the nominal rigidity of prices, reducing the size of menu costs required to induce nominal rigidity. The real rigidity might take the form of an efficiency wage model, for example, where the equilibrium determined the real wage which was not sensitive to the level of economic activity. On the empirical level, Ball et al. (1988) argued that the menu-cost theory had a clear prediction for the relation between inflation and the inflation–output trade-off. If steady-state inflation was higher, this would mean that for a given level of menu costs, firms would change prices more frequently (there is less nominal rigidity). This in turn would mean that changes in nominal demand would have less effect on output when inflation is higher. Thus the non-neutrality of money in the short run was higher in low-inflation economies than in high-inflation economies, which was confirmed in the data.

Whilst there has been until recently quite some difficulty in making state- dependent or menu-cost models tractable enough to model wage and price dynamics out of steady state, another class of models proved well suited to a dynamic setting. These were the *time-dependent* models of pricing, which focused on the notion of staggered wage-

and price-setting: Taylor (1979) and Calvo (1983). Indeed, these two models have become the workhorses of the NNS framework. John B. Taylor's model focused on wage-setting: the empirical evidence suggests that many wage contracts take the form of a nominal wage being set for a period of four quarters. However, wages in different sectors are negotiated at different times. It is usually assumed that there are four equally sized cohorts, one cohort resetting the wage each quarter. Whilst this framework does not explain why wage contracts last for a particular period, it does start out from firm empirical observation and works out the implications of this for the resultant process. What we find is that wages gradually adjust to their new steady state values. The reason for this is that when setting wages the current cohort is facing an aggregate price level partly determined by cohorts that have moved previously. At any one time, with four cohorts, three cohorts will not reset the wage: they reset their wages in the previous three quarters. When the union sets its wage, it looks at what the aggregate price level and demand will be over the period of the contract: in this sense the wage-setting rule is dynamic and forward looking. However, it is also looking back at the previous wages insofar as they are reflected in the current price. This results in a gradual adjustment of wages and prices in response to a nominal shock. Taylor (1999) provides a good survey of this approach.

Calvo's model of nominal rigidity is based on a constant hazard rate model: each period, the firm or union faces a given probability of resetting its price or wage. The expected duration of the price or wage when it is set is the reciprocal of the reset probability. When the firm sets its price it looks into the infinite horizon, and takes into account the future price with the probability that the current price being set will still be in force. Thus, if the reset probability is 0.25 per quarter, we will observe 25 per cent of firms resetting price in any one quarter. In setting the price, each firm expects that the price will last for four quarters, but there is en ever diminishing probability that it might last ever longer. If we look across all firms, the average contract length will be about twice the life expectation at birth (twice the life expectation at birth minus 1). Thus a reset probability of 0.25 implies an average lifetime of prices set by all firms across the economy of seven quarters (see Dixon and Kara 2006). The firms choose an optimal price in a dynamic setting, but the setting itself leaves the fundamental probability of resetting the price unexplained. However, the model is highly tractable and has since become very popular.

Other New Keynesian Themes

Whilst the theoretical microfoundation of nominal rigidity was the main theme of the new Keynesian economics, other themes aimed to establish the implications of imperfect competition and other market imperfections as an alternative equilibrium concept to perfect competition.

One theme that ran through the new Keynesian literature that did not involve nominal rigidity was the effect of imperfect competition on the government expenditure multiplier. Papers by Dixon (1987) and Mankiw (1988) found that in simple general equilibrium models an increase in the degree of imperfect competition reflected in a bigger markup of price over marginal cost meant that the balanced budget government multiplier was bigger. The intuition behind this result was that there was a profit feedback effect: as output increased, so did firms' profits, which were paid back to households in dividends, part of which were spent again, and so on. This feedback effect was bigger than the markup. In a constant returns to scale world, there were no profits in a perfectly competitive equilibrium, so the effect was completely absent. In a follow-up paper, Startz (1989) argued that whilst the Dixon–Mankiw result held in the short run with a fixed number of firms, in the long run free entry would eliminate profits and the relationship between profits and the multiplier would disappear. This argument turned out to be true in general only in the case of constant returns to scale. The point is that when you allow for a concave production function with diminishing marginal product of labour, a second mechanism comes into effect: as employment rises, the real wage falls, which tends to reduce consumption. In the Walrasian case of perfect competition, the real wage effect always dominates the profit effect: the long-run multiplier with free entry is always greater than the short-run multiplier. It follows that if there is only a little imperfect competition, this will still be true, as shown in Dixon and Lawler (1996). Startz's result holds because with a constant marginal product of labour the real wage mechanism is absent and only the profit feedback is present.

It should be noted that the fiscal multiplier is still always less than unity. What is happening is that in equilibrium imperfect competition leads to lower real wages (the markup in the product market leads to real wages being below the marginal product). Households react to this by choosing more leisure and less consumption for any given utility level (the level of economic activity is below the perfectly competitive level). Now, an increase in government expenditure financed by a lump-sum tax makes the household worse off; so the household reacts by reducing its consumption and leisure (less leisure means working harder). The reason the short- run multiplier tends to be larger when there is more imperfect competition is that the equilibrium ratio of leisure to consumption is larger, so the effect of the tax on labour supply is larger, resulting in a bigger overall increase in labour supply and hence less crowding out of consumption. The mechanism underlying this is essentially a supply side effect, which is not exactly what some people might think of as 'Keynesian'.

The notion of 'coordination failure' was also important in the new Keynesian thought. The idea arose out of the concept of strategic complementarity. Strategic complementarity occurs when the marginal benefit from the action of one agent is increasing in the level of activity chosen by other agents. Effectively, the reaction functions are upward sloping. Cooper and John (1988) applied this idea to several macroeconomic applications, including search models and demand spillovers in multi-sector economies, and the subsequent literature has applied this concept to almost any model with positive externalities. One interesting feature of the coordination failure approach is that there may be multiple equilibria: if this is so and the equilibria are symmetric the equilibria will be Pareto ranked. With positive externalities the high activity equilibria will Pareto dominate the low-level equilibria. The existence of multiple equilibria is not easy to establish: it requires as a necessary condition that the slope of the reaction function must be greater than 1 for some values in between the two symmetric equilibria.

In the labour market, there were several developments in the new Keynesian literature. Perhaps the most important was the development of efficiency wage models. Whilst the model of efficiency wages had a long pedigree, it was seen as a way of modelling how firms might set wages at a level different from the competitive level. In Shapiro and Stiglitz (1984), the internal monitoring problem faced by the firm is influenced by the level of unemployment, since the higher the level of unemployment the costlier it is for an employee to lose his or her job. Unemployment can therefore act as a disciplining device. This model predicts that firms will be forced to pay workers a higher wage when unemployment is lower, leading to a theoretical explanation of pro-cyclical wages.

The New Neoclassical Synthesis (NNS)

In the 1990s, the new Keynesian ideas become part of the NNS approach, which is a combination of the dynamic structures developed by the RBC theory with a nominal side to the economy, which is based on imperfect competition and nominal rigidity. One of the main contributions has been the new Keynesian Phillips curve: this can be derived from both the Calvo and Taylor models of dynamic pricing (see Roberts 1995). The equation relates current inflation to current output and expected inflation next period

$$\pi_t = \beta E_t \pi_{t+1} + \kappa y_t$$

where inflation is π_t, the discount rate is β and output (deviation from capacity) is y_t. This differs from the traditional Phillips curve in which the expectation of current inflation appears on the right-hand side. The coefficient on the output gap is related to the probability the firm can reset its price, the discount rate and a parameter capturing the sensitivity of marginal cost to output.

Empirically, the new Keynesian Phillips curve has not done very well. The evidence seems to support the idea that lagged inflation needs to be included as well (resulting in the so-called 'hybrid Phillips curve'). This has led to the idea that indexation might be important: in the periods when firms cannot set prices or wages explicitly, they are updated by a 'rule of thumb' using last periods inflation rate (see Christiano et al. 2005) which results in a hybrid Phillips curve.

The Keynesian notion of demand management is very much at the centre of the analysis of monetary policy: the central bank is seen as using interest rate policy to stabilize the economy in two senses. The overall policy design should be to stabilize expectations and rule out explosive or indeterminate solutions: the possibility of economic turbulence caused by sunspot equilibria is seen as welfare reducing and is to be avoided (this is called *extrinisic uncertainty*). Thus policy should give rise to a unique rational expectations equilibrium path. In most models, a necessary condition for a unique equilibrium path is that the interest rate policy satisfies the *Taylor principle*, which states that if nominal inflation rises the central bank should raise the nominal interest rate by more, so that the real interest rate rises. Monetary policy should also be designed to stabilize the economy in response to real shocks, the *intrinsic* uncertainty facing the economy. This has been dubbed by some the 'science of monetary policy' (see Clarida et al. 1999). Of course, the new Keynesian science is different from the old Keynesian art in that the interest rate is the only instrument and fiscal policy is reduced to providing a prudent and sustainable regime of expenditure and taxation. But the view is still Keynesian in that the economy needs and benefits from having an active monetary policy.

An Evaluation

The most lasting legacy of the new Keynesian economics was to put imperfect competition and non-competitive models at the heart of macroeconomics. For a long time many economists had been impatient with the assumption of market clearing/demand equals supply as a basis for macroeconomics. However, a quest for a rigorous and consistent alternative was in place since Keynes's *General Theory* in 1936 raised more questions than it had answered. Whilst the book had given rise to the notion of using fiscal and monetary policy to stabilize the economy, this remained a practical art without a proper theoretical framework to underpin it. The macroeconomic theory developed was not consistent with standard microeconomics and was in this sense unsatisfactory. The real achievement of the new Keynesian literature was to provide the theoretical alternative to demand and supply economics that was rigorous and microfounded. Economics has always been ideological as well as scientific. There are those free market ideologues who believe that the free market is almost always the best and that the state should intervene as little as possible in the market. There are also those who believe that although markets are pretty good at many things, they can also malfunction and so maybe there is a role for some sort of public policy. In macroeconomics this polarity was at its most obvious in the 1980s and 1990s. The real business cycle theorists used models with perfect markets and were largely of the 'free-market' variety of economists. The new Keynesian economics provided a rigorous alternative to the free-market perspective and as such has left a lasting legacy which we can see is firmly embedded in the way nominal rigidity is understood and monetary policy is practiced.

Further Reading

Insofar as there is a defining book of the new Keynesian macroeconomics, it is Mankiw and Romer's two-volume collection (1991). Some good surveys were made in the early 1990s: Gordon (1990) is one; Silvestre (1993) focuses on the issue of imperfect competition; Dixon and Rankin (1994) focused more on the implications for macroeconomic policy issues. There was also a *Journal of Economic Perspectives* symposium on 'Keynesian Economics Today' in 1993 (volume 7, number 1) which takes a broader view of new and old Keynesian macroeconomics.

On the NNS approach, the monetary policy aspects are well surveyed by Clarida et al. (1999), and for text book treatment of the modelling foundations turn to Walsh (2003, ch. 5) and Woodford (2003, ch. 3). There is also an excellent survey of several NNS models of nominal rigidity in Ascari (2003).

See Also

▶ Microfoundations
▶ Real Business Cycles
▶ Real Rigidities

Bibliography

Akerlof, G., and J. Yellen. 1985. A near-rational model of the business cycle with wage and price inertia. *Quarterly Journal of Economics* 100: 823–838.

Ascari, G. 2003. Price and wage staggering: A unifying framework. *Journal of Economic Surveys* 17: 511–540.

Ball, L., and D. Romer. 1990. Real rigidities and the non-neutrality of money. *Review of Economic Studies* 57: 179–198.

Ball, L., N.G. Mankiw, and D. Romer. 1988. The new Keynesian economics and the output-inflation trade-off. *Brookings Papers on Economic Activity* 1: 1–82.

Calvo, G. 1983. Staggered prices in a utility-maximizing framework. *Journal of Monetary Economics* 12: 383–398.

Caplin, A., and D. Spulber. 1987. Menu costs and the neutrality of money. *Quarterly Journal of Economics* 102: 703–726.

Christiano, L., M. Eichenbaum, and C. Evans. 2005. Nominal rigidity and the dynamics effects of a shock to monetary policy. *Journal of Political Economy* 113: 1–45.

Clarida, R., J. Gali, and M. Gertler. 1999. The science of monetary policy. *Journal of Economic Literature* 37: 1661–1707.

Cooper, R., and A. John. 1988. Coordinating coordination failures in Keynesian models. *Quarterly Journal of Economics* 103: 441–463.

Dixon, H.D. 1987. A simple model of imperfect competition with Walrasian features. *Oxford Economic Papers* 39: 134–160.

Dixon, H.D., and E. Kara. 2006. How to compare Taylor and Calvo contracts: A comment on Michael Kiley. *Journal of Money, Credit and Banking* 38: 1119–1126.

Dixon, H.D., and P. Lawler. 1996. Imperfect competition and the fiscal multiplier. *Scandinavian Journal of Economics* 98: 219–231.

Dixon, H.D., and N. Rankin. 1994. Imperfect competition and macroeconomics: A survey. *Oxford Economic Papers* 46: 171–199.

Gordon, R.J. 1990. What is new Keynesian economics. *Journal of Economic Literature* 28: 1115–1171.

Hart, O. 1982. A simple model of imperfect competition with Keynesian features. *Quarterly Journal of Economics* 97: 109–138.

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

Mankiw, N.G. 1985. Small menu costs and large business cycles: A macroeconomic model of monopoly. *Quarterly Journal of Economics* 100: 529–539.

Mankiw, N.G. 1988. Imperfect competition and the Keynesian cross. *Economics Letters* 26: 7–13.

Mankiw, N.G., and D. Romer. 1991. *New Keynesian economics*. Cambridge, MA: MIT Press.

Parkin, M. 1986. The output-inflation trade-off when prices are costly to change. *Journal of Political Economy* 94: 200–224.

Roberts, J.M. 1995. New Keynesian economics and the Phillips curve. *Journal of Money, Credit and Banking* 27: 975–984.

Shapiro, C., and J. Stiglitz. 1984. Equilibrium unemployment as a worker discipline device. *American Economic Review* 74: 433–444.

Sheshinski, E., and Y. Weiss. 1977. Inflation and costs of price adjustment. *Review of Economic Studies* 44: 287–303.

Silvestre, J. 1993. The market power foundations of macroeconomic policy. *Journal of Economic Literature* 31: 105–141.

Startz, R. 1989. Monopolistic competition as a foundation for Keynesian macroeconomic models. *Quarterly Journal of Economics* 104: 737–752.

Taylor, J. 1979. Staggered wage setting in a micro model. *American Economic Review* 69: 108–113.

Taylor, J. 1999. Staggered price and wage setting in macroeconomics. In *Handbook of macroeconomics*, vol. 1B, ed. J.B. Taylor and M. Woodford. Amsterdam: North-Holland.

Walsh, C.E. 2003. *Monetary theory and policy*, 2nd ed. Cambridge, MA: MIT Press.

Woodford, M. 2003. *Interest and prices*. Princeton: Princeton University Press.

New Open Economy Macroeconomics

Giancarlo Corsetti

Abstract

'New open economy macroeconomics' (NOEM) refers to a body of literature embracing a new theoretical framework for policy analysis in open economy, aiming to overcome the limitations of the Mundell–Fleming model while preserving the empirical wisdom and policy friendliness of traditional analysis. NOEM contributions have developed general equilibrium models with imperfect competition and nominal rigidities, to reconsider conventional views on the transmission of monetary and exchange rate shocks; they have contributed to the design of optimal stabilization policies, identifying international dimensions of optimal monetary policy; and they have raised issues about the desirability of international policy coordination.

Keywords

Disconnect puzzle; Elasticity of substitution; Exchange rate pass-through; Expenditure switching effects; Imperfect competition; Inflationary bias; International policy coordination; International portfolio diversification; International transmission mechanism; Jensen's inequality; Law of one price; Local currency pricing; Mundell–Fleming model; Neo-Wicksellian' monetary economics; New Keynesian macroeconomics; New neoclassical synthesis; New open economy macroeconomics; Nominal rigidities; Optimal monetary policy; Optimal tariffs; Output gap; Producer currency pricing; Stabilization; Staggered price setting; Sticky prices; Terms of trade; Trade costs; Uncertainty

JEL Classifications

F3; F4

The new open economy macroeconomics (NOEM) is a leading development in international economics that began in the early 1990s. Its objective is to provide a new theoretical framework for open economy analysis and policy design, overcoming the limitations of the Mundell–Fleming model, while preserving the empirical wisdom and the close connection to policy debates of the traditional literature. The new framework consists of choice-theoretic, general-equilibrium models featuring nominal rigidities and imperfect competition in the markets for goods or labour. In this respect, the NOEM has close links with related agendas pursued in closed-economy macro, such as the 'new neoclassical synthesis' and the 'neo-Wicksellian' monetary economics. The assumption of imperfect competition is logically consistent with the maintained hypothesis that firms and workers optimally choose prices and wages subject to nominal frictions, as well as with the idea that output is demand-determined over some range in which firms (workers) can meet demand at non-negative profits (surplus).

NOEM models differ from the Mundell–Fleming approach in at least two notable dimensions. First, all agents are optimizing, that is, households maximize expected utility and managers maximize firms' value. The expected utility of the national representative consumer thus provides a natural welfare criterion for policy evaluation and design. Second, general-equilibrium analysis paves the way towards further integration of international economics as a unified field, bridging the traditional gap between open macroeconomic and trade theory.

From a historical perspective, NOEM was launched by Obstfeld and Rogoff (1995), although Svensson and van Wijnbergen (1989) had also worked out a model with NOEM features as an open economy development of Blanchard and Kiyotaki (1987).

A specific goal of the NOEM agenda is to achieve the standards of tractability which made traditional models so popular and long-lived among academics and policymakers. For instance, many contributions have adopted the model specification by Corsetti and Pesenti (2001), which admits a closed-form solution by virtue of some educated restrictions on preferences (Tille 2001, explains the relation of this model to Obstfeld and Rogoff 1995). At the same time, the NOEM literature has promoted the construction of a new generation of large, multi-country quantitative models by international institutions and national monetary authorities. A leading example is the Global Economic Model (GEM) of the International Monetary Fund (see, for example, Laxton and Pesenti 2003).

This article first introduces a stylized NOEM model. Based on this model, it then provides a short selective survey of the NOEM literature, and its main advances in the analysis of the international transmission mechanism and policy design in open economies.

A Stylized NOEM Model

To illustrate the basic features of NOEM models, highlighting similarities and differences with the Mundell–Fleming model, it is useful to refer to the model by Corsetti and Pesenti (2001, 2005a, b) and Obstfeld and Rogoff (2000) (henceforth CP–OR). The economy consists of two countries, Home and Foreign, specialized in the production of one type of tradable goods, denoted H and F, respectively. Home consumption falls on both local goods and imports, that is, $C = C(C_H, C_F)$; the price level P includes both local goods and imports prices in Home currency, that is, $P = P(P_H, P_F)$. Preferences over local and imported goods are Cobb–Douglas with identical weights across countries: as the elasticity of substitution is equal to 1, any increase in domestic output is matched by a proportional fall in its price, so that terms-of-trade movements ensure efficient risk sharing. Furthermore, utility from consumption is assumed to be logarithmic, while disutility from labour l is linear.

Let μ index the Home monetary stance. Specifically, μ is the nominal value of the inverse of consumption marginal utility – for example, with log utility, $\mu = PC$. Whatever the instruments used by monetary authorities, μ indexes its ultimate effect on current spending. With competitive labour markets, the Households' optimality conditions imply that the nominal wage moves proportionally to μ, that is, $W = \mu$. Furthermore, abstracting from investment and government spending, μ indexes nominal aggregate demand. Similar definitions and conditions hold for the Foreign country, whose variables are denoted with a star, that is, $\mu^* = W^*$.

Let ε denote the nominal exchange rate, measured in units of Home currency per unit of Foreign currency. With perfect risk sharing, it is well known that the real exchange rate $\varepsilon P/P^*$ is equal to the ratio between the two countries' consumption marginal utilities (see Backus and Smith 1993). Rearranging this condition, the nominal exchange rate is equal to the ratio of Home to Foreign monetary stance, that is, $\varepsilon = \mu/\mu^*$. A Home expansion depreciates ε.

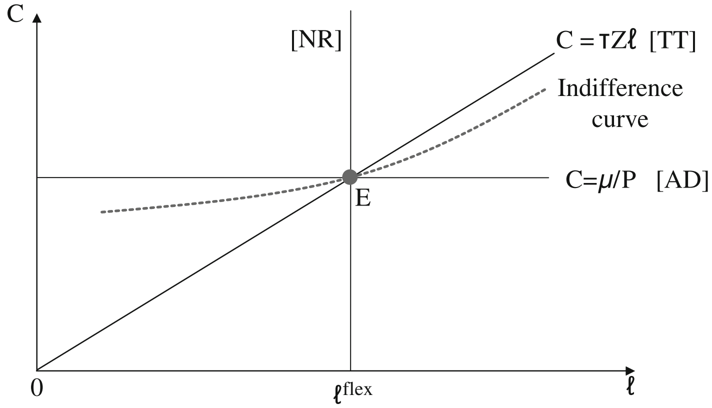
Goods are supplied by a continuum of firms, each being the only producer of a differentiated variety of the national good. For simplicity, production is linear in labour. With nominal rigidities, managers optimally set prices as to maximize the market value of the firm. (Since households are assumed to own firms, the discount factor used in calculating the present value is the growth in the marginal utility of consumption.) In the CP–OR model, prices are preset for one period and marginal costs coincide with unit labour costs $W/Z = \mu/Z$. In this model, optimal pricing actually takes a form that is very similar to textbook monopoly pricing: Home firms selling in the domestic market set P_H by charging the optimal markup over *expected* marginal costs, that is:

$$P_H = \text{markup} \cdot E \overset{\text{marginalcost}}{\left(\frac{\hat{\mu}}{Z}\right)}$$

where E denotes conditional expectations. If prices were flexible, the above would hold with current instead of expected costs.

When modelling nominal rigidities in the exports market, however, the following issue arises: are export prices sticky in the currency of the producers or in the currency of the destination market? In the NOEM literature, this issue has fed an extensive debate on the international transmission mechanism and the design of optimal stabilization policies, discussed in detail in the next sections.

The equilibrium allocation can be characterized in terms of three equilibrium relationships, labelled AD, TT and NR. In Fig. 1, these are drawn in the space 'consumption' vs. 'labour', C vs. l. The horizontal AD locus represents the Home aggregate demand in real terms, given by



New Open Economy Macroeconomics, Fig. 1

the ratio of the monetary stance to the price level: $C = \mu/P$. The upward-sloping TT locus shows the level of consumption that Home agents obtain (at market prices) in exchange for l units of labour. The slope of the TT locus depends on the (exogenous) productivity level Z, and the (endogenous) price of domestic GDP ($Y = Zl$), in terms of domestic consumption τ, that is, $C = \tau \cdot Z \cdot l$. Since agents consume both local goods and imports, τ rises with an improvement in the terms of trade of the Home country, conventionally defined as the price of imports in terms of exports. The vertical NR locus marks the equilibrium employment in the flexible prices (or natural rate) allocation, l^{flex}. Because of firms' monopoly power, l^{flex} is inefficiently low. To stress this point, Fig. 1 includes the indifference curve passing through the equilibrium point E, where it crosses the TT locus from above: with monopolistic distortions, the marginal rate of substitution between labour and consumption differs from the marginal rate of transformation.

With flexible prices, the macroeconomic equilibrium is determined by the NR locus and the TT locus. For a given μ, nominal prices adjustment ensures that demand is in equilibrium. With nominal rigidities, the equilibrium is instead determined by the AD locus and the TT locus. Depending on the level of demand, employment may fall short of or exceed the natural rate, opening employment and output gaps proportional to $(l^{flex} - l)$.

The International Transmission Mechanism and the Allocative Properties of the Exchange Rate

According to traditional open macroeconomic models, exchange rate movements play the stabilizing role of adjusting international relative prices in response to shocks, when frictions prevent or slow down price adjustment in the local currency. At the heart of this view is the idea that nominal depreciation transpires into real depreciation, making domestic goods cheaper in the world markets, hence redirecting world demand towards them: exchange rate movements therefore have 'expenditure switching effects'.

Consistent with this view, NOEM contributions after Obstfeld and Rogoff (1995) draws on the Mundell–Fleming and Keynesian tradition, and posits that export prices are sticky in the currency of the producers. Thus the nominal import prices in local currency move one-to-one with the exchange rate. This hypothesis is commonly dubbed 'producer currency pricing' (PCP).

Under PCP firms preset P_H and P_F^*, so the Home country's terms of trade $\varepsilon P_F^*/P_H$ deteriorate with unexpected depreciation. Moreover, as long as demand elasticities are identical in all markets, firms have no incentive to price discriminate: the price of exports obeys the law of one price, that is, $P_H^* = P_H/\varepsilon$ and $P_F = \varepsilon P_F^*$.

Monetary shocks have two distinct effects on the Home allocation and welfare. Expansions raise demand and output: because of monopolistic distortions in production, positive nominal shocks benefit domestic consumers by raising output towards its efficient (competitive) level. However, currency depreciation also raises the relative price of Foreign goods, reducing the real income of domestic consumers. In terms of Fig. 1, monetary expansions shift the AD locus upward and, due to currency depreciation, cause the TT locus to rotate clockwise. The new equilibrium may lie either above or below the indifference curve passing through E, the initial equilibrium. In other words, Home welfare may rise or fall, depending on the relative magnitude of monopoly power in production, vis-à-vis the terms-of-trade externality, in turn related to openness and the degree of substitutability between Home and Foreign tradables. (The size of the monetary shock also matters: by the same argument, by the theory of optimal tariffs a country never gains from monetary shocks which are large enough to raise output up to its competitive – Pareto-efficient – level.)

A noteworthy implication for policy analysis is that, in relatively open economies where terms-of-trade distortions are strong, benevolent policymakers may derive short-run benefits by implementing surprise monetary contractions, which appreciate the Home currency and boost the purchasing power of Home consumers. In these economies, monetary policy can have a deflationary bias.

In the Foreign country, welfare spillovers of a Home monetary expansion are unambiguously positive. Foreign consumers benefit from the terms-of-trade movement, which raises their income in real terms: the Foreign TT rotates counterclockwise. In addition, cheaper imports reduce inflation, raising aggregate demand for a given monetary stance μ^*: the Foreign AD shifts upward.

The high elasticity of import prices to the exchange rate underlying the above analysis is, however, at odds with a large body of empirical studies showing that the exchange rate pass-through on import prices is far from complete in the short run, and deviations from the law of one price are large and persistent (see, for example, Engel and Rogers 1996; Goldberg and Knetter 1997; Campa and Goldberg 2005). This evidence has motivated a thorough critique of the received wisdom on the expenditure switching effects of the exchange rate. Specifically, Betts and Devereux (2000) and Devereux and Engel (2003), among others, posit that firms preset prices in the currency of the markets where they sell their goods. This assumption, commonly dubbed 'local currency pricing' (LCP), attributes local currency price stability of imports mainly to nominal frictions, with far-reaching implications for the role of the exchange rate in the international transmission mechanism (see Engel 2003).

To the extent that import prices are sticky in the local currency, a Home depreciation does not affect the price of Home goods in the world markets; hence, it has no expenditure switching effects. Instead, it raises *ex post* markups on Home exports: at given marginal costs, revenues in domestic currency from selling goods abroad rise. In contrast with the received wisdom, nominal depreciation strengthens a country's terms of trade: if P_F and P_H^* are preset during the period, the Home terms of trade $P_F/\varepsilon P_H^*$ improve when the Home currency weakens. In Fig. 1, with LCP, a Home monetary expansion shifts aggregate demand AD upward and rotates the TT counterclockwise.

It follows that monetary authorities cannot derive short-run welfare benefits from surprise contraction. As currency depreciation improves the terms of trade, the inflationary bias in policymaking is even stronger than in a closed economy.

International spillovers from Home monetary expansions are detrimental to Foreign welfare. If prices in local currency remain constant, a Home expansion does not at all affect the aggregate demand in the Foreign country. Yet the adverse terms-of-trade movement forces Foreign agents to work more to sustain an unchanged level of consumption: for a given AD, the TT locus rotates clockwise.

An interesting case with asymmetric transmission is one in which the prices of exports are all preset in one currency, so that Home firms adopt PCP while Foreign firms adopt LCP (see, for example, Devereux et al. 2003).

While the NOEM literature has encompassed additional real and financial aspects in the analysis of the transmission mechanism, the PCP versus LCP debate identifies essential building blocks of optimal stabilization policy.

International Dimensions of Optimal Monetary Policies

A defining question of open economy macroeconomics is whether monetary and fiscal policy should react to international variables, such as the exchange rate or the terms of trade, beyond the influence that these variables have on the domestic output gap (for example, via external demand) and domestic inflation (for example, via import prices). This is a research area where choice-theoretic NOEM models have comparative advantages over the traditional literature. Indeed early NOEM contributions have established a set of original and provocative results, setting benchmarks for further analytical and quantitative studies.

To account for these results, consider the stabilization problem in a CP–OR economy with country-specific productivity uncertainty. In a flexible price environment (corresponding to the long run of the CP–OR model), a positive productivity shock in the Home country causes the world price of Home goods to fall. This raises both domestic and foreign demand for Home output, and worsens the Home terms of trade. With sticky prices, by contrast, unexpected gains in productivity simply translate into lower employment: given μ and μ^* (hence given the exchange rate), current demand is satisfied with a lower labour input. (In Fig. 1, a higher Z rotates the TT locus counterclockwise. With the AD and the TT loci held fixed, the equilibrium employment is below the natural rate. A fall in domestic prices would shift the AD locus up, while offsetting part of the rotation of the TT locus. The flexible price equilibrium always lies on the NR locus.)

However, under the hypothesis of PCP, it is easy to see that monetary policy in a sticky-price environment can support the flexible price allocation. Posit that monetary rules satisfy $\mu = \Gamma Z$, where Γ denotes a (possibly time-varying) variable indexing the level of nominal variables in the Home country. When such rules are implemented, any gain in productivity is matched by a proportional expansion of the monetary stance, which raises Home demand and depreciates the Home currency. Marginal costs remain constant in nominal terms (since $\mu/Z = \Gamma$): hence product prices in domestic currency would remain fixed even if there were no nominal rigidities. At the same time, however, exchange rate movements adjust international relative prices, as monetary policy moves ε in proportion to productivity changes.

A first benchmark result is that, in economies with the CP–OR features, monetary policy rules supporting the flexible price allocation are optimal: no rule welfare-dominates complete marginal cost and output gap stabilization. This is true under different assumptions regarding nominal rigidities, including staggered price setting and partial adjustment (see, for example, Clarida et al. 2002). Optimal monetary rules are completely 'inward-looking': welfare-maximizing central banks stabilize the GDP deflator while letting the consumer price index (CPI) fluctuate with movements in the relative price of imports. There is no need for monetary policies to react to international variables.

The result that monetary rules supporting a flexible price allocation are optimal, however, does not hold in general. In the presence of multiple distortions monetary authorities are generally able to exploit nominal rigidities and improve welfare relative to such allocation (Benigno and Benigno 2003; Corsetti and Dedola 2005). Yet, holding PCP, it is unclear whether and under which conditions deviating from full domestic stabilization could yield significant welfare gains.

A second result concerns the costs of inefficient stabilization. The New Keynesian theory has emphasized welfare costs from relative price dispersion when private pricing decisions are not synchronized (see, for example, Galí and Monacelli 2003). Early NOEM contributions have instead pioneered the analysis of the effect of uncertainty on the level of prices and economic activity. A simple example illustrates this point. Suppose that monetary policy responds to productivity shocks according to rule: $\mu = \Gamma Z^{\gamma}$. When $\gamma < 1$, marginal cost uncertainty due to insufficient stabilization implies $E(\mu/Z) = \Gamma E(1/Z^{1-\gamma}) > \Gamma$: by a straightforward application of Jensen's inequality, expected marginal costs are higher than under complete stabilization. Higher costs transpire into higher prices both in nominal terms and relative to wages, reducing the average supply of domestic goods, thus exacerbating monopolistic distortions in the economy (see, for example, Sutherland 2005, and Kollman 2002, for a quantitative assessment).

Similar effects, with potentially stronger welfare implications, are caused by a noisy conduct of monetary policy and exchange rate variability (Obstfeld and Rogoff 1998). Notably, Broda (2006) provides evidence consistent with the (NOEM) prediction that incomplete stabilization *and* monetary/exchange rate noise transpire into higher price levels and real appreciation.

A third result, derived on the assumption of LCP, defines a clear-cut argument in favour of policies with an international dimension. To the extent that exporters' revenues and markups are exposed to exchange rate uncertainty, firms' optimal pricing strategies internalize the monetary policy of the importing country. In the CP–OR model, for instance, Foreign firms optimally preset the price of their goods in the Home market P_F by charging the equilibrium markup over expected marginal costs *evaluated in Home currency*, that is,

$$P_F = \text{markupg} \cdot E\left(\varepsilon \frac{\mu^*}{Z^*}\right) = \text{markupg} \cdot E\left(\frac{\mu}{Z^*}\right).$$

Clearly, the price of Home imports depends on the joint distribution of Home monetary policy and Foreign productivity shocks.

Suppose that Home monetary authorities ignore the influence of their decisions on the price of Home imports. For the reason discussed above, import prices will tend to be inefficiently high. On the other hand, if Home monetary authorities want to stabilize Foreign firms' marginal costs, they can only do so at the cost of raising costs and markup uncertainty for Home producers, resulting in higher Home good prices. It follows that, to maximize Home welfare, Home policymakers should optimally trade off the stabilization of marginal costs of all producers (domestic and foreign) selling in the Home markets.

When foreign firms' profits are exposed to exchange rate uncertainty, optimal monetary rules are no longer inward-looking. The importance of Foreign shocks in the conduct of monetary policy depends on the degree of openness of the economy, measured by the overall share of imports in the CPI (see Corsetti and Pesenti 2005a, and Sutherland 2005, for a discussion of intermediate degrees of pass-through, and Smets and Wouters 2002, and Monacelli 2005, for models with staggered price setting).

Notably, the case for an international dimension in monetary policy described above transpires into limited exchange rate variability. Since with LCP optimal monetary policies respond to both domestic and foreign shocks, national monetary stances tend to be more correlated than in the case of inward-looking stabilization of output gaps. This implies lower exchange rate volatility. In the baseline CP–OR model, the optimal policy rules actually prevent *any* short-run fluctuations of the exchange rate, a point stressed by Devereux and Engel (2003). But this exact result holds only when the weights of Home and Foreign goods in final expenditure are assumed to be identical across countries: Home and Foreign monetary authorities de facto stabilize the same weighted average of marginal costs. The presence of non-traded goods or some Home bias in consumption would obviously imply asymmetries in the optimal monetary stances, which would be incompatible with a fixed exchange rate (Duarte and Obstfeld 2007; Corsetti 2006). Even if, with LCP, exchange rate variability does not perform any role in adjusting international prices, a fixed rate regime would impose unwarranted constraints on the efficient conduct of monetary policy.

A fourth result concerns the desirability of international policy coordination. Leading NOEM contributions have fed considerable scepticism on this issue. At the core of this scepticism is the disappointing quantitative assessment of welfare gains from coordination. By using the CP–OR model, for instance, it is possible to build economies with either PCP or LCP behaviour, where optimal monetary rules are identical whether national policymakers act independently or cooperatively (maximizing an equally weighted sum of national welfare functions). When this exact result breaks down (depending on the elasticity of substitution between Home and Foreign tradables, and/or sector-specific shocks in the presence of non-tradables), gains from coordination usually remain quite small (see, for example, Pappa 2004; Benigno and Benigno 2006).

The lesson from the NOEM literature, stressed by Obstfeld and Rogoff (2002), is a new welfare-based argument against coordination: once policymakers independently pursue efficient stabilization policies in their own country (that is, they 'keep their house in order'), the room for improving welfare through cooperation is quite limited (see Canzoneri et al. 2005, for a discussion).

The results reviewed above were first derived in highly stylized economies. A critical question directing current NOEM research is whether they would still hold in richer models with good quantitative performance.

Challenges to the NOEM Literature

The above debate on the role of exchange rate in the international transmission has motivated further empirical and theoretical work on market segmentation along national borders and on its implications for international macroeconomic adjustment. As stressed by Obstfeld and Rogoff (2001), despite the ongoing process of real and financial globalization, frictions and imperfections appear to keep national economies 'insular'.

An important issue is the extent to which the evidence of local currency price stability of imports can be explained by nominal rigidities. It is well understood that the low elasticity of import prices with respect to the exchange rate is in large part due to the incidence of distribution (Burstein et al. 2007). Several macro and micro contributions have emphasized the role of optimal destination-specific markup adjustment by monopolistic firms depending on market structure (Dornbusch 1987; Goldberg and Verboven 2001), or vertical interactions between producers and retailers (Corsetti and Dedola 2005).

The main point is that low pass-through is not necessarily incompatible with expenditure switching effects (see, for example, Obstfeld 2002). In this respect, Obstfeld and Rogoff (2000) emphasizes that, in the data (and consistent with the received wisdom), nominal depreciation does tend to be associated with deteriorating terms of trade. This piece of evidence clearly sets an empirical hurdle for LCP models, if we assume a high degree of price stickiness in local currency (see Corsetti et al. 2005, for a quantitative assessment). Interestingly, estimates of LCP models downplaying price discrimination, distribution and other real determinants of incomplete pass-through predict that the degree of price stickiness is implausibly higher for imports than for domestic goods, a result suggesting model misspecification (see, for example, Lubik and Schorfheide 2006).

Moreover, the currency denomination of exports prices should be treated as an endogenous choice by profit maximizing firms (see, for example, Bacchetta and Van Wincoop 2005; Devereux et al. 2004). To appreciate the contribution by the NOEM literature on this issue, recall that, in the CP–OR model above, expansionary monetary shocks unrelated to productivity raise nominal wages and marginal costs while depreciating the currency. For a firm located in a country with noisy monetary policy, pricing its exports in foreign currency (that is, choosing LCP) is therefore quite attractive: it ensures that revenues from exports in domestic currency will tend to rise in parallel with nominal marginal costs, with stabilizing effects on the markup. This may help explain why exporters from emerging markets with relatively unstable domestic monetary policies prefer to price their exports to advanced countries in the importers' currency. The same argument, however, suggests that LCP is not necessarily optimal for exporters producing in countries where monetary policy systematically stabilizes marginal costs (see Goldberg and Tille 2005, for empirical evidence).

New waves of studies are building models with trade costs where goods tradability is endogenous, and/or new varieties are created at business cycle frequencies. Trade and transaction costs are also at the heart of recent attempts to integrate current account and macroeconomic dynamics with international portfolio diversification in a unified analytical framework (see, for example, Devereux and Sutherland 2007).

The discussion above is far from exhausting the range of topics and issues analysed by the NOEM literature, which has marked a radical change of paradigm in international macroeconomics. Many authors have undertaken a systematic reconsideration of classical themes in the new framework. A partial list of themes includes overshooting (for example, Hau 2000); current account, debt and exchange rate dynamics (Cavallo and Ghironi 2002; Ganelli 2005; Ghironi 2006); exchange rate uncertainty and trade (Bacchetta and Van Wincoop 2000); and fiscal policy (Adao et al. 2006). An important set of papers delves into empirical analysis of NOEM models (for example, Bergin 2003; Lubik and Schorfheide 2006).

Yet most NOEM contributions so far specify models which predict a counterfactually high degree of consumption risk sharing: even when financial markets are incomplete, intertemporal trade and terms-of-trade spillovers ensure that the consumption risk of productivity shocks is contained, and the market allocation is not too distant from the efficient one (see, for example, Chari et al. 2002). Not only this is inconsistent with a large body of evidence (see Backus and Smith 1993); most crucially, a counterfactually high degree of risk sharing built in NOEM models may limit their capacity to comprehend significant cross-border spillovers and policy trade-offs. Similarly, in most models the exchange rate is tightly related to fundamentals, at odds with a large body of evidence showing that the relation between the exchange rate and virtually any macroeconomic aggregate is exceedingly weak – the so-called disconnect puzzle.

Further progress in these areas is crucial towards the fulfilment of the NOEM research agenda.

See Also

- ▶ International Finance
- ▶ International Policy Coordination
- ▶ International Real Business Cycles
- ▶ Nominal Exchange Rates
- ▶ Price Discrimination (Empirical Studies)

Bibliography

Adao, B., M.I. Horta Correia, and P. Teles. 2006. On the relevance of exchange rate regimes for stabilization policy. Discussion paper no. 5797. CEPR.

Bacchetta, P., and E. Van Wincoop. 2000. Does exchange rate stability increase trade and welfare? *American Economic Review* 50: 1039–1109.

Bacchetta, P., and E. Van Wincoop. 2005. A theory of the currency denomination of international trade. *Journal of International Economics* 67: 295–319.

Backus, D.K., and G.W. Smith. 1993. Consumption and real exchange rates in dynamic economies with non–-traded goods. *Journal of International Economics* 35: 297–316.

Benigno, G., and P. Benigno. 2003. Price stability in open economy. *Review of Economic Studies* 70: 743–764.

Benigno, G., and P. Benigno. 2006. Designing targeting rules for international monetary policy cooperation. *Journal of Monetary Economics* 53: 473–506.

Bergin, P. 2003. Putting the new open economy macroeconomics to a test. *Journal of International Economics* 60: 3–34.

Betts, C., and M. Devereux. 2000. Exchange rate dynamics in a model of pricing to market. *Journal of International Economics* 50: 215–244.

Blanchard, O., and N. Kiyotaki. 1987. Monopolistic competition and the effects of aggregate demand. *American Economic Review* 77: 647–666.

Broda, C. 2006. Exchange rate regimes and national price levels. *Journal of International Economics* 70: 52–81.

Burstein, A., M. Eichenbaum, and S. Rebelo. 2007. Modeling exchange rate pass-through after large devaluations. *Journal of Monetary Economics* 54: 346–368.

Campa, J., and L. Goldberg. 2005. Exchange rate pass through into import prices. *Review of Economics and Statistics* 87: 679–690.

Canzoneri, M.B., R. Cumby, and B. Diba. 2005. The need for international policy coordination: What's old, what's new, what's yet to come? *Journal of International Economics* 66: 363–384.

Cavallo, M., and F. Ghironi. 2002. Net foreign assets and the exchange rates: Redux revived. *Journal of Monetary Economics* 49: 1057–1097.

Chari, V.V., P.J. Kehoe, and E. McGrattan. 2002. Can sticky prices generate volatile and persistent real exchange rates? *Review of Economic Studies* 69: 633–663.

Clarida, R., J. Galí, and M. Gertler. 2002. A simple framework for international policy analysis. *Journal of Monetary Economics* 49: 879–904.

Corsetti, G. 2006. Openness and the case for flexible exchange rates. *Research in Economics* 60: 1–21.

Corsetti, G., and L. Dedola. 2005. A macroeconomic model of international price discrimination. *Journal of International Economics* 67: 129–156.

Corsetti, G., and P. Pesenti. 2001. Welfare and macroeconomic interdependence. *Quarterly Journal of Economics* 116: 421–446.

Corsetti, G., and P. Pesenti. 2005a. International dimension of optimal monetary policy. *Journal of Monetary Economics* 52: 281–305.

Corsetti, G., and P. Pesenti. 2005b. The simple geometry of transmission and stabilization in closed and open economies. Working paper no. 11341. Cambridge, MA: NBER.

Corsetti, G., L. Dedola, and S. Leduc. 2005. DSGE models of high exchange rate volatility and low pass through. Discussion paper no. 5377. CEPR.

Devereux, M., and C. Engel. 2003. Monetary policy in open economy revisited: Price setting and exchange rate flexibility. *Review of Economic Studies* 70: 765–783.

Devereux, M., and A. Sutherland. 2007. Monetary policy and portfolio choice in an open economy macro model. *Journal of the European Economics Association* 5: 491–499.

Devereux, M., C. Engel, and C. Tille. 2003. Exchange rate pass-through and the welfare effects of the euro. *International Economic Review* 44: 223–242.

Devereux, M., C. Engel, and P. Storgaard. 2004. Endogenous exchange rate pass-through when nominal prices are set in advance. *Journal of International Economics* 63: 263–291.

Dornbusch, R. 1987. Exchange rates and prices. *American Economic Review* 77: 93–106.

Duarte, M., and M. Obstfeld. 2007. Monetary policy in the open economy revisited: The case for exchange-rate flexibility restored. Working paper. Online. Available at http://elsa.berkeley.edu/~obstfeld/DO_JIMF-2.pdf. Accessed 29 Nov 2007.

Engel, C. 2003. Expenditure switching and exchange rate policy. In *NBER macroeconomics annual 2002*, vol. 17, 231–272.

Engel, C., and J. Rogers. 1996. How wide is the border? *American Economic Review* 86: 1112–1125.

Galí, J., and T. Monacelli. 2003. Monetary policy and exchange rate volatility in a small open economy. *Review of Economic Studies* 72: 707–734.

Ganelli, G. 2005. The new open economy macroeconomics of government debt. *Journal of International Economics* 65: 167–184.

Ghironi, F. 2006. Macroeconomic interdependence under incomplete markets. *Journal of International Economics* 76: 428–450.

Goldberg, P.K., and M.M. Knetter. 1997. Goods prices and exchange rates: What have we learned? *Journal of Economic Literature* 35: 1243–1272.

Goldberg, L., and C. Tille. 2005. Vehicle currency use in international trade. Staff report no. 200. Federal Reserve Bank of New York.

Goldberg, P.K., and F. Verboven. 2001. The evolution of price dispersion in the European car market. *Review of Economic Studies* 68: 811–848.

Hau, H. 2000. Exchange rate determination: The role of factor price rigidities and nontradeables. *Journal of International Economics* 50: 421–447.

Kollman, R. 2002. Monetary policy rules in the open economy: Effects on welfare and business cycles. *Journal of Monetary Economics* 49: 989–1015.

Laxton, D., and P. Pesenti. 2003. Monetary rules for small, open, emerging economies. *Journal of Monetary Economics* 50: 1109–1146.

Lubik, T., and F. Schorfheide. 2006. A Bayesian look at new open economy macroeconomics. In *NBER macroeconomics annual 2005*, vol. 20, 313–366.

Monacelli, T. 2005. Monetary policy in a low pass-through environment. *Journal of Money Credit and Banking* 6: 1047–1066.

Obstfeld, M. 2002. Inflation-targeting, exchange rate pass-through, and volatility. *American Economic Review* 92: 102–107.

Obstfeld, M., and K. Rogoff. 1995. Exchange rate dynamics redux. *Journal of Political Economics* 102: 624–660.

Obstfeld, M., and K. Rogoff. 1998. Risk and exchange rates. In *Contemporary economic policy: Essays in honor of Assaf Razin*, ed. H. Helpman and E. Sadka. Cambridge/New York: Cambridge University Press.

Obstfeld, M., and K. Rogoff. 2000. New directions for stochastic open economy models. *Journal of International Economics* 50: 117–153.

Obstfeld, M., and K. Rogoff. 2001. The six major puzzles in international finance: Is there a common cause? In *NBER macroeconomics annual 2000*, vol. 15, 339–390.

Obstfeld, M., and K. Rogoff. 2002. Global implications of self-oriented national monetary rules. *Quarterly Journal of Economics* 117: 503–536.

Pappa, E. 2004. Do the ECB and the Fed really need to cooperate? Optimal monetary policy in a two-country world. *Journal of Monetary Economics* 51: 753–779.

Smets, F., and R. Wouters. 2002. Openness, imperfect exchange rate pass-through and monetary policy. *Journal of Monetary Economics* 49: 947–981.

Sutherland, A. 2005. Incomplete pass-through and the welfare effects of exchange rate variability. *Journal of International Economics* 65: 375–400.

Svensson, L.E.O., and S. van Wijnbergen. 1989. Excess capacity, monopolistic competition and international transmission of monetary disturbances. *Economic Journal* 99: 785–805.

Tille, C. 2001. The role of consumption substitutability in the international transmission of shocks. *Journal of International Economics* 53: 421–444.

Newcomb, Simon (1835–1909)

Milton Friedman

Keywords
Charitable giving; Equation of exchange; Newcomb, S.; Quantity theory of money; Wages fund

JEL Classifications
B31

Newcomb entitled his autobiography *Reminiscences of an Astronomer* (1903), devoted only 10 pages out of 416 to his activities in economics, and remarked: ‘Being sometimes looked upon as an economist, I deem it not improper to disclaim any part in the economic research of today’ (p. 408). The 1913 *Encyclopaedia Britannica* in a lengthy article describes him as ‘one of the most distinguished astronomers of his time’ and includes one sentence, ‘He also wrote on questions of finance and economics.’ The 1970 edition of the *Encyclopaedia* describes him as ‘the greatest American astronomer of the 19th century’ and repeats the remark that ‘he wrote on finance and economics’.

Those may well be correct evaluations of the *relative* importance of Newcomb’s work in astronomy and economics. Yet they give a wholly misleading impression of the *absolute* importance of his contribution to economics. He wrote two classics of economic science: *A Critical Examination of Our Financial Policy during the Southern Rebellion* (1865) and *Principles of Political Economy* (1885). The first ‘contains the most sophisticated, original, and profound analysis of the theoretical issues involved in Civil War finance that we have encountered, regardless of date of publication’ (Friedman and Schwartz 1963, p. 18). The second contains what Irving Fisher, in his obituary note on Newcomb, regarded as ‘his chief and most fruitful contribution to economic science’, namely.

> the distinction he applied in particular to what he called ‘societary circulation’, or the equation of exchange between money and goods. So far as I am aware, he was the first definitely to enunciate this equation, expressing the fact that the quantity of money multiplied by its velocity of circulation is equal to price-level multiplied by volume of business transactions. This equation, with due amplifications, represents the so-called ‘quantity theory of money’ in its highest form. He also employed this same distinction
>
> . . . to expose the fallacy of ‘the wage-fund’. (Fisher 1909, p. 642)

Another notable item in the *Principles* is its final chapter, ‘Of Charitable Effort’, an economic analysis of charity that is highly relevant to modern problems of the welfare state, in part because Newcomb writes about currently sensitive issues with a frankness and plainness that is absent from contemporary literature.

In addition to these two books, Newcomb published well over 50 popular magazine articles on economic issues, some of which formed the basis for two popular books: *The ABC of Finance* (1877) and *A Plain* Man’s *Talk on the Labor Question* (1886). The latter, which was still in print in the 1980s, remains today an extraordinarily persuasive and effective exposition of the basic principles of a market economy and the effects of labour union activity on the interests of the worker.

Had these items constituted the whole of Newcomb’s canon, instead of only a surprising few out of a total of well over 500 items, including not only major works in astronomy but also textbooks in mathematics, important contributions to statistics, and even a science fiction novel, he would have come to be regarded as one of the leading American economists of the 19th century. Irving Fisher noted that one reason ‘his economic writings did not attract the attention among economists which they deserved . . . is that . . . once a man’s name becomes associated with a particular department of knowledge like astronomy, any attempts to contribute to other departments encounter a prejudice which it is difficult to overcome’ (Fisher 1909, p. 641). Perhaps also it was not irrelevant that he was completely self-taught in economics, as in much else.

Newcomb was born on 12 March 1835 in a small town in Nova Scotia, son of an impecunious country school teacher. He died on 11 July 1909, and was buried with all the military pomp due to his congressionally conferred rank of Rear Admiral, a remarkable transition due entirely to Newcomb's own talents, character and persistence. Though his only formal schooling consisted of occasional attendance at his father's schools, he early displayed unusual intellectual interests and capacities. As what seemed in that remote region and time the only avenue of further instruction, he was apprenticed at the age of 16 to a herb doctor for a five-year period. The doctor turned out to be a quack who treated Simon as a slave, and provided no training whatsoever.

After two years, Simon finally summoned up the courage to run away, hiding in the woods as his erstwhile master sought to track him down. He joined his father, who had gone to New England after the death of Simon's mother, and father and son made their way to the Eastern Shore of Maryland, where both found employment as country teachers. Despite being entirely self-taught, Simon started to write articles on mathematical and astronomical subjects, one of which he sent to Professor Henry, Secretary of the Smithsonian Institute. This led to Professor Henry's becoming interested in Newcomb and ultimately recommending him for a job as a 'computer' at the Nautical Almanac in Cambridge, Massachusetts – an event which Newcomb described as 'an epoch – an entrance into a new world'. Employment at the Nautical Almanac enabled him to take courses at the Lowell Scientific School of Harvard University, where he received a degree in 1857. In 1861 he received a commission as professor of mathematics at the US Naval Observatory; in 1877 he was appointed superintendent of the Nautical Almanac, and in 1884 professor of mathematics and astronomy at the Johns Hopkins University, a position he held concurrently with his posts at the Naval Observatory and the Nautical Almanac.

In addition to his prodigious written output, Newcomb served for many years as editor of the *American Journal of Mathematics* and was active in the National Academy of Sciences, and the American Association for the Advancement of Science, of which he was president in 1877. Truly a Renaissance man.

See Also

- ▶ Equation of Exchange
- ▶ Quantity Theory of Money

Bibliography

Archibald, R.C. 1924. Simon Newcomb 1835–1909. Bibliography of his life and work. *Memoirs of the National Academy of Sciences* 17: 19–69.

Campbell, W.W. 1924. Biographical memoir: Simon Newcomb. *Memoirs of the National Academy of Sciences* 17: 1–18.

Fisher, I. 1909. Obituary. Simon Newcomb. *Economic Journal* 19: 641–644.

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

Stigler, S.M. 1973. Simon Newcomb, Percy Daniell, and the history of robust estimation 1885–1920. *Journal of the American Statistical Association* 68: 872–879.

Newmarch, William (1820–1882)

D. P. O'Brien

Newmarch was born in Thirsk, Yorkshire, and died in Torquay. He had little formal education, but rose from the position of bank clerk to be a force in the City of London, being manager of Glyn Mills from 1862 to 1881. An excitable but effective speaker, he was a member of the Political Economy Club from 1852 (Treasurer 1855–1882), and a considerable force in the (Royal) Statistical Society of which he was Secretary 1854–1862 and President 1869–1871. He was also elected a Fellow of the Royal Society and wrote for the *Morning Chronicle* and the *Economist*.

Newmarch was important as the principal author of the last two volumes of Tooke's monumental *History of Prices* (though, oddly, this publication, unlike Newmarch's own work in the *Economist*, did not employ index numbers), and as an economist in his own right for exploring the effects of the gold discoveries, public debt, and questions of monetary control. He was one of the leading opponents of the Currency School and the Bank Act of 1844, arguing that causality ran from prices to note issue, so long as the notes were convertible. He believed that monetary base control, as embodied in the 1844 Act, was not only ineffective – following the work of William Leatham he showed that the actual number of bills of exchange increased in times of monetary contraction – but that it produced, at times, both unnecessary stringency and harmful fluctuations in the rate of interest. Though his position was analytically underdeveloped his work is still of considerable interest.

See Also

▶ Tooke, Thomas (1774–1858)

Selected Works

1851. An attempt to ascertain the magnitude and fluctuations of the amount of bills of exchange (Inland and Foreign) in circulation at one time in Great Britain, in England, in Scotland, in Lancashire, and in Cheshire, respectively, during each of the twenty years 1828–1847, both inclusive; and also embracing in the inquiry bills drawn upon foreign countries. *Journal of the Royal Statistical Society* 14: 143–183.

1857. (With T. Tooke) *A history of prices and of the state of the circulation, during the nine years 1848–1856. In two volumes; forming the fifth and sixth volumes of the history of prices from 1792 to the present time*. London: Longmans.

Bibliography

Ashbee, R.A. 1979. William Newmarch and Glyns. *Three Banks Review* 122: 49–60.

Bonar, J., and H.W. Macrosty. 1934. *Annals of the royal statistical society 1834–1934*. London: Royal Statistical Society.

Gregory, T.E. 1928. *An introduction to Tooke and Newmarch's 'A history of prices and of the state of the circulation'*. London: P.S. King.

O'Brien, D. P., ed. 1971. *The correspondence of Lord Overstone*, 3 vols. Cambridge: Cambridge University Press.

Wood, E. 1939. *English theories of central banking control 1819–1858*. Cambridge, MA: Harvard University Press.

News Shocks

Nir Jaimovich
University of Zurich, Department of Economics,
Zurich, Switzerland

Abstract

News shocks are shocks that are useful for predicting future fundamentals but do not affect current fundamentals. While the idea of "news shocks" as a driver of economic fluctuations has been present since the early work on business cycles it had been formalized and assessed in the last decade. This entry discusses both the theoretical impact of news shocks on the economy and their empirical relevance for business cycles.

Keywords

Business cycles; Shocks; Information

JEL Classification

E13; E20; E32

Introduction

What are the forces that lead the economy to experience booms and busts in aggregate economic activity? This question has been central within (i) the economic profession, (ii) policy makers and (iii) the general public.

The modern approach to business cycle analysis relies on the methodological breakthroughs in the 1980s of the real business cycle (hereafter RBC) framework in particular, and the dynamic stochastic general equilibrium approach (hereafter DSGE). Central to this framework is that it studies the effect of various shocks to the economy (such as monetary, fiscal, trade, oil and “animal-spirits” shocks).

Throughout many iterations of the DSGE framework, it has been argued that a key shock in generating business cycle is a “technology/total-factor-productivity (TFP) shocks” – i.e., shocks that directly affect the production function. The fact that a key “suspect” in generating business cycle is a shock that directly affects the production function has proven to be controversial for various reasons (see the discussion in Rebelo (2005)). Two key criticisms have been as follows. First, it is hard to accept the idea that recessions are driven by negative TFP shocks as this would imply that the economy simply “forgot” how to produce. Second, in the DSGE framework, assuming that the economy has some advance knowledge on new technologies yields predictions that, ex ante, seem counterintuitive. For example, “positive news” about the arrival of new technologies sends, immediately as the news arrive, the economy into a recession!

These shortcomings lead researchers to consider a new class of shocks in the last 10 years, “news shocks”. These are shocks that are useful for predicting future fundamentals but do not affect current fundamentals. While the idea of “news shocks” has been present since the early work on business cycles (e.g., Pigou (1927)), it was dormant until the work of Beaudry and Portier (2004, 2006). Specifically, Beaudry and Portier (2004) proposed a modern DSGE framework where news shocks can yield (i) recessions without having to rely on negative TFP shocks, and where (ii) positive (negative) news about the future can lead to an expansion (recession) in the current period. In addition to this theoretical work, Beaudry-Portier (2006) studied the relevance of news shocks from an empirical point of view. They developed an empirical framework according to which “news shocks” were found to have the effects as in Beaudry-Portier (2004) and showed the quantitative importance of news shocks.

In the decade that followed these seminal contributions, there has been a burst of theoretical and empirical research studying the effects of news shocks. On the theory side, research aimed at exploring the theoretical conditions under which news shocks can be a key shock that drives the economy. On the empirical front, the key identification problem is that, naturally, news shocks are not directly observable. This has lead to different empirical specifications and approaches to study their effect. Currently, the empirical evidence on the plausibility and relevance of news shocks is still mixed.

The rest of this entry proceeds as follows. For simplicity, we consider throughout the entry only the reaction of the economy to positive news about the future. In almost all models the response to a negative shock is simply the opposite of the response to a good news. The next section sketches a simple model that analyses the impact of news shocks and describes the building blocks of more advanced and recent work in the literature. Section “Modern Approach” then moves to the more sophisticated current work and especially discusses the current empirical approaches aimed at identifying the impact of news. Section “Conclusions” concludes.

A Simple Model of News Shocks

While the basic premise that good news about the future can generate an expansion sounds intuitive, it is not present in the most basic modern macro models. Specifically, in this line of models, good news about the future leads in fact to a fall in employment and output! We begin this section by describing the basic intuition of why positive news about the future could lead to a decline in economic activity. We then formalize this example in a simple consumer choice problem. This will serve as benchmark for the discussion of how modern macroeconomic models overcame this prediction.

Specifically, consider a consumer who derives utility from consuming a product (say bananas) and leisure (say watching TV). It is common in Economics to assume that these are both normal goods; that is, holding everything else constant, the richer the consumer is, the more bananas and leisure she wants to consume.

Consider the case that suddenly the consumer faces a temporary increase in her current hourly wage. How would she react to this? On the one hand, this temporary increase in her hourly wage makes taking time off for leisure more costly; for example, instead of watching 1 h of TV she could be working an extra hour and take advantage of the temporary higher wage rate. In Economics, this effect is termed as the "substitution effect" where consumers shy away from a good (in this case leisure), if its price increases (in this case the wage rate). On the other hand, since her current hourly wage is higher, and thus, holding everything else constant, she is richer, the consumer would like to consume more of the things she enjoys: i.e., more bananas and more TV. In Economics, this effect is termed as the "income effect" where consumers consume more of the goods they care about as they get richer.

Overall, in this example, whether the consumer will end up working more or less depends on different assumptions. However, practically, in almost all modern macroeconomics, the substitution effect (i.e., the "working more") tends to dominate, and hence, the consumer would end up working more, taking advantage of the current temporary increase in the hourly rate.

Consider now the case when this consumer suddenly learns that her future, rather than the contemporaneous, hourly wage is about to increase. What will she do? The consumer understands that her lifetime resources have increased, and hence she is richer. This implies that she would like to consume more of all the normal goods she cares about. Hence she will consume immediately more bananas and more TV watching, even though she did not receive the increase in income at the current period.

Since in this example there is no immediate increase in her current salary, there is no offsetting substitution effect that makes her work more. Thus, in response to "good news" about the future, the consumer ends up eating more bananas and spending more time watching TV, implying that she will work less and employment falls. Since employment is an input in the production function, then a fall in employment leads to a fall in output and since consumption increases, then it must be that savings (and thus investment) falls. Hence, overall good news about the future will lead to an immediate contraction in output and employment!

A Two Period Example

In what follows we formalize this intuition in a simple two period model. Specifically, consider the above consumer to maximize her utility from consumption over two periods (we will later add her utility from leisure to the analysis). That is, the consumer cares about consumption today (which we denote by a utility function $U(c_t)$) and consumption tomorrow (which we denote by a utility function $U(c_{t+1})$).[1] We assume that the consumer likes to consume more (i.e., the first derivative of the utility function is positive), but at a declining rate (i.e., the second derivative of the utility function is negative).[2] Naturally, absent a budget constraint the consumer would like to consume infinite amounts. Thus, the consumer needs to be facing a budget constraint. Specifically, at the first period (i.e., period t) the consumer's budget constraint is given by

$$c_t + a_{t+1} = y_t,$$

where y_t denotes her income at that period and where a_{t+1} denotes any savings she transfers from the first per

iod to the second period (i.e., $t + 1$).[3] Then, in the second period, the consumer's budget constraint is given by

[1] For simplicity, without loss of generality, we assume no discounting, and a gross interest that equals one.

[2] That is, U is a strictly concave function.

[3] Without loss of generality we assume that the consumer begins the period with no assets.

$$c_{t+1} = y_{t+1} + a_{t+1}.$$

That is, the consumer's resources are her income (y_{t+1}) and the savings she transferred from the first period.[4] We can combine these two budget constraints into one "lifetime" budget constraint:

$$c_t + c_{t+1} = y_t + y_{t+1}.$$

This last equation simply reflects the fact that over her lifetime, the consumer's total consumption must equal her total lifetime income.

What is the optimal consumption path of the consumer? Maximizing the consumer's utility with respect to consumption today and consumption tomorrow, and denoting by "prime" sign the first derivative of the utility function, it follows that she will equate the marginal utility of consumption in both periods, that is,

$$U'(c_t) = U'(c_{t+1})$$

Moreover, given the assumption that U is a strictly concave function this simply implies that

$$c_t = c_{t+1} = c^*.$$

That is, the optimal consumption path is to consume the same amounts of bananas in each period, which we denote by c^*. Using this result in the budget constraint, we thus get

$$c^* = \frac{y_t + y_{t+1}}{2},$$

that is, the consumer splits her lifetime income by two and consumes this amount at each period.

Consider now the case, as in the above discussion, where the consumer learns a period in advance that her next period income, i.e., y_{t+1}, will increase with certainty. Then, as the equation above suggests, contemporaneous consumption, i.e., c_t, will increase immediately (as the consumer wants to spreads her lifetime income over the two periods). However, since her current income (i.e., y_t) did not increase, then it must be that her current savings (i.e., a_{t+1}, which also equal to investment in this example) will fall. Thus, this simple example captures the above intuition; in the presence of goods news about future income, contemporaneous consumption and investment must move in opposite ways.

Now, in order to investigate the impact of news on the labour market, we add to the above problem an endogenous decision on employment. Specifically, as is common in the literature, we add a disutility from working; denoting the number of hours worked in a period by h_t, the utility function then becomes

$$U(c_t) - V(h_t) + U(c_{t+1}) - V(h_{t+1})$$

where V is a convex function. That is, given the negative sign in front of the V function we assume that both the first and second derivatives of V are positive; i.e., the consumer derives a disutility from working at an increasing rate. In this case, the budget constraint of the consumer is given by

$$c_t + c_{t+1} = w_t h_t + w_{t+1} h_{t+1},$$

where w_t and w_{t+1} denote the wage rate at period t and $t + 1$, respectively. Then, with some algebra, one can show that the optimal allocation is such the following equation holds in each period

$$\frac{U'(c_t)}{V'(h_t)} = w_t$$

Then, consistent with the discussion above, assume that the contemporaneous wage rate, w_t, does not change. Rather, the consumer learns that tomorrow wage rate, w_{t+1}, will increase. With the same logic as above, her optimal consumption reaction is to increase consumption immediately, implying that the numerator in the above equation falls (recall that U is a strictly concave function so if c_t increases then $U'(c_t)$ falls). Then, since we assume there is no change in current wage, it must be that the denominator falls. Given the assumptions made above regarding V, then it must be that the amount of hours worked falls in order to make the equation hold.

[4] Note that since the consumer lives for only two periods, she has no incentives to save in the last period, $t + 1$.

To summarize, the simple model discussed above predicts that in response to good news about the future, consumption increases, while investment, hours worked, and thus output fall. While the above discussion was based on simplified "toy model", these insights and predictions are present in more advanced modern sophisticated macroeconomic models. That is good (bad) news about the future leads to a recession (expansion). Prima facie, these results suggest that news shocks cannot be a basic driving force of the business cycle.

Modern Approach

In the last decade, many different channels and "modifications" to the benchmark model have been proposed in the literature where good (bad) news shocks about the future lead to an expansion (recession). Given the abundance of theoretical models where news shocks can indeed be a driver of the business cycle, it is beyond the scope of this entry to review all models and the interested reader is encouraged to read a thorough and more technical review of the existing work in Beaudry and Portier (2014).

However, a common theme is that the different channels proposed in the literature need to "overcome" the three basic forces that make the economy react negatively to good news. These are (i) the income effect that makes consumers want to consume more leisure when they receive good news about the future, which leads to a fall in employment and output, (ii) the lack of a reaction from the current labour demand from firms in response to good news about the future that allows the economy expand, and (iii) the lack of incentives to invest and build the capital stock in response to good news, before they actually materialize.

The Empirical Evidence

While the last 10 years have seen the advance of theoretical analysis where news shocks can be a driver of the business cycle, their empirical relevance is still an open question. The ambiguity is due to the fact that news shocks are essentially consumers' and firms' expectations and perceptions about the future. As such, they are inherently hard to measure. This identification challenge implies that there is no one unified way to measure the impact and effects of news shocks. Broadly speaking, during the last decade, three distinct methods have been used to tackle this challenge. In what follows we discuss these methods and their findings.

Reduced Form Vector-Auto-Regression Evidence

The key idea in this literature is to control for news by having a variable that is forward looking in its behaviour and thus is likely to react to news. This is the central idea in the seminal contribution of Beaudry and Portier (2006) who argue that stock prices are likely to contain news and expectations about the future. Under different scenarios, this assumption allows the researchers to identify news as innovations to stocks prices that are not driven by contemporaneous shocks to the economy. Beaudry and Portier (2006, 2014) show how under this identification, positive news shocks lead to an expansion in the economy where consumption, investment, GDP and hours worked all increase on impact.

This approach has been challenged by Barsky and Sims (2011) who propose an alternative statistical way to measure news. In their approach, news shocks lead to a persistent fall in hours worked. Hence, in fact, this pattern is consistent with the discussion in section "A Simple Model of News Shocks" where news shocks lead to a fall in employment and output. According to these results, there is no "puzzle" to be resolved and no need for a new theoretical paradigm since the existing one predicts the correct response of the economy to news.

Overall, this literature has been exploring the role of the different identifying assumptions. Hence there are different plausible combination of variables and identification methods that yield significantly different results. The effects of news shocks on the economy in this approach remain an open question.

Natural Experiments

These challenges have lead researchers to adopt a different, more direct approach to the identification of news shocks. Specifically, the idea is that from time to time, there are identifiable "natural experiments" that generate news shocks in markets. These events can then be used as a direct measurement of news shocks. As before, in this line of work, the results with respect to the effects of news are mixed.

For example, Bruckner and Pappa (2015) study the aggregate effects of bidding for the Olympic Games using panel data for 188 countries during the period 1950–2009. They find that investment, consumption and output significantly increased years before the actual event in bidding countries. Similarly, Alexopoulos (2011) studies periods where there is new information on technological developments that are not yet implemented. Alexopoulos (2011) finds that economic activity tends to pick up after these news events.

In contrast, Arezki et al. (2017) use oil and gas discoveries as a directly observable measure of news shocks about future income and output. Since there is usually a delay of about 5 years between a discovery and production, these discoveries serve as a natural candidate for news shocks. The authors find that after the news arrives, investment rises, employment falls, while GDP does not increase. Similarly, Mertens and Ravn (2012) use tax legislation as a way to measure news; specifically, when the difference between an announcement on a tax policy and its implementation is large enough, the authors consider that to be a news shock. In this work, they find that a pre-announced tax cut leads to different reaction than surprise tax cuts as the former leads to a decline in aggregate output, investment and hours worked, with no effect on consumption.

Overall, the "natural experiment" approach has an important advantage over the reduced form approach discussed above since the shocks are "identifiable". However, most of this literature focuses on shocks that are not cyclical in nature, making their implications for the relevance of news shocks to the business cycle an open question.

Maximum Likelihood Model Based Estimation

The third prominent approach is one where researchers use dynamic general equilibrium models to evaluate the importance of different shocks to economic fluctuations. In this line of work, researchers study modern equilibrium models where various shocks are considered. Through statistical methods the importance of news shocks can be assessed. The pioneering work in this area is Schmitt-Grohe′ and Uribe (2012) who find that news shocks account for roughly half of output fluctuations. Follow-up work in this area produced different results, and overall, the effects of news shocks on the economy within this approach remain an open question.

Overall, the maximum likelihood approach has an advantage since it formally embeds news shocks into state-of-the-art macroeconomic models which allow the researchers to conduct a "horse race" between different shocks to the economy. However, this alternative approach also has its limitations; the resulting decompositions and importance of news shocks are model-based and thus depend critically on the specific assumptions of the model. Hence the final conclusions are not "model free" and crucially depend on various modelling assumptions.

Conclusions

News shocks offer an attractive theory of expansions and recessions. In response to good news about the future, the economy "gears up" and the expansion is immediate. Similarly, in response to negative news about the future, the economy slides into a recession. While this story sounds plausible to many, it has proven surprisingly difficult to capture it in a modern theoretical business cycle model.

In the last decade, modern statistical and theoretical methods have been used to address this old question. This has sharpened our views on the contribution of news shocks to cyclical fluctuations. On the theoretical side, researchers have suggested many mechanisms via which news shocks can be a driver of the business cycle. On the empirical side, the evidence in support of the importance of news shocks is still an open question due to the inherent difficulty of identification.

Future work is required to assess the qualitative and quantitative importance of news shocks.

Bibliography

Alexopoulos, Michelle. 2011. Read all about it!! What happens following a technology shock? *American Economic Review* 101 (4): 1144–1179.

Arkezi, Rabah, Valerie A. Ramey, and Liugang Sheng. 2017. News shocks in open economies: Evidence from giant oil discoveries. *Quarterly Journal of Economics* 132 (1): 103–155.

Barsky, Robert, and Eric Sims. 2011. News shocks and business cycles. *Journal of Monetary Economics* 58 (3): 273–289.

Beaudry, Paul, and Franck Portier. 2004. An exploration into Pigou's theory of cycles. *Journal of Monetary Economics* 51: 1183–1216.

Beaudry, Paul, and Franck Portier. 2006. News, stock prices, and economic fluctuations. *American Economic Review* 96 (4): 1293–1307.

Beaudry, Paul, and Franck Portier. 2014. The news view of business cycles: Insights and challenges. *Journal of Economic Literature* 52 (4): 993–1074.

Brückner, Markus, and Evi Pappa. 2015. News shocks in the data: Olympic games and their macroeconomic effects. *Journal of Money, Credit and Banking* 47 (7).

Mertens, Karel, and Morten Ravn. 2012. Empirical evidence on the aggregate effects of anticipated and unanticipated US tax policy shocks. *American Economic Journal: Economic Policy* 4 (2): 145–181.

Pigou, A.C. 1927. *Industrial fluctuations*. London: MacMillan.

Rebelo, Sergio. 2005. Real business cycle models: Past, present and future. *Scandinavian Journal of Economics* 107 (2): 217–238.

Schmitt-Grohe′, Stephanie, and Martin Uribe. 2012. What's news in business cycles. *Econometrica* 80 (6): 2733–2764.

Nicholls, William Hord (1914–1978)

Karl A. Fox

Born on 19 July 1914 in Lexington, Kentucky, Nicholls was the son of a respected agricultural economist. After completing a liberal arts degree at the University of Kentucky, Nicholls took up graduate work in economics at Harvard (1934–7), where he was strongly influenced by John D. Black and Edward Chamberlin. He held faculty positions at Iowa State College during 1938–44, at the University of Chicago during 1945–8, and at Vanderbilt University from 1948 until his death on 4 August 1978.

Nicholls's *Imperfect Competition within Agricultural Industries* (1941) introduced a generation of agricultural economists to theories of imperfect competition. Chamberlin (1933) had concentrated on the selling side of imperfect markets. Nicholls demonstrated the prevalence and importance of similar structures on the buying side of agricultural markets, where a few large firms in a given industry faced many uncoordinated farmer-sellers. Moreover, these large firms were simultaneously processors and distributors, confronting farmers as oligopsonists and consumers as oligopolists. Nicholls called this industry structure 'oligopoly-oligopsony', and his detailed analysis of it was his most important contribution. The quality of his theoretical analysis was at least equal to Chamberlin's and he used it as a framework for proposed empirical research, an orientation with particular appeal to agricultural economists.

Nicholls also made an important contribution to the estimation of labour productivity functions (1948), and worked in such diverse fields as the study of price policies in the cigarette industry (1951), the formulation of development policies for the Southern Region of the United States (1960, 1961) and agricultural-industrial development policies for Brazil (1969). He was one of the most versatile and distinguished agricultural economists of his generation.

Selected Works

1941. *Imperfect competition within agricultural industries: A theoretical analysis of imperfect competition with special application to the agricultural industries.* Ames: Iowa State University Press.

1948. *Labor productivity functions in meat packing.* Chicago: University of Chicago Press.

1951. *Price policies in the Cigarette Industry: A study of 'concerted action' and its social control, 1911–1950.* Nashville: Vanderbilt University Press.

1960. *Southern tradition and regional progress.* Chapel Hill: University of North Carolina Press.

1961. Industrialization, factor markets, and agricultural development. *Journal of Political Economy* 69: 319–340.

1969. The transformation of agriculture in a presently semi-industrialized country: The case of Brazil. In *The role of agriculture in economic development,* ed. Erik Thorbecke. New York: Columbia University Press.

References

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

Nicholson, Joseph Shield (1850–1927)

B. F. Kiker

Nicholson was born in Wrawby, Lincolnshire and educated at Cambridge. An influential economist and prolific writer, he held the Chair of Political Economy at the University of Edinburgh from 1880 to 1925.

In the tradition of Smith, Ricardo and J.S. Mill, his *Principles of Political Economy* (1893), although eclectic and dwarfed by Marshall's work, was thought by Schumpeter to be a 'creditable achievement' (1954, p. 830). Perhaps his most original work was in the area of monetary economics and capital theory – particularly, human capital (Kiker 1966). Nicholson's *Treatise on Money and Essays on Present Monetary Problems* (1888) provides an excellent treatment of the state of monetary theory at the turn of the century, containing the best discussion in support of bimetallism to be found at that time. Later, in an essay titled *Inflation* (1919) he strongly criticized the fiat monetary standard that Great Britain had adopted during World War I and advocated a return to the gold standard that existed in 1914. In *The Effects of Machinery on Wages* (1892), Nicholson argued that the excess supply of goods resulting from industrialization would have an unfavourable impact on real wages and employment. In his concept of capital, he treated the acquired skills and abilities of man as capital – emphasizing the importance of education and training on the productivity of labour and the necessity for viewing a unit of productive labour from the point of view of a lifetime, rather than one productive period (Nicholson 1893, 1922). Although his methodology was crude, Nicholson estimated the value of the stock of human capital in the United Kingdom in order to provide some insight into the necessary relationship between labour and capital and the historical progress of man (Nicholson 1891).

See Also

► Human Capital

Selected Works

1888. Treatise on money and essays on present monetary problems. London: W. Blackwood & Sons.

1891. The living capital of the United Kingdom. Economic Journal 1: 95–107.

1892. The effects of machinery on wages. London: Swan Sonnenschein & Co.

1893. Principles of political economy. London: Adam & Charles Black.

1919. Inflation. London: P.S. King & Son.

1922. Elements of political economy. Reprint of 2nd edn of 1906, London: A. & C. Black.

Bibliography

Kiker, B.F. 1966. The historical roots of the concept of human capital. *Journal of Political Economy* 74: 481–499.

Schumpeter, J.A. 1954. *History of economic analysis*. New York: Oxford University Press.

Nikaido, Hukukane (1923–2001)

Kazuo Nishimura

Keywords

Existence of general equilibria; Factor substitution; Gale–Nikaido lemma; Harrod–Domar model; Imperfect competition; Knife-edge property; Minimax theorem; Monopolistic competition; Morishima, M.; Nikaido, H.; Objective demand functions; Solow, R.; Turnpike theorems; von Neumann growth model

JEL Classifications

B31

Hukukane Nikaido graduated from the Department of Mathematics, University of Tokyo, in 1949. While he was a university student he became interested in economics and studied Marx's *Das Kapital*, Hicks's *Value and Capital* and Samuelson's *Foundations of Economic Analysis*. After graduating from the University of Tokyo and becoming an Associate Professor of Mathematics at the Tokyo College of Science, he wrote papers concerning the von Neumann growth model and the minimax theorem (1954a, b; 1955). He also worked on the existence of equilibria in the general equilibrium model with many firms and many consumers. His paper had been completed independently of Arrow and Debreu (1954) and McKenzie (1954), and was published in *Metroeconomica* (1956a). One of his results in the existence proof is now known as Gale–Nikaido lemma. These achievements led him to visit Stanford in 1955–6 at the invitation of Kenneth Arrow. At Stanford, Nikaido started to work on the existence of general equilibria for an economy with infinitely many commodities (1956b; 1957), and then published in the *Journal of the Mathematical Society of Japan* (1959a). His contributions on the existence of general equilibria in the infinite dimensional space had long remained unknown.

After he returned from Stanford he was invited by Michio Morishima to join the Institute of Socioeconomic Research at Osaka University. There he began to work on the stability of general equilibria (1959b, 1960, 1964a). Osaka was very active in research in those days, and many well-known economists from abroad visited Osaka University. One was John Hicks, who in those days was interested in the turnpike theorem of multi-sector economic growth. Turnpike theorems had been proved by Morishima (1961) and Radner (1961). Radner's result was improved by Nikaido (1964b). David Gale also visited Osaka in 1961. He and Nikaido wrote a joint paper (1965) on the uniqueness of solutions of nonlinear simultaneous equations; the condition used in the paper has been called the Gale–Nikaido condition. In 1969 Nikaido moved from Osaka University to Hitotsubashi University in Tokyo. His previous research was published in a book, *Convex Structures and Economic Theory* (1968), which has been read by many graduate students and researchers worldwide.

After moving to Hitotsubashi University, he began to work on general equilibrium combined with monopolistic competition. Nikaido recognized, however, that the demand function, a partial equilibrium theoretic construction, involves inconsistencies in a general equilibrium situation. By introducing the concept of 'objective demand functions' Nikaido explored the existence of monopolistically competitive equilibria (1974). His research was published in *Monopolistic Competition and Effective Demand* (1975).

Thereafter Nikaido developed his previous work on imperfect competition into a dynamic model (1978, 1979, 1980a). His main concern was in the theory of out-of-equilibrium adjustments. Nikaido also re-examined the knife-edge property in the Harrod–Domar model and the stability property in Solow's neoclassical growth model. He showed that the stability of Solow's model depends on the assumption that an investment is equal to a saving, rather than the smooth factor substitution as had been generally believed, and that, if an intended investment is not the same as a realized investment, the steady state solution is not necessarily stable and the imbalance is not solved even with flexible factor substitution (1975, 1980b).

In 1983 Nikaido joined Tsukuba University and later Tokyo International University. From then on he spent most of his time working on Marxian economics and then Keynesian models, using dynamic analysis developed in his earlier research. They were problems that he had been concerned with when he was a young university student.

See Also

► Existence of General Equilibrium

Selected Works

1954a. Note on the general economic equilibrium for nonlinear production functions. *Econometrica* 22: 49–53.

1954b. On von Neumann's minimax theorem. *Pacific Journal of Mathematics* 4(1): 65–72.

1955. New aspects of von Neumann's model with special regard to computational problems. *Annals of the Institute of Statistical Mathematics* 6: 223–230.

1956a. On the classical multilateral exchange problem. *Metroeconomica* 8(2): 135–145.

1956b. On the existence of competitive equilibrium for infinitely many commodities. Technical Report No. 34, Department of Economics, Stanford University.

1957. Existence of equilibrium based on Walras' Law. ISER Discussion Paper No.2, Institute of Social and Economic Research, Osaka University.

1959a. Coincidence and some systems of inequalities. *Journal of the Mathematical Society of Japan* 11: 354–373.

1959b. Stability of equilibrium by the Brown-von Neumann differential equation. *Econometrica* 27: 654–671.

1960. (With H. Uzawa.) Stability and nonnegativity in a Walrasian tâtonnement process. *International Economic Review* 1(1): 50–59.

1964a. Generalized gross substitutability and extremization. In *Advances in game theory*, ed. M. Dresher, L. Shapley and A. Tucker. Princeton: Princeton University Press.

1964b. Persistence of continual growth near the von Neumann Ray: A strong version of the Radner turnpike theorem. *Econometrica* 32: 151–162.

1965. (With D. Gale.) The Jacobian matrix and global univalence of mappings. *Mathematische Annalen* 159(2): 81–93.

1968. *Convex structures and economic theory.* New York: Academic Press.

1974. What is an objective demand function? *Zeitschrift für Nationalökonomie* 34: 291–307.

1975. *Monopolistic competition and effective demand*. Princeton: Princeton University Press.

1975. Factor substitution and Harrod's knife-edge. *Zeitschrift für Nationalökonomie* 35: 149–154.

1978. (With S. Kobayashi.) Dynamics of wage-price spirals and stagflation in the Leontief–Sraffa system. *International Economic Review* 19: 83–102.

1979. Wage-price spiral under monopoly: 'What is an objective demand function?' set in motion. *Zeitschrift für Nationalökonomie* 39: 299–313.

1980a. Prices and income distribution in a Leontief economy. *Journal of Economic Behavior and Organization* 1: 61–79.

1980b. Harrodian pathology of neoclassical growth. *Zeitschrift für Nationalökonomie* 40: 111–134.

Bibliography

Arrow, K., and G. Debreu. 1954. Existence of an equilibrium for a competitive economy. *Econometrica* 22: 265–290.

McKenzie, L. 1954. On equilibrium in Graham's model of world trade and other competitive systems. *Econometrica* 22: 147–161.

Morishima, M. 1961. Proof of a turnpike theorem: The 'no joint production case'. *Review of Economic Studies* 28: 89–97.

Radner, R. 1961. Paths of economic growth that are optimal states: A turnpike theorem. *Review of Economic Studies* 28: 98–104.

No Trade Theorems

Ricardo Serrano-Padial

Abstract

No trade theorems represent a class of results showing that, under certain conditions, trade in asset markets between rational agents cannot be explained on the basis of differences in information alone. They pose a challenge to provide a theoretical justification of the high trade volumes observed in financial markets. This article overviews existing no trade theorems and discusses alternative approaches to modelling information-based trade.

Keywords

Asset markets; Demand shocks; No trade theorems; Trade volume; Uncertainty

JEL Classifications

C72; D44; D82

The very high levels of daily trading activity observed in many financial markets are often attributed to speculation: agents hold different views about how much assets are worth – for instance, they may have different expectations about future prices – and these differences should lead them to trade. Rational agents typically hold distinct opinions if they privately observe different information. Thus, as this reasoning goes, the arrival of asymmetric information should induce agents to trade. No trade theorems challenge this premise by showing that, if the initial asset allocation is commonly known to be efficient, then any proposed trade after the arrival of new information cannot lead to a Pareto improvement over the initial allocation as long as traders interpret information in a similar fashion. As a consequence, agents do not have an incentive to trade after receiving the new information.

The logic behind these results goes as follows. Consider the market of an asset in which it is common knowledge, before the arrival of new information, that the initial allocation is efficient. This implies that the asset is allocated to the agents that value it the most. Otherwise, there would be a mutually beneficial trade in which some agents holding units of the asset sell them to agents with higher valuations, contradicting the fact that the initial allocation is efficient. With the arrival of new information, some agents may value the asset more than those holding it if the former receive a more positive signal about the asset value than the latter *and* if agents only consider their own signal when updating their beliefs about the asset. Thus, without considering any additional information, they could get to a better allocation by trading. However, the fact that an agent is willing to acquire units of the asset at a given price conveys some information about the signal she received. Since traders interpret signals in a similar fashion, this *additional* information should be taken into account by the potential sellers, who then revise upwards their beliefs about the asset and become unwilling to trade at the proposed price, even though they would have agreed to trade had they only considered their own signals.

No trade theorems have proven to be a major hurdle not only in the modelling of information-based trade, but also in analysing other aspects of trade under incomplete information, such as information aggregation or information acquisition in markets. In particular, they highlight a well-known paradox associated with the efficient markets hypothesis (Grossman and Stiglitz 1980): if market prices reflect all the relevant information possessed by agents about asset values, traders sharing common prior beliefs cannot take advantage of their private information and thus have no incentive to acquire it in the first place.

Theoretical Results

There are three basic elements in no trade theorems: efficiency of the initial allocation, common knowledge of the institutional and informational environment, and some degree of agreement in they way new information should be interpreted. Efficiency of the initial allocation implies that, before the arrival of information, there is no alternative allocation that Pareto dominates it. Common knowledge requires that agents have correct beliefs about the beliefs of others and about their equilibrium behavior and all this is commonly known by all agents. Finally, traders need to exhibit some similarities in the way they interpret new information. The strongest assumption implies agents having common priors about the distribution of asset values and private information. In this context, rational agents cannot 'agree to disagree' (Aumann 1976; Geanakoplos 1994), i.e. they cannot hold different beliefs after observing the same information. Weaker notions include noisy versions of concordant beliefs (priors about the value of the asset may differ but traders share the same beliefs about the distribution of information conditional on asset values) and consistent beliefs (receiving the same information leads to the same posterior beliefs).

The gist of no trade theorems is to show that efficiency of the allocation before the arrival of new information leads to efficiency of the same allocation after agents receive new information. Accordingly, the stronger the notion of efficiency, the stronger the compatibility requirements on agents' beliefs (see Holmstrom and Myerson (1983) for a classification of efficiency notions). It turns out that the relevant type of efficiency depends on the notion of market equilibrium, which establishes which types of trade are considered feasible. Most no trade theorems focus on three different equilibrium notions: common knowledge trade; incentive compatible trade; and rational expectations equilibria. Common knowledge trade refers to the case in which agents do not behave strategically, trades are public, and markets are complete (i.e. there exist a complete set of Arrow–Debreu securities), so that trades contingent on the true state of the world are possible. In this context, it will be common knowledge for rational traders that when a public trade is carried out it is because it is feasible and mutually beneficial. If agents behave strategically feasible trades are those that induce agents to truthfully reveal their private information (incentive compatible trade). Finally, rational expectations equilibrium refers to the case in which agents are non-strategic and trades are contingent on prices, which potentially reflect agents' private information. Morris (1994) identifies the relevant types of efficiency of the initial allocation that lead to no trade for these alternative equilibrium concepts and provides the corresponding restrictions on traders' beliefs.

Early no trade theorems (Rubinstein 1975; Kreps 1977; Tirole 1982; Sebenius and Geanakoplos 1983) focus on the common prior assumption to show that if agents are risk averse no mutually beneficial trade exists after the arrival of information. Milgrom and Stokey (1982) show that concordant beliefs are sufficient to rule out common knowledge trade. Generalising this result, Dow et al. (1990) show that, if the initial allocation is *ex ante* efficient with respect to prior beliefs, common knowledge trade is ruled out regardless of the nexus between prior and posterior beliefs, as long as agents are rational. On the other hand, Morris (1994) shows that consistent beliefs are associated with no trade when agents are strategic (incentive compatible trade). There are also no trade theorems dealing with the case where all information is public (Hakansson et al. 1982).

One may be tempted to regard no trade theorems as theoretical artifacts without much empirical content. After all, agents face uncertainty in most asset markets about the number of participants, their beliefs and the sources of information they may have access to, rendering the common knowledge assumption too restrictive. However, as we explore ways to break no trade results it becomes apparent that speculative trade does not trivially follow from weakening the conditions underlying no trade theorems.

Alternative Sources of Information-based Trade

There are different routes taken by the literature to elicit trade in models of asset markets under asymmetric information. The most frequent approaches either weaken the common knowledge assumption or exogenously introduce 'liquidity' in the market, i.e. make the initial allocation inefficient due to demand shocks. Other approaches allow agents to ‘agree to disagree’ by introducing bounded rationality. Finally, some models introduce uncertainty in the market. Here is a brief overview:

- Lack of common knowledge: There are different ways in which the common knowledge assumption can be relaxed. For instance, traders may not have common knowledge about potential disagreements over the asset value that asymmetric information creates, or they may not have common knowledge about the rationality of other traders. One way of relaxing it is to require that traders have common beliefs rather than common knowledge (see Monderer and Samet (1989)) for a definition of common beliefs), which implies that agents are not completely certain about other agents’ beliefs or their rationality. In the context of common knowledge trade, Neeman (1996) shows that common belief of potential disagreements leads to trade only if rationality is not common knowledge. One interpretation of this result is that speculative trade can occur if agents exhibit some overconfidence: even if all traders are rational, some believe it is possible that other traders may not be so, or that other traders may (erroneously) think so. Another approach to relax common knowledge is to let agents interpret information in a dissimilar fashion or to introduce doubts about how to interpret information, whether public or private. Differences in interpretation of public information among bayesian agents may arise with common priors if agents additionally receive private information (Andreoni and Mylovanov 2010) or when they hold different priors and are uncertain about how to interpret some signals (Acemoglu et al. 2009).
- Demand shocks/noise traders: Many theoretical models aimed at studying information aggregation, insider trading and other interesting phenomena in financial markets get around no trade by introducing exogenous sources of liquidity, that is, positive demand/supply for the asset at any given price. This is done by either having aggregate demand shocks (Hellwig 1980; Diamond and Verrecchia 1981; Kyle 1985, 1989) or by introducing agents with immediate (exogenous) liquidity needs willing to sell/buy at current prices (Glosten and Milgrom 1985; Easley and O’Hara 1992).
- Bounded rationality: There are many ways in which bounded rationality and psychological biases in the way agents update beliefs can elicit trade. Geanakoplos (1989) characterizes the conditions on the information structures associated to bounded rationality under which speculative trade is possible, which basically require agents’ information structures not being represented by a partition of the space of possible states of the world (see Rubinstein and Wolinsky (1990) for an example of speculative trade when information structures are non-partitional).
- Uncertainty: A potential way to elicit trade is to introduce (Knightian) uncertainty by letting agents hold multiple priors rather than a single one. In this context, Kajii and Ui (2009) show that for certain classes of preferences under uncertainty the updating rule mapping priors to posteriors is the key determinant of the existence of speculative trade. A no trade theorem still applies if the set of posterior beliefs is the collection of all conditional probability distributions of the priors (*full bayesian updating*). If, on the other hand, the set of posteriors is the collection of all conditional distributions that maximize the likelihood of the observed private information (*maximum likelihood updating*) speculative trade can happen. Dow et al. (1990) provide an early example of trade in the presence of uncertainty when the arrival of information completely resolves all of the initial uncertainty.

Bibliography

Acemoglu, D., V. Chernozhukov, and M. Yildiz. 2009. Fragility of asymptotic agreement under Bayesian learning. *Mimeo*.

Andreoni, J., and T. Mylovanov. 2010. Diverging opinions. *Mimeo*.

Aumann, R.J. 1976. Agreeing to disagree. *Annals of Statistics* 4: 1236–1239.

Diamond, D.W., and R.E. Verrecchia. 1981. Information aggregation in a noisy rational expectations economy. *Journal of Financial Economics* 9: 221–235.

Dow, J., Madrigal, V., and Werlang, S.d.C. 1990. Preferences, common knowledge and speculative trade. IFA Working Paper, pp. 125–190.

Easley, D., and M. O'Hara. 1992. Time and the process of security price adjustment. *Journal of Finance* 47(2): 577–605.

Geanakoplos, J.D. 1989. Game theory without partitions, and applications to speculation and consensus. Cowles Foundation Discussion Paper No. 914.

Geanakoplos, J. 1994. Common knowledge. In *Handbook of Game Theory II*, ed. R. Au-mann and S. Hart, 1437–1496. Amsterdam: Elsevier Science.

Glosten, L.R., and P.R. Milgrom. 1985. Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics* 14: 71–100.

Grossman, S.J., and J.E. Stiglitz. 1980. On the impossibility of informationally efficient markets. *American Economic Review* 70(3): 393–408.

Hakansson, N.H., J.G. Kunkel, and J.A. Ohlson. 1982. Sufficient and necessary conditions for information to have social value in pure exchange. *Journal of Finance* 37: 1169–1181.

Hellwig, M.F. 1980. On the aggregation of information in competitive markets. *Journal of Economic Theory* 22: 477–498.

Holmstrom, B., and R.B. Myerson. 1983. Efficient and durable decision rules with incomplete information. *Econometrica* 51(6): 1799–1819.

Kajii, A., and T. Ui. 2009. Interim efficient allocations under uncertainty. *Journal of Economic Theory* 144(1): 337–353.

Kreps, D.M. 1977. A note on 'fulfilled expectations' equilibria. *Journal of Economic Theory* 14: 32–44.

Kyle, A.S. 1985. Continuous auctions and insider trading. *Econometrica* 53(6): 1315–1336.

Kyle, A.S. 1989. Informed speculation with imperfect competition. *Review of Economic Studies* 56(3): 317–355.

Milgrom, P.R., and N. Stokey. 1982. Information, trade and common knowledge. *Journal of Economic Theory* 26: 17–27.

Monderer, D., and D. Samet. 1989. Approximating common knowledge with common beliefs. *Games and Economic Behavior* 1: 170–190.

Morris, S. 1994. Trade with heterogeneous prior beliefs and asymmetric information. *Econometrica* 62(6): 1327–1347.

Neeman, Z. 1996. Common beliefs and the existence of speculative trade. *Games and Economic Behavior* 16: 77–96.

Rubinstein, A. 1975. Security market efficiency in an Arrow–Debreu economy. *American Economic Review* 65: 812–824.

Rubinstein, A., and A. Wolinsky. 1990. On the logic of 'agreeing to disagree' type results. *Journal of Economic Theory* 51: 184–193.

Sebenius, J.K., and J.D. Geanakoplos. 1983. Don't bet on it: contingent agreements with asymmetric information. *Journal of the American Statistical Association* 78: 424–426.

Tirole, J. 1982. On the possibility of speculation under rational expectations. *Econometrica* 50(5): 1163–1182.

Noise Traders

James Dow and Gary Gorton

Abstract

Noise traders are agents whose theoretical existence has been hypothesized as a way of solving certain fundamental problems in financial economics. We briefly review the literature on noise traders.

Keywords

'No trade' theorem; Arbitrage; Grossman, S.; Hedging; Imperfect information revelation; Information aggregation and prices; Information costs; Insurance motive; Liquidity traders; Market microstructure; Market selection hypothesis; Noise; Noise traders; Private information; Rational expectations equilibrium

JEL Classifications

G14

'Noise traders' are economic agents who trade in security markets for noninformation-based reasons. The existence of noise traders was theoretically posited as a solution to the 'no trade' or 'no speculation' results of Grossman and Stiglitz (1980) and Milgrom and Stokey (1982). These authors showed that it is impossible under most circumstances for an agent with superior information to profit from that information by trading. The intuition for the 'no trade' result is as follows. A buyer of an asset is prepared to pay a seller a price p only if the buyer believes that, conditional on the seller agreeing to sell the asset, the value of the asset exceeds p. But then the seller, knowing this, is at least as well off keeping the asset. So no one trades.

But we do observe trade in the world. Moreover, no trade is difficult to reconcile with the notion of asset market efficiency, in which prices allegedly contain all available information. If some agents produce costly private information and then trade on their private information, security prices will reflect some or all of the information and hence become more informationally efficient. To explain how informed traders can cover the costs of information production when they trade in securities markets, someone in the market must lose money trading against them. 'Noise traders' or 'liquidity traders' are the names given to the traders who lose money, on average, when they trade. Their trade then provides the subsidy to cover the informed traders' cost of information production.

The idea that there are traders who systematically lose money trading securities leads to obvious questions. Do noise traders really exist? Who exactly are noise traders in reality? How do noise traders survive and persist when they are losing money trading?

Rational Expectations and Efficient Security Markets

In security markets, prices are alleged to reflect 'all available information'. But how does this come about? What is the information, and how is it aggregated into the price? The concept of a rational expectations equilibrium (REE) gave formal content to the notion of 'market efficiency', which has been a central concept in financial economics since the 1960s. The idea is that, if agents understand the economy and understand how markets work, they know that current prices reflect the information which is known to some agents but maybe not to others. The uninformed agents understand the link between current prices and the information of the informed agents, and so can infer something about the information in prices. When the prices that prevail in equilibrium coincide with what the uninformed agents can learn from the prices and with the actions taken by the informed agents, who trade on their information knowing that the uninformed agents will infer (some or all) of the information, then the equilibrium is said to be a rational expectations equilibrium. The idea that prices can convey information, in the sense of REE, is due to Lucas (1972). (See also Green 1977; Radner 1979. Grossman 1981, provides a brief intellectual history of REE; see also Allen and Jordan 1998.)

But, when all the information of the informed agents is revealed in a fully revealing REE, there is a problem if information acquisition is costly. Grossman (1976) considers a model of the stock market in which there are two types of traders: 'informed' and 'uninformed'. Informed traders take positions in the market based on their information. Uninformed traders have no information but know that prices will reflect the information of the informed traders. Grossman shows that the equilibrium prices aggregate and reveal the information perfectly, 'but in doing this the price system eliminates the private incentive for collecting the information' (1976, p. 574). Grossman is quite clear in identifying the paradox, but he also proposes a solution:

> When a price system is a perfect aggregator of information it removes private incentives to collect information. If information is costly, there must be *noise* in the price system so that traders can earn a return on information gathering. If there is no *noise* and information collection is costly, then a perfect competitive market will break down because no equilibrium exists where one collects information. (1976, p. 574; emphasis added)

Beja (1976) also argues that REE and costly endogenous information acquisition are not compatible when agents are strategic and that consequently asset prices cannot be efficient.

So 'noise' is required if agents are to acquire and trade on their costly information. But what is this 'noise'? The example of 'noise' that Grossman points to is 'an uncertain total stock of the risky asset' (1976, p. 574). He describes 'noise' simply as 'many other factors' (1977, p. 431). The device of adding a random noise term to the aggregate supply of the asset is used in Grossman and Stiglitz (1980). They show that, when information production is costly and there is noise in the asset supply, then some traders will acquire information and trade, but rational expectations prices will not be fully revealing.

If there is uncertainty about the supply of the asset in the market, or about the level of demand, or about the risk aversion of other traders, then uninformed traders cannot be sure that prices reflect the information of the informed traders. The basic idea is that the uninformed traders confuse the private information with uncertainty about the other unknown variables. It is this additional uncertainty, or noise, which makes it possible for the informed traders to trade without perfectly revealing their information, and hence profit from its production.

The device of adding a noise term to aggregate supply does result in REE that are only partially revealing. Unfortunately, there were two problems with this approach as a general matter. First, the partially revealing REE models require somewhat special assumptions. Second, it was not clear what the proposed noise shock to aggregate supply really corresponds to in reality. (There are other problems as well. Hellwig 1980, pointed out that REE requires traders to act rationally with respect to information, yet they ignore the effect of their transactions on the price. This was deemed the 'Schizophrenia problem': '... Grossman's agents are slightly schizophrenic: (Hellwig 1980, p. 478'. The model in Kyle 1985, avoided this problem.) On the first point, Green's (1977) non-existence example uses a noise term on the traders' endowments, and suggests that this will not be a suitable basis for a general approach. The general equilibrium literature did develop a number of generalizations, including, for example, the difference between the dimensions of the signals and the dimension of the prices (see, for example, Jordan 1983; Ausubel 1990). Others have provided slightly different models that have partially revealing equilibria, but still there seems to be no general approach (see, for example, Allen 1981; Allen and Jordan 1998, for a discussion).

Noise Traders

REE models assume that traders maximize expected utility with *rational* beliefs, where rational beliefs are defined to be consistent with the model itself. There may be 'noise', but this was not viewed as emanating from incorrect beliefs. (There is the issue of how traders come to understand the model, that is, how they learn. On that question see, for example, Blume et al. 1982; Blume and Easley 2004.) In general, the notion of 'noise' in the REE literature was somewhat vague and corresponded to a random error term added to the aggregate excess demand function. Understanding the role of 'noise' appeared to require leaving the REE world and explicitly detailing the origin of noise. This was done by Kyle (1985).

Kyle posited the existence of 'uninformed noise traders who trade randomly' (1985, p. 1315). (In private correspondence, Kyle said that he did not coin the term 'noise trading' but attributes it to Sanford Grossman.) Kyle identified certain people as trading in a way which made noise in the sense that their trade was not based on information. That is, he explicitly posited the existence of a class of agents – people – who traded in a certain way so as to fulfil the role of 'noise'. By explicitly introducing noise traders, Kyle focused attention on the details of the trading process. This became the foundation for the study of market microstructure. (Garman 1976, appears to have been the first to use the term 'market microstructure'. See Easley and O'Hara 2003, for a survey of the microstructure literature.) Around the same time Kyle, Glosten and Milgrom introduced a similar class of agents: '...we assume that there are informed investors and purely "liquidity" traders' (1985, p. 76). Earlier, Treynor (1971, under the pseudonym W. Bagehot) talked about 'liquidity-motivated' traders.

In REE models agents do not act strategically; the process of learning from prices occurs in equilibrium (as opposed to happening in real time), and the details of trading are treated in reduced form (agents submit demand functions to an auctioneer). Kyle and Glosten and Milgrom changed this by specifying the trading process in a way that was not possible in REE models. In both papers there is a competitive market-maker who receives orders from traders, at least one of whom has superior information. The market-maker must infer the information of the informed trader from the order flow. The market-maker knows that some traders are privately informed, and that others are not trading based on any superior information (the noise traders). Inference about information occurs as the market-maker learns by watching the order flow. Gradually, the market-maker changes his price to reflect the information.

Still, the noise traders in this new type of model were not well-motivated. In fact, their motives are not explained. They earn a lower-than-average return than the informed agents, who earn an above-average return. If the uninformed noise traders could at least buy the market portfolio, then they could earn the average return on the market. But in fact they are not allowed to buy the market portfolio. That is their root problem (see Dow and Gorton 1995).

Diamond and Verrecchia (1981) suggest adding a noise term to agents' risk exposures (their endowments). Risk-averse agents will then have an insurance motive for trading. DeMarzo and Duffie (1999) propose a model where different traders have different discount rates. Shocks to their discount rates provide an incentive to trade that other traders cannot distinguish from speculative trading intended to profit from information about the liquidation value of the asset. These papers solve the theoretical problem of finding a logically consistent model that can be used as a basis for economic analysis, including welfare statements, of markets with imperfect information revelation. Papers that have applied these models in various settings include Biais and Mariotti (2005) (for the DeMarzo and Duffie model) and Dow and Rahi (2000, 2003) (for the Diamond and Verrecchia approach).

But is it really plausible to believe that there is a significant demand for individual stocks or bonds based on an insurance motive? Stock indexes, exposure to the yield curve, or foreign currency could experience demand variations due to insurance motives, but there are close substitutes for individual stocks and bonds from a risk point of view. Also, if investors do start off with different discount factors, one would expect them to trade these differences away.

In other words, plausibly the demand curve for an individual asset should be almost perfectly elastic. The price at which it becomes elastic (given the prices of all other assets) should be almost identical for all agents. Hence, we revert to the situation where the asset has a unique fundamental value that all agents will agree on if they have the same information about the asset's cash flows. So the question of who noise traders actually are remains open.

Who Are the Noise Traders?

The details of the identity of noise traders or liquidity traders were initially left vague. For example, Glosten and Milgrom write of exogenous events motivating their trade, like 'job promotions or unemployment, deaths or disabilities...' (1985, p. 77). These shocks were not well identified. Notably, noise traders were modelled as equally likely to be buying or selling securities, which, while making models technically tractable, is counter-intuitive. Exogenous reasons for needing money and hence having to sell securities seems more natural than exogenous reasons for having to buy securities.

The details of the identity of noise traders are important, because if noise traders are simply irrational there is clearly an incentive for 'smart money' to take advantage of them, and eventually eliminate them from the market. The 'market selection hypothesis' holds that irrational traders will eventually be driven out of the market. Noise traders should not survive, and so cannot play the role envisioned for them. In fact, it has long been argued that rational traders will eliminate irrational traders from the market by taking their money when they trade at incorrect prices. This process is what causes prices to be driven to (or close to) fundamental values (see, for example, Friedman 1953).

Noise traders can survive only if there are some frictions or barriers preventing them from being eliminated by the smart money. That is, there must be some limits to arbitrage. One possibility is that the smart money has a limited horizon over which trade can occur. With a limited horizon, the noise traders could cause losses to the smart money by moving prices further away from fundamentals. This is the idea in DeLong et al. (1990), Dow and Gorton (1994), and Shleifer and Vishny (1997). These papers argue that there are 'limits to arbitrage', providing an explanation for the persistence of noise trade.

Still, the question remains: who are the noise traders? On one view, noise traders are simply individuals who are less than rational; they are subject to behavioural biases and fads. For example, Shiller (1984) argued that some investors rely on 'popular models' which are wrong, and also that they can be subject to fads. Along the same lines, Shleifer and Summers (1990, p. 19) wrote: 'their demand for assets is affected by their beliefs or sentiments that are not fully justified by fundamental news.' A large literature argues that individual investor trading is subject to a myriad of psychological biases, and that such individuals may use various heuristics, 'popular models', as the basis for their investment decisions. This literature is surveyed in Barberis and Thaler (2003).

A second rationale for noise trading focuses not on individual investors but on professional traders and money managers ('funds') hired by principals/investors. Funds do not invest and trade their own money; they work for others. This creates a potential conflict of interest or agency problem. This notion is developed by Dow and Gorton (1997). They argue that churning by funds, which occurs when they do not become informed and want to pretend that they have, is 'noise' in a setting where all market participants are rational. Among the other agents in the market are hedgers. Noise trading, being a manifestation of agency problems, reduces the profitability of traders to the employers of the traders and money managers. But it benefits hedgers who earn more when they hedge. Consequently, they hedge more, which in turn can support more informed fund trading. Dow and Gorton (1997) show that a 'small' amount of hedging demand can result in a 'large' noise. Irrationality is not needed to explain significant amounts of noise.

Summary

Noise traders play an essential role in modern finance theory, but their identities, motivations, and ability to persist remain topics of research.

We thank Pete Kyle for comments.

Bibliography

Allen, B. 1981. A class of monotone economies in which rational expectations equilibria exist but prices do not reveal all information. *Economics Letters* 7: 227–232.

Allen, B., and J. Jordan. 1998. The existence of rational expectations equilibrium: A retrospective. Research Department Staff Report No. 252, Federal Reserve Bank of Minneapolis.

Ausubel, L. 1990. Partially-revealing rational expectations equilibrium in a competitive economy. *Journal of Economic Theory* 50: 93–126.

Bagehot, W. (pseudonym for Jack Treynor). 1971. The only game in town. *Financial Analysts Journal* 22: 12–14.

Barberis, N., and R. Thaler. 2003. A survey of behavioral finance. In *Handbook of the economics of finance: Financial markets and asset pricing*, ed. G. Constantinides, M. Harris, and R. Stulz, vol. 1B. Amsterdam: North-Holland.

Beja, A. 1976. The limited information efficiency of market processes. Working Paper No. 43, Research Program in Finance, University of California, Berkeley.

Biais, B., and T. Mariotti. 2005. Strategic liquidity supply and security design. *Review of Economic Studies* 72: 615–649.

Blume, L., and D. Easley. 2004. If you're so smart, why aren't you rich? Belief selection in complete and incomplete markets. *Econometrica* 74: 929–966.

Blume, L., M. Bray, and D. Easley. 1982. Introduction to the stability of rational expectations equilibrium. *Journal of Economic Theory* 26: 313–317.

DeLong, J., A. Shleifer, L. Summers, and R. Waldman. 1990. Noise trader risk in financial markets. *Journal of Political Economy* 98: 703–738.

DeMarzo, P., and D. Duffie. 1999. A liquidity-based model of security design. *Econometrica* 67: 65–99.

Diamond, D., and R. Verrecchia. 1981. Information aggregation in a noisy rational expectations economy. *Journal of Financial Economics* 9: 221–235.

Dow, J., and G. Gorton. 1994. Arbitrage chains. *Journal of Finance* 49: 819–850.

Dow, J., and G. Gorton. 1995. Profitable informed trading in a simple general equilibrium model of asset pricing. *Journal of Economic Theory* 67: 327–369.

Dow, J., and G. Gorton. 1997. Noise trading, delegated portfolio management, and economic welfare. *Journal of Political Economy* 105: 1024–1050.

Dow, J., and R. Rahi. 2000. Should speculators be taxed? *Journal of Business* 73: 89–107.

Dow, J., and R. Rahi. 2003. Informed trading, investment, and economic welfare. *Journal of Business* 76: 430–454.

Easley, D., and M. O'Hara. 2003. Microstructure and asset pricing. In *Handbook of the economics of finance: Financial markets and asset pricing*, ed. G. Constantinides, M. Harris, and R. Stulz. Amsterdam: North-Holland.

Friedman, M. 1953. The case for flexible exchange rates. In *Essays in positive economics*. Chicago: University of Chicago Press.

Garman, M. 1976. Market microstructure. *Journal of Financial Economics* 3: 257–275.

Glosten, L., and P. Milgrom. 1985. Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics* 14: 71–100.

Green, J. 1977. The non-existence of informational equilibria. *Review of Economic Studies* 44: 451–463.

Grossman, S. 1976. On the efficiency of competitive stock markets where traders have diverse information. *Journal of Finance* 32: 573–585.

Grossman, S. 1977. The existence of futures markets, noisy rational expectations and informational externalities. *Review of Economic Studies* 44: 431–449.

Grossman, S. 1978. Further results on the informational efficiency of competitive stock markets. *Journal of Economic Theory* 18: 81–101.

Grossman, S. 1981. An introduction to the theory of rational expectations under asymmetric information. *Review of Economic Studies* 48: 541–559.

Grossman, S., and J. Stiglitz. 1980. On the impossibility of informationally efficient markets. *American Economic Review* 70: 393–408.

Hellwig, M. 1980. On the aggregation of information in competitive markets. *Journal of Economic Theory* 22: 477–498.

Jordan, J. 1983. On the efficient market hypothesis. *Econometrica* 51: 1325–1343.

Kyle, A. 1985. Continuous auctions and insider trading. *Econometrica* 53: 1315–1335.

Lucas, R. 1972. Expectations and the neutrality of money. *Journal of Economic Theory* 4: 103–124.

Milgrom, P., and N. Stokey. 1982. Information, trade and common knowledge. *Journal of Economic Theory* 26: 17–27.

Radner, R. 1979. Rational expectations equilibrium: Existence and information revealed by prices. *Econometrica* 47: 370–391.

Shiller, R. 1984. Stock prices and social dynamics. *Brookings Papers on Economic Activity* 1984(2): 457–498.

Shleifer, A., and L. Summers. 1990. The noise trader approach to finance. *Journal of Economic Perspectives* 4: 19–33.

Shleifer, A., and R. Vishny. 1997. The limits of arbitrage. *Journal of Finance* 51: 35–55.

Nominal Exchange Rates

Richard T. Baillie

Abstract

The nominal exchange rate is the rate at which the currency of one country can be exchanged for that of another. The overall value of a currency can be summarized through the 'effective nominal exchange rate', which is a weighted average of a country's nominal bilateral exchange rates. Following the advent of freely floating exchange rates in 1973 there has been intense research on understanding the mechanisms of nominal exchange rate determination and the search for an adequate model, but no model has so far withstood rigorous empirical tests.

Keywords

Bretton Woods system; Bubbles; Cointegration; Currency crises; Euro; European Monetary System; Exchange rate determination; Exchange rate target zones; Exchange rate volatility; Fixed exchange rates; Floating exchange rates; Forward premium anomaly; GARCH models; Law of one price; Monetary transmission mechanism; Money supply; Nominal exchange rate determination; Nominal exchange rates; Nominal interest rates; Open market operations; Overshooting; Peso problem; Purchasing power parity; Rational expectations; Real balances; Real exchange rates; Real interest differential; Realized volatility; Risk premium; Specification problems in econometrics; Spot exchange rates; Sterilized intervention; Transversality condition; Uncovered interest parity

JEL Classifications

F31

The nominal exchange rate is the price at which the money of one country can be exchanged for another. Usually, nominal exchange rates are bilateral, which means they denote the number of units of one country in terms of one unit of another; for example, two US dollars to one UK pound; or 0.50 UK pounds to one US dollar. Bilateral exchange rates can be expressed either in terms of spot rates, which are prices for immediate delivery, or in terms of forward contracts for delivery in the future. Some foreign exchange (FX) markets also trade currency options and futures. The worldwide FX market transacted approximately $1,700bn a day in 2006, making it by far the largest financial market. On most weeks it operates for 156 of the 168 hours available, with New York, London and Tokyo being considered as the most important and heavily traded markets. Surveys of FX market participants generally suggests that 98 per cent or more of currency transactions are motivated by speculation, arbitrage and international capital movements, rather than for the purposes of importing or exporting goods.

The overall value of a particular currency can be summarized through the ‘effective nominal exchange rate’, which is a weighted average of a country’s nominal bilateral exchange rates. A number of international financial institutions regularly report these effective rates, with different weights being used dependent on which criteria – for example, the patterns of trade – are being emphasized.

A real exchange rate is the nominal bilateral exchange rate divided by the ratio of the price indices for the two countries. Usually consumer price indices (CPIs) are used for this purpose, although trade weighted price indices are also sometimes used.

Historical Perspective

Following the end of the Second World War in 1945, a conference in Bretton Woods, New Hampshire established a system of fixed exchange rates based on the US dollar, with the US dollar in turn being convertible to gold at a fixed gold standard. However, continuing trade imbalances and apparent exchange rate misalignments led to a collapse of the Bretton Woods fixed exchange rate system in March 1973. Since then the international monetary system has generally followed what is best characterized as a managed or ‘dirty floating’ regime, with governments and/or central banks occasionally intervening to attempt to influence the value of the currencies and volatility of the market. Until the early 1990s the nominal rates between the three major regions of North America, Western Europe and Japan were formally freely floating. However, many bilateral rates were pegged under various arrangements. In particular, the European Monetary System (EMS) allowed individual countries currencies to move in a narrow band, named the ‘snake’ around par rates for each member country’s bilateral rate vis-à-vis the German Deutschmark. After several periods of apparent instability, such as the autumn of 1992 when the UK pound exited the EMS, and also the autumn of 1993 when the bands were widened to plus and minus 15 per cent of par rate; the new euro currency was introduced in 1999. Originally ten member countries of the EMS surrendered their sovereign currencies to form the euro area.

The other major development, converse to the formation of the euro, has been the collapse of communism in the late 1980s and the early 1990s, which has led many of the previously fixed exchange rates of eastern Europe and Asia to become floating rates.

As of 2007 the currencies of the US dollar, Japanese yen, euro, British pound, Swiss franc and Canadian dollar are the most actively traded, freely floating currencies.

Empirical Behaviour

To a large extent nominal exchange values and returns behave in similar manner to other asset prices. On denoting the spot exchange rate at time t as S_t, then $\Delta s_t = \Delta \ln(S_t)$ is the approximately continuously compounded rate of return. Many empirical studies have found that the hypothesis of a unit root in $\ln(S_t)$ cannot be rejected, so that returns appear to be stationary. Furthermore, returns generally appear to be approximately serially uncorrelated, so that the returns appear to be close to a martingale difference sequence, which is consistent with the theory of weak form efficiency. This has led to the one of the most striking empirical properties of high frequency, daily, weekly, or even monthly nominal exchange rate returns, concerning their apparent lack of predictability in their conditional mean. Numerous studies such as Meese and Rogoff (1983), using forward rates, surveys of market participant's expectations, and nonlinear time series models have been unable in the MSE sense to improve on random walk predictions of the nominal exchange rate.

However, the unconditional distribution of short-term nominal spot exchange rate returns is non-Gaussian and has substantial excess kurtosis; that is, they are leptokurtic. Also, returns generally exhibit time-dependent volatility, which can be well represented by various types of generalized autoregressive conditionally heteroskedastic (GARCH) models. These models represent the autocorrelated nature of volatility, which is generally considered to be due to arrival of news and to the patterns of trading volume. See Baillie and Bollerslev (1989), who estimate and discuss these models for different levels of temporal aggregation. The degree of non-Gaussianity and the level of persistence of the volatility in GARCH models are particularly high for daily returns and decreases for lower frequencies of returns. Andersen and Bollerslev (1997a, b) have used high-frequency data to examine returns and the volatility process of nominal spot exchange rate returns. They find particular stylized patterns of worldwide FX market volatility which characterizes the volatility process for each spot returns series. Andersen et al. (2003) consider the concept of realized volatility, which is an observable measure of (daily) volatility obtained from aggregating information on high-frequency returns within the day. For example, the sum of squared high-frequency returns is often used to measure daily realized volatility. The daily realized volatility is generally found to be almost pure fractional white noise, with the long-memory parameter generally being in the range of 0.30–0.40.

Purchasing Power Parity

The theory of purchasing power parity (PPP) is sometimes known as the law of one price and is to be found in the work of Ricardo in the 18th century and by Cassel in the 1920s. If S_t denotes the spot exchange rate, measured in terms of the dollar–yen rate at time t, P_t is the domestic US price level and P_t^* is the foreign country (Japan's) price level, then continuous PPP requires $P_t = S_t$ P_t^*. The real exchange rate is defined as Q_t where $Q_t = \left(S_t\, P_t^*\right)/P_t$ and, if PPP held continuously, the real exchange rate would be constant over time. In general, empirical real exchange rates since 1973 have been found to exhibit highly persistent autocorrelation and may possibly be non-stationary (see Abauf and Jorion, 1990). An important area of research in international finance has been to understand the duration of the effect of shocks to the real exchange rate, and the evidence for whether the real exchange rate returns to equilibrium in 'finite' time and restores PPP. More empirical work has re-established PPP holding in the long run, but with significant deviations (see Frankel and Rose 1996).

Uncovered Interest Rate Parity

On denoting domestic interest rates as i_t and foreign rates as i_t^* it is known that covered interest rate parity holds exactly apart from very small transaction costs and brokerage fees so that $(1+i_t)S_t = \left(1+i_t^*\right)F_t$, where F_t is the forward exchange rate. This relationship implies that the forward premium is equivalent to the interest rate differential. An important extension is the theory of uncovered interest rate parity (UIP), where $(1+i_t) = \left(1+i_t^*\right)E_t(S_{t+1}/s_t)$, which implies that the interest rate differential is approximately the expected rate of appreciation (depreciation) of a currency. Hence,

$$E_t \Delta S_{t+1} \approx \left(i_t - i_t^*\right),$$

where E_t represents the expectation operator conditioned on a sigma field of information available at time t. Hence the country with the higher rate of interest is expected to have the currency depreciation. The UIP hypothesis requires the joint assumptions of rational expectations, risk neutrality, free capital mobility and the absence of taxes on capital transfers. The theory can be derived from the solution of an Euler equation where expected real returns in the forward market are hypothesized to be zero.

Models of Exchange Rate Determination

Following the advent of freely floating exchange rates in 1973 there has been intense research on understanding the mechanisms of nominal exchange rate determination and the search for an adequate model.

Earlier work by Mundell (1963) and Fleming (1962) emphasized a Keynesian approach and considered the relative advantages of fixed versus floating nominal exchange rates. In particular, monetary policy was shown to be ineffective as a policy tool under a fixed exchange rate, while fiscal policy is effective. Conversely, monetary policy was shown to be effective under a flexible exchange rate, and fiscal policy to be ineffective under flexible exchange rates. The dominant modern paradigm is the asset market approach, which implies that the nominal exchange rate is the value of one country's money supply against another. The simplest version of the monetary model assumes PPP to hold continuously and for the existence of stable and static demand for real balances for one and possibly both countries. If the demand for real balances in the United States is $m_t - p_t = \varphi y_t - \alpha i_t$ where the lower case letters m_t, p_t and y_t represent the natural logarithms of money, prices and income respectively, and where i_t is the level of nominal interest rates; where φ is the elasticity of the demand for real balances with respect to income and α is the semi-elasticity with respect to the nominal rate of interest. The combination of PPP, uncovered interest rate parity and the demand for real balances equation is sufficient to generate a first-order rational expectations equation of the form

$$S_t = \left(\frac{1}{1+\alpha}\right) Z_t + \left(\frac{\alpha}{1+\alpha}\right) E_t S_{t+1}$$

where z_t are the fundamentals, and in this case are $z_t = \left(m_t - \varphi y_t - p_t^* + \alpha i_t^*\right)$, and asterisks denote foreign equivalents. On assuming the transversality condition which eliminates bubbles, the forward looking solution is

$$S_t = \left(\frac{1}{1+\alpha}\right) \times \sum_{j=0}^{\infty} \left(\frac{\alpha}{1+\alpha}\right) E_t \left(m_{t+j} - \varphi y_{t+j} - p_{t+j}^* + \alpha i_{t+j}^*\right).$$

A more intuitive solution is to further assume the same demand for real balances equation for the foreign country, in which case the solution is

$$S_t = \left(\frac{1}{1+\alpha}\right) \times \sum_{\mathrm{j}=0}^{\infty} \left(\frac{\alpha}{1+\alpha}\right) E_t \left[\left(m_{i+j} - m_{i+j}^*\right) - \varphi \left(y_{t+j} - y_{t+j}^*\right)\right]$$

Similarly to the Keynesian approach, the monetary model implies an equivalent depreciation of the exchange rate with respect to an increase in US money supply and prices. However, the model also implies dollar appreciation following an increase in US incomes, and a dollar depreciation following an increase in US nominal interest rates; both implications are contrary to the Keynesian approach.

The empirical realization that nominal exchange rates did not move in perfect synchronization with relative prices and money supplies generated attempts to loosen the constraints of the model. Frankel (1979) introduced the real interest differential (RID) model, while the celebrated concept of overshooting was due to Dornbusch (1976). Under rational expectations, the overshooting model is very similar to an alternative model of Woo (1985), which assumes flexible prices but dynamic adjustments in the demand for real balances. The solution paths for both models are obtained from the forward solution of a second-order forward-looking rational expectations equation.

It has been hard to find rigorous empirical support for any of the models. While the log of the exchange rate and most of the macroeconomic fundamentals appear to be well approximated by integrated processes, there has been an absence of cointegration. This rejects the long-run properties of the basic monetary model as well as the Dornbusch overshooting formulation. In fact, the macro fundamentals are found to add little explanatory power to the model. This again is consistent with the findings of Meese and Rogoff (1983), who found that most models and forecasting methods were inferior, in the sense of *ex ante* MSE forecasting comparisons, to a simple random walk model. Mark (1995) and Mark and Choi (1997) have found some evidence that fundamentals have increased explanatory power when predicting exchange rates a year or more ahead. An alternative approach considering the possibility of nonlinear adjustment to equilibrium has been advocated by Engel and Hamilton (1990) and Taylor and Peel (2000), and is likely to remain an active area of research for the forseeable future.

High-frequency analyses by Anderson and Bollerslev (1998) and Andersen et al. (2003) have examined the role of macro news announcements on exchange rate returns. Some explanatory power has been detected, but not as much as would be suggested by the macro models. These findings tend to support perceived wisdom in the FX market concerning the fact that traders react less to macroeconomic news than previously expected.

Forward Premium Anomaly

The forward premium or forward discount anomaly refers to the widespread result that the returns on freely floating exchange rates are invariably negatively correlated with the lagged forward premium. One of the most widespread tests of uncovered interest rate parity is based on the regression of future spot returns on the lagged forward premium, or equivalently the lagged interest rate differential,

$$\varDelta s_{t+1} = \alpha + \beta(f_t - s_t) + \varepsilon_{t+1},$$

where ε_{t+1} is the regression disturbance. While the theory of uncovered interest rate parity would suggest that $\alpha = 0$, $\beta = 1$ and ε_{t+1} uncorrelated, a substantial body of empirical work has found the estimate of the slope coefficient β to be negative. Interestingly, this result is found for different currencies, different numeraire currencies and over different sample periods, including the 1920s. As discussed by Baillie and Bollerslev (2000), the estimated β coefficient is time varying and can be as low as -13 for periods within the 1980s. Possible explanations of the forward premium anomaly have included ‘peso problem’ effects, the role of learning and heterogeneous beliefs on the part of agents; while the most dominant explanation has been in terms of the presence of a time-dependent risk premium, ρ_{t+1} which is defined as

$$E_t \Delta s_{t+1} = (f_t - s_t) - \rho_{t+1}.$$

Fama (1984) has shown that a $\hat{\beta} < 0$ implies that $Cov(E_t\,\Delta s_{t+1}\rho_{t+1}) < 0$, so that the expected rate of appreciation is negatively correlated with the risk premium, and also $Var(\rho_{t+1}) > Var(E_t\ \Delta s_{t+1})$, so that the variability of the risk premium must exceed that of the expected rate of appreciation. Models of the time-dependent risk premium are generally motivated by versions of the Lucas–Breeden asset pricing approach (see Lucas 1978). Following Domowitz and Hakkio (1985) many parametric models for the risk premium have been formulated from micro theoretic models. See Hodrick (1989) for one of the most detailed formulations, which is discussed in detail by Engel (1996). These models generally represent the risk premium in terms of the second conditional moments of fundamentals, and there has been little definitive empirical support for these models. However, Baillie and Kilic (2006) have found evidence for nonlinear smooth transition regime adjustment to uncovered interest rate parity with threshold variables, such as the conditional variability of US money growth, and the interest rate differential, which are variables derived for risk premium from theoretical models.

Other authors have noted that the problem of econometric specification with uncorrelated returns being regressed on the forward premium or interest rate differential appears to have very persistent, or ‘long-memory’ autocorrelation. Baillie and Bollerslev (2000) and Maynard and Phillips (2001) discuss some of the specification issues that result.

Target Regimes and Intervention

There has been considerable research on the implementation of target zones for nominal exchange rates. In particular, Krugman (1991) has considered the differential equations behind monetary policy-style intervention at the bands of the target zone; while Neely (1999) documents some of the statistical properties of such returns. Complications due to intra-marginal intervention have also been considered, and the empirical success of the models is discussed by Bekaert and Gray (1998). Perhaps most work in this area has been done on trying to understand the transmission mechanism of sterilized intervention, where open market operations by a central bank are designed to maintain levels of money supply following their purchase (sale) of domestic currency. Such intervention is generally officially motivated as an attempt to either move a nominal exchange rate closer to a target level, and/or to reduce FX market volatility. The empirical results are controversial with relatively small effects being detected, although Baillie and Osterberg (1997) use an extension of Hodrick (1989) to motivate intervention affecting the risk premium, and find quite strong supportive econometric evidence. The reasons for currency crises and the possibility of early warning corrective actions that may be taken to avoid crises have also attracted attention (see Kaminsky et al. 1998; Kaminsky and Schumaker 2000; Rose and Svensson 1995).

See Also

- ▶ Exchange Rate Target Zones
- ▶ Purchasing Power Parity
- ▶ Real Exchange Rates
- ▶ Uncovered Interest Parity

Bibliography

Abauf, N., and P. Jorion. 1990. Purchasing power parity in the long run. *Journal of Finance* 45: 157–174.

Andersen, T.G., and T. Bollerslev. 1997a. Intraday periodicity and volatility persistence in financial markets. *Journal of Empirical Finance* 4: 115–158.

Andersen, T.G., and T. Bollerslev. 1997b. Heterogeneous information arrivals and return volatility dynamics: Uncovering the long-run in high frequency returns. *Journal of Finance* 52: 975–1005.

Andersen, T.G., and T. Bollerslev. 1998. Deutsche mark–dollar volatility: Intraday activity, patterns, macroeconomic announcements and longer run dependence. *Journal of Finance* 53: 219–265.

Andersen, T.G., T. Bollerslev, F.X. Diebold, and F. Labys. 2003. Modeling and forecasting realized volatility. *Econometrica* 71: 579–625.

Baillie, R.T., and T. Bollerslev. 1989. The message in daily exchange rates: A conditional variance tale. *Journal of Business and Economic Statistics* 7: 297–305.

Baillie, R.T., and T. Bollerslev. 2000. The forward premium anomaly is not as bad as you think. *Journal of International Money and Finance* 19: 471–488.

Baillie, R.T., and W.P. Osterberg. 1997. Central bank intervention and risk in the forward premium. *Journal of International Economics* 43: 483–497.

Baillie, R.T., and R. Kilic. 2006. Do asymmetric and nonlinear adjustments explain the forward premium anomaly? *Journal of International Money and Finance* 25: 22–47.

Bekaert, G., and S.F. Gray. 1998. Target zones and exchange rates: An empirical investigation. *Journal of International Economics* 45: 1–35.

Domowitz, I., and C.S. Hakkio. 1985. Conditional variance and the risk premium in the foreign exchange market. *Journal of International Economics* 19: 47–66.

Dornbusch, R. 1976. Expectations and exchange rate dynamics. *Journal of Political Economy* 84: 1161–1176.

Engel, C. 1996. The forward discount anomaly and the risk premium: A survey of recent evidence. *Journal of Empirical Finance* 3: 123–192.

Engel, C.M., and J.D. Hamilton. 1990. Long swings in the dollar: Are they in the data and do the markets know it? *American Economic Review* 80: 689–713.

Fama, E.F. 1984. Spot and forward exchange rates. *Journal of Monetary Economics* 14: 319–338.

Fleming, M.J. 1962. Domestic financial policies under fixed and floating exchange rates. *IMF Staff Papers* 9: 369–379.

Frankel, J.A. 1979. On the mark: Theory of floating exchange rates based on real interest rate differentials. *American Economic Review* 69: 610–622.

Frankel, J.A., and A.K. Rose. 1996. A panel data project on purchasing power parity: Mean reversion within and between bands. *Journal of International Economics* 40: 209–224.

Hodrick, R.J. 1989. Risk, uncertainty and exchange rates. *Journal of Monetary Economics* 23: 433–459.

Kaminsky, G., S. Lizondo, and C.M. Reinhart. 1998. Leading indicators of currency crisis. *IMF Staff Papers* 45: 1–48.

Kaminsky, G.L., and S.L. Schumaker. 2000. What triggers market jitters? A chronicle of the Asian crisis. *Journal of International Money and Finance* 18: 537–560.

Krugman, P. 1991. Target zones and exchange rate dynamics. *Quarterly Journal of Economics* 106: 669–682.

Lucas, R.E. 1978. Asset prices in an exchange economy. *Econometrica* 46: 1429–1445.

Mark, N.C. 1995. Exchange rates and fundamentals: Evidence on long-horizon predictability. *American Economic Review* 85: 201–218.

Mark, N.C., and D.-Y. Choi. 1997. Real exchange rate prediction over long horizons. *Journal of International Economics* 43: 29–60.

Maynard, A., and P.C.B. Phillips. 2001. Rethinking an old empirical puzzle: Econometric evidence on the forward discount anomaly. *Journal of Applied Econometrics* 16: 671–708.

Meese, R.A., and K.R. Rogoff. 1983. Exchange rate models of the seventies: Do they fit out of sample? *Journal of International Economics* 14: 3–24.

Mundell, R.A. 1963. Capital mobility and stabilization policy under fixed and flexible exchange rates. *Canadian Journal of Economics and Political Science* 29: 475–485.

Neely, C.J. 1999. Target zones and conditional volatility: The role of realignments. *Journal of Empirical Finance* 6: 177–192.

Rose, A.K., and L.E.O. Svensson. 1995. Expected and predicted realignments: The FF/DM exchange rate during the European Monetary System 1979–1993. *Scandinavian Journal of Economics* 97: 175–200.

Taylor, M.P., and D.A. Peel. 2000. Nonlinear adjustments and long run equilibrium and exchange rate fundamentals. *Journal of International Money and Finance* 19: 33–53.

Woo, W.T. 1985. The monetary approach to exchange rate determination under rational expectations. *Journal of International Economics* 18: 1–16.

Non-clearing Markets in General Equilibrium

Jean-Pascal Bénassy

Abstract

In this article we study models with non-clearing markets in a full general equilibrium framework. The theories we describe synthesize three major schools of thought, Walrasian, Keynesian and imperfect competition. This synthesis is notably achieved by introducing quantity signals in addition to price signals into the traditional general equilibrium model. This considerably enlarges the scope of traditional general equilibrium, allowing us not only to construct equilibria with various price rigidities but also to endogenize prices in a decentralized imperfect competition framework.

Keywords

Active demand; Budget constraint; Dual decision method; Fixprice equilibria; Fixpricefixwage macroeconomic model; Frictions; General equilibrium; Imperfect competition; Keynesianism; Market efficiency; Monopolistic competition; Non-clearing markets in general equilibrium; Non-tâtonnement processes; Objective demand curve; Overbidding; Price making; Price signals; Price taking; Quantity constraints; Quantity signals; Quasi-equilibrium; Rational expectations; Rationing; Subjective demand curves; Walrasian equilibrium

JEL Classifications

D5

In this article we study how to model situations of non-clearing markets in a full general equilibrium framework. As we shall see from the historical discussion at the end, the theories we obtain synthesize three major schools of thought: (*a*) the Walrasian school, as Walras was the first to study a fully fledged general equilibrium system; (*b*) the Keynesian school, as Keynes emphasized the importance of quantity adjustments in reaching a macroeconomic equilibrium with at least one non-clearing market (that is, the labour market); and (*c*) the imperfect competition school, which endogenized prices through explicit price making by agents internal to the system.

This synthesis is notably achieved by introducing quantity signals into the traditional general equilibrium model. These quantity signals are quantity constraints which tell each agent the maximum quantity he can trade in each market. As we shall see, the introduction of these quantity signals in addition to price signals considerably enlarges the scope of traditional general equilibrium since they allow us not only to treat equilibria with various price rigidities, but also to endogenize prices in a decentralized imperfect competition framework.

The plan of the entry is the following. In the next three sections we describe the general concepts. The fourth section gives a brief historical outline of this line of thought.

Non-clearing Markets and Quantity Signals

In this section and the next two we describe various concepts in the framework of a monetary exchange economy where one good, money, serves as numéraire, medium of exchange and reserve of value (similar concepts have been developed for barter economies – see Bénassy 1975b, 1982 – but the formalization gets quite clumsy). There are ℓ markets in the period considered, where non-monetary goods indexed by $h = 1, \ldots, \ell$; 'are exchanged against money at the price p_h. We call p the vector of these prices.

Agents are indexed by $i = 1, \ldots, n$; n. In market h agent i may make a purchase $d_{ih} \geq 0$ or a sale $s_{ih} \geq 0$. Define his net transaction of good h, $z_{ih} = d_{ih} - s_{ih}$, and z_i the ℓ-dimensional vector of these net transactions.

At the beginning of the period agent i holds quantities $\overline{m}_i$ of money, and ω_{ih} of good h. Call ω_i the vector of the ω_{ih}. As a result of his trades z_i, agent i ends up with final holdings of non-monetary goods and money, x_i and m_i, given respectively by:

$$x_i = \omega_i + z_i \; m_i = \overline{m}_i - pz_i$$

We assume that agent i has a utility function on these final holdings $U_i(x_i, m_i) = U_i(w_i + z_i, m_i)$, which we assume throughout strictly concave in its arguments.

Walrasian Equilibrium

In order to contrast it with the non-Walrasian equilibrium concepts that will follow, let us describe briefly the Walrasian equilibrium of this economy (Arrow and Debreu 1954; Debreu 1959). Each agent i receives (from the implicit auctioneer) a price signal p. As a response he expresses a Walrasian net demand given by the function $z_i(p)$, solution in z_i of the following program:

$$\text{Maximize} \quad U_i(\omega_i + z_i, m_i)\text{s.t.}$$
$$pz_i + m_i = \overline{m}_i$$

A Walrasian equilibrium price vector p^* is defined by the condition that all markets clear, that is:

$$\sum_{i=1}^{n} z_i\left(p^*\right) = 0$$

The vector of transactions realized by each agent i is $z_i(p^*)$.

Demands and Transactions

As we will be studying non-clearing markets, we must now make an important distinction, that between demands and supplies on the one hand, and the resulting transactions on the other.

Transactions, that is, purchases or sales of goods, denoted d_{ih}^* and s_{ih}^*, are exchanges actually made, and must thus identically balance on each market, that is:

$$D_h^* = \sum_{i=1}^{n} d_{ih}^* = \sum_{i=1}^{n} s_{ih}^* = S_h^* \quad \text{for all } h \qquad (1)$$

On the other hand, demands and supplies, denoted $\tilde{d}_{ih}$ and $\tilde{s}_{ih}$, are signals transmitted to the market (that is, to the other agents) before exchange takes place. They represent as a first approximation the exchanges the agents wish to make on each market. So they do not necessarily match in a specific market, and no identity like (1) applies to them:

$$\tilde{D}_h = \sum_{i=1}^{n} \tilde{d}_{ih} \neq \sum_{i=1}^{n} \tilde{s}_{ih} = \tilde{S}_h$$

In order to shorten notation, we often work in what follows with net demands and net transactions defined respectively by:

$$\tilde{z}_{ih} = \tilde{d}_{ih} - \tilde{s}_{ih} \;\; z^*_{ih} = d^*_{ih} - s^*_{ih}$$

The equality of aggregate purchases and sales Eq. (1) is then rewritten:

$$\sum_{i=1}^{n} z^*_{ih} = 0 \;\; \text{for all } h \tag{2}$$

Rationing Schemes

In each market h the exchange process must generate consistent transactions (that is, transactions satisfying Eq. (1) or (2)) from any set of possibly inconsistent demands and supplies. Some rationing will necessarily occur, which may take various forms, such as uniform rationing, queuing, priority systems, proportional rationing, and so forth . . . depending on the particular organization of each market. We call *rationing scheme* the mathematicalrepresentation of each specific organization. To be more precise, the rationing scheme in market h is defined by a set of n functions:

$$z^*_{ih} = F_{ih}(\tilde{z}_{1h}, \ldots, \tilde{z}_{nh}) i = 1, \ldots, n \tag{3}$$

such that:

$$\sum_{i=1}^{n} F_{ih}(\tilde{z}_{1h}, \ldots, \tilde{z}_{nh}) = 0 \;\; \text{for all } \; \tilde{z}_{1h}, \ldots, \tilde{z}_{nh}$$

We assume that F_{ih} is continuous, non-decreasing in $\tilde{z}_{ih}$ and non-increasing in the other arguments. Before examining the possible properties of these rationing schemes, let us take a most simple example with two agents. Agent 1 emits a demand $\tilde{d}_{1h}$, agent 2 a supply $\tilde{s}_{2h}$. Then a natural rationing scheme, implicit in most macroeconomic models, is to take the level of transactions as equal to the minimum of demand and supply, that is:

$$d^*_{1h} = s^*_{2h} = \min\left(\tilde{d}_{1h}, \tilde{s}_{2h}\right) \tag{4}$$

Properties of Rationing Schemes

We first study two possible properties that a rationing scheme may satisfy: voluntary exchange and market efficiency.

The first property is actually an extremely natural one in a free market economy: We shall say that there is *voluntary exchange* in market h if no agent can be forced to purchase more than he demands, or to sell more than he supplies, which is expressed by:

$$d^*_{ih} \le \tilde{d}_{ih} s^*_{ih} \le \tilde{s}_{ih} \;\; \text{for all } \; i$$

or equivalently in algebraic terms:

$$|z^*_{ih}| < |\tilde{z}_{ih}| \tilde{z}_{ih} \cdot \tilde{z}_{ih} \ge 0 \;\; \text{for all } i.$$

Most markets in reality meet this condition, and we henceforth assume that it always holds. This allows us to classify agents in a market h in two categories: unrationed agents for which $z^*_{ih} = \tilde{z}_{ih}$, and rationed agents who trade less than they wanted.

The second property we study here is that of market efficiency, or absence of frictions, which corresponds to the idea of exhaustion of all mutually advantageous exchanges: a rationing scheme on a market h is *efficient*, or *frictionless*, if one cannot find simultaneously a rationed demander and a rationed supplier in market h. The intuitive idea behind this is that in an efficiently organized market a rationed buyer and a rationed seller would meet and exchange until one of the two is not rationed. Together with voluntary exchange, it implies the 'short-side rule', according to which agents on the 'short side' of the market can realize their desired transactions:

$$\tilde{D}_{\hbar} \ge \tilde{S}_h \Rightarrow s^*_{ih} \; \text{for all } i \tilde{S}_h \ge \tilde{D}_h \Rightarrow d^*_{ih} - \tilde{d}_{ih} \; \text{for all } \; i$$

This rule also implies that the global level of transactions on a market h will be equal to the minimum of aggregate demand and supply:

$$D^*_h = S^*_h = \min\left(\tilde{D}_h, \tilde{S}_h\right)$$

We should note that the market efficiency assumption may not always hold, notably if one considers a fairly wide and decentralized market, because some demanders and suppliers might not meet pairwise. In particular, the market efficiency property is usually lost by aggregation of sub-markets, whereas the voluntary exchange property remains intact in the aggregation process. So we must keep in mind that it does not always hold. Fortunately, this hypothesis is not necessary for most of the microeconomic concepts presented in the next sections.

Quantity Signals

Now it is clear that at least rationed agents must perceive a quantity constraint in addition to the price signal. As it turns out, these quantity signals appear quite naturally in the formulation of a number of rationing schemes called *non-manipulable*, which can be written under the form:

$$d_{ih}^{*} = \min\left(\tilde{d}_{ih}, \tilde{d}_{ih}\right) s_{ih}^{*} = \min\left(\tilde{s}_{ih}, \tilde{s}_{ih}\right) \qquad (5)$$

where the quantity signals $\overline{d}_{ih}$ and $\overline{s}_{ih}$ are functions only of the demands and supplies of the other agents. As an example, we can note that the rationing scheme corresponding to Eq. (4) above is of this type with:

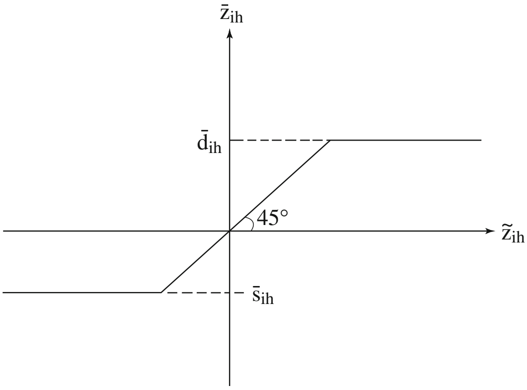
$$\overline{d}_{1h} = \tilde{s}_{2h} \overline{s}_{2h} = \tilde{d}_{1h}$$

For non-manipulable schemes the relation between z_{ih}^{*} and $\tilde{z}_{ih}$ looks as in Fig. 1, in which we see where the term ‘non-manipulable’ comes from: once rationed, the agent cannot increase, or ‘manipulate’, the level of his transactions by increasing his demand and supply.

To make things a little more precise, let us rewrite the rationing scheme in market h Eq. (3) under the form:

$$z_{ih}^{*} = F_{ih}\left(\tilde{z}_{ih}, \tilde{z}_{-ih}\right) \qquad (6)$$

where $\tilde{z}_{-ih}$ is the set of all net demands on market h, except that of agent i, that is, $\tilde{z}_{-ih} = \left\{\tilde{z}_{jh} | j \neq i\right\}$. The rationing scheme is non-manipulable if it can be rewritten as in Eq. (5), or algebraically:



Non-clearing Markets in General Equilibrium, Fig. 1

$$F_{ih}\left(\tilde{z}_{ih}, \tilde{z}_{-ih}\right) = \begin{cases} \min\left(\tilde{z}_{ih}, \tilde{d}_{ih}\right) \tilde{z}_{ih} \ \geq 0 \\ \max\left(\tilde{z}_{ih}, \ -\overline{s}_{ih}\right) \quad \tilde{z}_{ih} \geq 0 \end{cases}$$

where $\overline{d}_{ih}$ and $\tilde{s}_{ih}$ are functions of all demands and supplies in market h, except that of agent i, which we shall write as:

$$\tilde{d}_{ih} = G_{ih}^{d}\left(\tilde{z}_{-ih}\right) \geq 0 \tilde{s}_{ih} = G_{ih}^{s}\left(\tilde{z}_{-ih}\right) \geq 0 \qquad (7)$$

Note that the functions $G_{ih}^{d}\left(\tilde{z}_{-ih}\right)$ and $G_{ih}^{s}\left(\tilde{z}_{-ih}\right)$ are not arbitrary, but are related to the rationing scheme F_{ih} through:

$$G_{ih}^{d}\left(\tilde{z}_{-ih}\right) = \max\left\{\tilde{z}_{ih} | F_{ih}\left(\tilde{z}_{ih}, \tilde{z}_{-ih}\right) = \tilde{z}_{ih}\right\} \qquad (8)$$

$$G_{ih}^{s}\left(\tilde{z}_{-ih}\right) = -\min\left\{\tilde{z}_{ih} | F_{ih}\left(\tilde{z}_{ih}, \tilde{z}_{-ih}\right) = \tilde{z}_{ih}\right\} \qquad (9)$$

where it appears clearly that these quantity constraints are indeed the maximum purchase and sale that agent i can make in market h.

We may note that some rationing schemes, called *manipulable*, such as the proportional rationing scheme, cannot be written under this form. The phenomenon of manipulation through demand and supply leads then to a perverse phenomenon of overbidding, and to the non-existence of an equilibrium unless additional constraints are put on demands and supplies (Bénassy 1977b, 1982).

Most rationing schemes in the real world are actually non-manipulable through demand and supply, and we thus from now on study only such rationing schemes as can be characterized by Eq. (5) or (7). The variables $\overline{d}_{ih}$ and $\overline{s}_{ih}$ in (5) and (7) are *quantity constraints*. These are the quantity signals that each agent receives, and they play a fundamental role in both quantity and price determination, as we see in the next two sections. Before moving to the study of these problems and to the definition of non Walrasian equilibria, it is useful to rewrite Eqs. (6) and (7) pertaining to an agent i under vector form:

$$z_i^* = F_i(\tilde{z}_i, \tilde{z}_{-i})\overline{d}_i = G_i^d(\tilde{z}_{-i})\overline{s}_i = G_i^s(\tilde{z}_{-i}) \quad (10)$$

where $\tilde{z}_i$ is the vector of $\tilde{z}_{ih}, h = 1, \ldots, \ell$, and $\tilde{z}_{-i}$ is the set of all such vectors, except that of agent i himself, i.e. $\tilde{z}_{-i} = \{\tilde{z}_j | j \neq i\}$.

Fixprice Equilibria

We now study a first concept of non-Walrasian equilibrium, that of fixprice equilibrium. This concept is of interest for several reasons. First, it gives us a very large class of consistent market allocations, since we shall find that under very standard conditions a fixprice equilibrium exists for every positive price system and every set of rationing schemes (we may note that Walrasian allocations are particular fixprice allocations, specifically those corresponding to a Walrasian price vector). Second, as we see in the next section, fixprice equilibria are a very important building block in constructing other non-Walrasian equilibrium concepts with flexible prices.

We thus assume that the price system p is given. As indicated, we assume that the rationing schemes in all markets are non-manipulable. Accordingly, transactions and quantity signals are generated in all markets according to the formulas seen above Eq. (10). We immediately see that all that remains to be done in order to obtain a fixprice equilibrium concept is to determine how demands themselves are formed, a task to which we now turn.

Effective Demands and Supplies

Demands and supplies are signals that agents send to the 'market' (that is, to the other agents) in order to obtain the best transactions. Consider thus an agent i faced with a price vector p and vectors of quantity constraints, $\overline{d}_i$ and $\overline{s}_i$. He knows that his transactions will be related to his demands and supplies by formulas (5) seen above, namely,

$$d_{ih}^* = \min\left(\tilde{d}_{ih}, \overline{d}_{ih}\right) s_{ih}^* = \min(\tilde{s}_{ih}, \overline{s}_{ih})$$

Now the problem is to choose a vector of net effective demands $\tilde{z}_i$ which will lead him to the best possible transactions. As it turns out, there exists a simple and workable definition which generalizes Clower's (1965) original 'dual decision' method: the effective demand of agent i on market h is the trade which maximizes his utility subject to the budget constraints and to the quantity constraints on the *other* markets. Formally the effective demand $\tilde{z}_{ih}$ is solution in z_{ih} of the following programme:

$$\text{Maximize } U_i(\omega_i + z_i, m_i)\text{s.t.}$$
$$\begin{cases} pz_i + m_i = \overline{m}_i \\ -\overline{s}_{ik} \leq z_{ik} \leq \overline{d}_{ik} k \neq h \end{cases}$$

Because of the strict concavity of U_i, we obtain a function, denoted $\tilde{\xi}_{ih}\left(p, \tilde{d}_i, \tilde{s}_i\right)$. Repeating the operation for all markets $h = 1, \ldots \ell$, we obtain a vector function of effective demands $\tilde{\xi}_i\left(p, \overline{d}_i, \overline{s}_i\right)$. This vector of effective demands has two good properties. First, it leads to the best transactions that it is possible to attain given the price vector p and the quantity constraints $\overline{d}_i$ and $\overline{s}_i$. Second, whenever a constraint is binding on a market h, the corresponding demand or supply is greater than the quantity constraint, which thus 'signals' to the market that the agent trades less than he would want. Such signals are useful to avoid trivial equilibria where no one would trade because nobody else signals that he wants to trade.

Fixprice Equilibrium

With the above definition of effective demand, we are now ready to give a first definition of a fixprice equilibrium, found in Bénassy (1975a, 1982).

Definition 1 *A fixprice equilibrium associated with a price system p and rationing schemes represented by functions F_i , $i = 1, \ldots, n$, is a set of effective demands $\tilde{z}_i$, transactions z_i^* and quantity constraints $\overline{d}_i$ and $\overline{s}_i$ such that:*

(a) $\tilde{z}_i = \widetilde{\xi}_i\left(p, \overline{d}_i, \overline{s}_i\right) \; i = 1, \ldots, n$

(b) $z_i^* = F_i(\tilde{z}_i, \tilde{z}_{-i}) i = 1, \ldots, n$

(c) $\overline{d}_i = G_i^d(\tilde{z}_{-i}) \overline{s}_i = G_i^s(\tilde{z}_{-i}) i = 1, \ldots, n$

Equilibria defined in this way exist for all positive prices and all rationing schemes satisfying voluntary exchange and non-manipulability (Bénassy 1975a, 1982). The 'exogenous' data are the price system p and the rationing schemes $F_i, i = 1,\ldots n$. One may wonder whether for given such exogenous data the equilibrium is likely to be unique. A positive answer has been given by Schulz (1983), who showed that the equilibrium is globally unique, provided the 'spillover' effects (there is a spillover effect when a binding constraint in one market modifies the effective demand in another market) are less than 100 per cent in value terms. For example in the simplest Keynesian model this would amount to a propensity to consume strictly smaller than 1.

In what follows we assume that the Schulz conditions hold, and denote by $\tilde{Z}_i(p), Z_i^*(p), \overline{D}_i(p)$ and $\overline{S}_i(p)$ the functions giving the values of $\tilde{z}_i, z_i^*, \overline{d}_i$, and $\overline{s}_i$ at a fixprice equilibrium corresponding to p (the market organization, and thus the rationing schemes, being assumed invariant).

An Alternative Concept

We shall now present an alternative concept of fixprice equilibrium, due to Drèze (1975) (who actually dealt with the more general case of prices variable between fixed limits), and which we shall recast using our notations. That concept does not separate demands from transactions, and thus considers directly the vectors of transactions z_i^* and quantity constraints $\overline{d}_i$ and $\overline{s}_i$. The original concept actually assumed uniform rationing, so that the vectors di and si were the same for all agents.

Definition 2 *A fixprice equilibrium for a given set of prices p is defined as a set of transactions* z_i^* and quantity constraints $\overline{d}_i$ and $\overline{s}_i$ such that:

(a) $\sum_{i=1}^n z_{ih}^* = 0 \, \forall \, h$

(b) *The vector* z_i^* *is solution in* z_i *of:*

$$\text{Maximize } U_i(\omega_i + z_i, m_i)\text{s.t.}$$
$$\begin{cases} pz_i + m_i = \overline{m}_i \\ -\overline{s}_{ih} \leq z_{ih} \leq \overline{d}_{ih} \forall \, h \end{cases}$$

(c) $\forall h \, z_h^* = \overline{d}_{ih}$ *for some i implies* $z_{jh}^* > -\overline{s}_{jh}$

$\forall \, j z_{ih}^* = -\overline{s}_{ih}$ *for some i implies* $z_{jh}^* < -d_{jh} \forall j$

Let us now interpret these conditions. Condition (a) is the natural requirement that transactions should balance in each market. Condition (b) says that transactions must be individually rational, that is, they must maximize utility subject to the budget constraint and the quantity constraints on all markets. We may note at this stage that using quantity constraints under the form of upper and lower bounds on trades implicitly assumes rationing schemes which exhibit voluntary exchange and non-manipulability, as we saw when studying rationing schemes. Condition (c) says that rationing may affect either supply or demand, but not both simultaneously. We recognize here with a different formalization the condition of market efficiency which is thus built into this definition of equilibrium, whereas it is not in the previous definition.

Drèze (1975) proved that an equilibrium according to def 2 exists for all positive price systems and for uniform rationing schemes under the standard concavity assumptions for the utility functions. The concept is easily extended to non-uniform bounds (Grandmont and Laroque 1976; Greenberg and Muller 1979), but in this last case it is not specified in the concept how shortages are allocated. Because of this there will be usually an infinity of equilibria corresponding to a given price vector, as soon as there are two rationed agents on one side of a market.

As we noted above, the two concepts of fixprice equilibrium we described in this section are based, implicitly or explicitly, on a representation of markets under the form of rationing schemes satisfying voluntary exchange and non-manipulability. This suggests that, if in the first definition we further assume that all rationing schemes are efficient or frictionless, the two definitions should yield similar sets of equilibrium allocations for a given price system. This was indeed proved by Silvestre (1982, 1983) for both exchange and production economies. The relation between the two concepts has been further explored by D'Autume (1985).

Price Making and Equilibrium

As this stage we still need a description of price making by agents internal to the system. We describe in this section a concept dealing with that problem and we shall see that, just as in demand and supply theory, quantity signals play a prominent role. It is indeed quite intuitive that quantity constraints must be a fundamental part of the competitive process in a decentralized economy: it is the inability to sell as much that they want which leads suppliers to propose, or accept from other agents, a lower price, and conversely it is the inability to purchase as much as they want that leads demanders to propose, or accept, a higher price.

Various modes of price making integrating these aspects can be envisioned. We deal here with a realistic organization of the pricing process where agents on one side of the market (most often the suppliers) quote prices and agents on the other side act as price takers. The general idea relating the concepts in this section to those of the previous one is that price makers change their prices so as to 'manipulate' the quantity constraints they face (that is, so as to increase or decrease their possible sales or purchases). As we shall see, this model of price making is quite reminiscent of the imperfect competition line (Chamberlin 1933; Robinson 1933; Triffin 1940; Bushaw and Clower 1957; Arrow 1959), and more particularly of the theories of general equilibrium with monopolistic competition, as developed notably by Negishi (1961).

The Framework

We thus now assume that agent i controls the prices of a (possibly empty) subset H_i of goods. Goods are distinguished both by their physical characteristics and by the agent who sets their price. We thus consider two goods sold by different sellers as different goods, a fairly natural assumption since these goods differ at least by location, quality, and so on, so that:

$$H_i \cap H_j = \{\varnothing\} i \neq j$$

We denote by p_i the set of prices controlled by agent i and p_{-i} the rest of prices, that is:

$$p_i = \{p_h | h \in H_i\} p_{-i} = \{p_h | h \notin H_i\}$$

Each agent chooses his price vector p_i taking the other prices p_{-i} as given. The equilibrium structure is thus that of a Nash equilibrium in prices, corresponding to an idea close to that of monopolistic competition. The basic idea behind the modelling of price making itself in such models is, as we indicated above, that each price maker uses the prices he controls to 'manipulate' the quantity constraints he faces. Consider the markets whose price are determined by agent i, and subdivide further H_i into H_i^d (goods demanded by i) and H_i^s (goods supplied by i). We may note in passing that, although agent i appears formally as a monopolist in markets $h \in H_i^s$ or a monopsonist in markets $h \in H_i^d$, his actual 'monopoly power' may be very low due to the fact that other agents sell or buy products which are extremely close substitutes to those he controls. Because the price makers are alone on their side of the markets where they set prices, their quantity constraints on these markets have the simple form:

$$\bar{s}_{ih} = \sum_{j \neq i} \tilde{d}_{jh} \quad h \in H_i^s \qquad \bar{d}_{ih} = \sum_{j \neq i} \tilde{s}_{jh} \quad h \in H_i^d$$

that is, the maximum quantity that price setter i can sell is the total demand of the others, and conversely if he is a buyer. All we need to know, in order to pose the problem of price setting as a standard decision problem, is the relation, as perceived by the price maker, between the quantity constraints he faces and the prices he sets. Several approaches allow us to treat this problem and to link it with the concepts seen previously. The first, based on Negishi's (1961) subjective demand curve approach, was developed in Bénassy (1976, 1982). The second is an objective demand curve approach, developed in Bénassy (1987, 1988), and which we shall now briefly describe.

Objective Demand Curves

The implicit idea behind the objective demand curve approach (Gabszewicz and Vial 1972; Marschak and Selten 1974; Nikaido 1975) is that each price maker knows the economy well enough to be able to compute under all circumstances the actual quantity constraints he will face. Since we are considering a Nash equilibrium, he must be able to perform this computation for any set pi of prices he chooses as well as for any set $p-\ i$ of the other prices; that is, he must be able to compute his constraints for any vector of prices, once all feedback effects have been accounted for.

But we know from the previous section that, for a given organization of the economy (that is, notably for given rationing schemes), and for a given set of prices p, the quantity constraints agent i faces are given by the functions $\overline{D}_i(p)$ and $\overline{S}_i(p)$. If the agent has full knowledge of the parameters of the economy (a strong assumption, of course, but which is embedded in the notion of an objective demand curve), then he knows this and the objective demand and supply curves will be respectively given by the functions $\overline{S}_i(p)$ and $\overline{D}_i$ (p) We may note that the objective demand curve $\overline{S}_i(p)$ is denoted as a constraint on agent i's supply, which is natural since the sum of all other agents' demands acts as a constraint on the sales of agent i, and symmetrically with the objective supply curve $\overline{D}_i(p)$.

Price Making and Equilibrium

If agent i knows the two vector functions $\overline{D}_i(p)$ and $\overline{S}_i(p)$, the programme giving his optimal price p_i is the following:

$$\text{Maximize}\, U_i(\omega_i + z_i, m_i)\text{s.t.}\\ \begin{cases} pz_{\underline{i}} + m_i = \overline{m}_i \\ -\overline{S}_i(p) \leq z_i \leq \overline{D}_i(p) \end{cases}.$$

which yields the optimum price p_i chosen by agent i as a function of the other prices p_{-i}.

$$p_i = \psi_i(p_{-i})$$

This naturally leads us to the definition of an equilibrium with price makers:

Definition 3 *An equilibrium with price makers is characterized by a set of prices p_i^*, net demands $\tilde{z}_i$, transactions z_i^* and quantity constraints $\overline{d}_i$ and $\overline{s}_i$ such that*:

(a) $p_i^* = \psi_i(p_{-i}^*)$

(b) $\tilde{z}_i, z_i^*, \tilde{d}_i, \overline{s}_i$ are equal respectively to $\tilde{Z}_i(p^*), Z_i^*(p^*), \overline{D}_i(p^*), \overline{S}_i(p^*)$.

Condition (a) indicates that we have a Nash equilibrium in prices, given each agent's optimal price responses. Condition (b) says that the various quantities form a fixprice equilibrium (according to def 1) for the price vector p*. Further discussion and conditions for existence can be found in Bénassy (1988, 1990).

Bibliographical References

So far we have concentrated in this entry on the microeconomic concepts allowing us to deal with non-clearing markets at a general equilibrium level. We now indicate further bibliographical references both on the early history of the domain and on macroeconomic applications.

History

The field we described in this entry has a triple ancestry. On one hand Walras (1874) developed a model of general equilibrium with interdependent markets where adjustment was made through prices. This model, in its modern reformulation (Arrow and Debreu 1954; Arrow 1963; Debreu 1959) has become the basic benchmark concept in microeconomics. On the other hand Keynes (1936) and Hicks (1937) built, at the macroeconomic level, a concept of equilibrium where adjustment was made by quantities (the level of national income) as well as by prices. Finally, following the contributions by Chamberlin (1933) and Robinson (1933), progress was made on the treatment of imperfect competition. Notably, Negishi (1961) formalized imperfect competition with subjective demand curves in a general equilibrium framework.

A few isolated contributions in the post-war period made some steps towards modern theories of non-clearing markets. Bent Hansen (1951) introduced the ideas of active demand, close in spirit to that of effective demand, and of quasi-equilibrium where persistent disequilibrium created steady inflation. Patinkin (1956, ch. 13) considered the situation where firms might not be able to sell all their Walrasian output. Hahn and Negishi (1962) studied non-tâtonnement processes where trade could take place before a general equilibrium price system was reached.

A stimulating impetus came from the contributions of Clower (1965) and Leijonhufvud (1968), who reinterpreted Keynesian analysis in terms of market rationing and quantity adjustments. These insights were included in the first fixpricefixwage macroeconomic model by Barro and Grossman (1971, 1976).

The main subsequent development was the construction of rigorous microeconomic concepts allowing us to deal with non-clearing markets and imperfect competition in a full multimarket general equilibrium setting, as described above. Notably, Drèze (1975) and Bénassy (1975a, 1977b, 1982) bridged the gap between the Walrasian and Keynesian lines of thought by generalizing the Walrasian equilibrium concept to integrate non-clearing markets and quantity signals. The link between this new line of work and the imperfect competition equilibrium concepts in the Negishi (1961) line was made in Bénassy (1976, 1977a, 1988). These contributions led to the unified framework we set out in the previous sections. Of course, since one of the main goals of this line of research was to bridge the gap between microeconomics and macroeconomics, there were a number of macroeconomic applications of the above concepts, which we now briefly describe.

Macroeconomic Applications

As indicated above, the first fully worked out fixprice-fixwage macroeconomic model embedding the notions set out above is that of Barro and Grossman (1971, 1976). Early attempts are found in Glustoff (1968) and Solow and Stiglitz (1968). Further developments of the model were made in Bénassy (1977a, 1982, 1986), Malinvaud (1977), Hildenbrand and Hildenbrand (1978), Muellbauer and Portes (1978), Honkapohja (1979), Neary and Stiglitz (1983), and Persson and Svensson (1983). Most of these models concentrated on the problem of employment and policy. Other problems have been treated with this methodology, including notably foreign trade (Dixit 1978; Neary 1980; Cuddington et al. 1984), growth (Ito 1980; Picard 1983; D'Autume 1985), business cycles (Bénassy 1984), as well as the specific problems of planned socialist economies (Portes 1981).

An important part of this line of macroeconomic modelling is that concerned with the explicit introduction of price making and imperfect competition in the macro-setting. Models of that type can be found notably in Bénassy (1977a, 1982, 1987, 1990, 1991), Negishi (1977, 1979), Hart (1982), Snower (1983) Weitzman (1985), Svensson (1986), Blanchard and Kiyotaki (1987), Dixon (1987), Sneessens (1987), Silvestre (1988) and Jacobsen and Schultz (1990).

Now the concepts described in this entry are full general equilibrium models in the tradition of, say, Arrow and Debreu (1954) and Debreu (1959). Contemporaneously to these developments, other authors developed, under the initial name of real business cycles, dynamic stochastic models based on the hypothesis of rational expectations. At some point these two lines of work were synthesized, and the result of this synthesis is described in the dictionary article 'dynamic models with non-clearing markets'.

See Also

- ▶ Dynamic Models with Non-clearing Markets
- ▶ Fixprice Models

Bibliography

Arrow, K. 1959. Towards a theory of price adjustment. In *The allocation of economic resources*, ed. M. Abramowitz. Stanford: Stanford University Press.

Arrow, K. 1963. The role of securities in the optimal allocation of risk-bearing. *Review of Economic Studies* 31: 91–96.

Arrow, K., and G. Debreu. 1954. Existence of an equilibrium for a competitive economy. *Econometrica* 22: 265–290.

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

Barro, R., and H. Grossman. 1976. *Money, employment and inflation*. Cambridge: Cambridge University Press.

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

Bénassy, J.-P. 1975b. Disequilibrium exchange in barter and monetary economies. *Economic Inquiry* 13: 131–156.

Bénassy, J.-P. 1976. The disequilibrium approach to monopolistic price setting and general monopolistic equilibrium. *Review of Economic Studies* 43: 69–81.

Bénassy, J.-P. 1977a. A Neo-Keynesian model of price and quantity determination in disequilibrium. In *Equilibrium and disequilibrium in economic theory*, ed. G. Schwödiauer. Boston: Reidel Publishing Company.

Bénassy, J.-P. 1977b. On quantity signals and the foundations of effective demand theory. *Scandinavian Journal of Economics* 79: 147–168.

Bénassy, J.-P. 1982. *The economics of market disequilibrium*. New York: Academic Press.

Bénassy, J.-P. 1984. A non-Walrasian model of the business cycle. *Journal of Economic Behavior and Organization* 5: 77–89.

Bénassy, J.-P. 1986. *Macroeconomics: An introduction to the non-Walrasian approach*. Orlando: Academic Press.

Bénassy, J.-P. 1987. Imperfect competition, unemployment and policy. *European Economic Review* 31: 417–426.

Bénassy, J.-P. 1988. The objective demand curve in general equilibrium with price makers. *Economic Journal* 98: 37–49.

Bénassy, J.-P. 1990. Non Walrasian equilibria, money and macroeconomics. In *Handbook of monetary economics*, ed. B. Friedman and F. Hahn. Amsterdam: North-Holland.

Bénassy, J.-P. 1991. Microeconomic foundations and properties of a macroeconomic model with imperfect competition. In *Issues in contemporary economics, volume 1: Markets and welfare*, ed. K. Arrow. London: Macmillan.

Blanchard, O., and N. Kiyotaki. 1987. Monopolistic competition and the effects of aggregate demand. *American Economic Review* 77: 647–666.

Bushaw, D., and R. Clower. 1957. *Introduction to mathematical economics*. Homewood: Richard D. Irwin.

Chamberlin, E. 1933. *The theory of monopolistic competition*. 7th ed, 1956. Cambridge, MA: Harvard University Press.

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

Cuddington, J., P.-O. Johansson, and K. Löfgren. 1984. *Disequilibrium macroeconomics in open economies*. Oxford: Basil Blackwell.

D'Autume, A. 1985. *Monnaie, Croissance et Déséquilibre*. Paris: Economica.

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

Dixit, A. 1978. The balance of trade in a model of temporary equilibrium with rationing. *Review of Economic Studies* 45: 393–404.

Dixon, H. 1987. A simple model of imperfect competition with Walrasian features. *Oxford Economic Papers* 39: 134–160.

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

Gabszewicz, J., and J.-P. Vial. 1972. Oligopoly 'A la Cournot' in a general equilibrium analysis. *Journal of Economic Theory* 42: 381–400.

Glustoff, E. 1968. On the existence of a Keynesian equilibrium. *Review of Economic Studies* 35: 327–334.

Grandmont, J.-M., and G. Laroque. 1976. On Keynesian temporary equilibria. *Review of Economic Studies* 43: 53–67.

Greenberg, J., and H. Muller. 1979. Equilibria under price rigidities and externalities. In *Game theory and related topics*, ed. O. Moeschlin and D. Pallaschke. Amsterdam: North-Holland.

Hahn, F., and T. Negishi. 1962. A theorem on non tatonnement stability. *Econometrica* 30: 463–469.

Hansen, B. 1951. *A study in the theory of inflation*. London: Allen and Unwin.

Hart, O. 1982. A model of imperfect competition with Keynesian features. *Quarterly Journal of Economics* 97: 109–138.

Hicks, J. 1937. Mr. Keynes and the classics: A suggested inpt. *Econometrica* 5: 147–159.

Hildenbrand, K., and W. Hildenbrand. 1978. On Keynesian equilibria with unemployment and quantity rationing. *Journal of Economic Theory* 18: 255–277.

Honkapohja, S. 1979. On the dynamics of disequilibria in a macro model with flexible wages and prices. In *New trends in dynamic system theory and economics*, ed. M. Aoki and A. Marzollo. New York: Academic Press.

Ito, T. 1980. Disequilibrium growth theory. *Journal of Economic Theory* 23: 380–409.

Jacobsen, H.-J., and C. Schultz. 1990. A general equilibrium macro model with wage bargaining. *Scandinavian Journal of Economics* 92: 379–398.

Keynes, J.M. 1936. *The general theory of money, interest and employment*. New York: Harcourt Brace.

Leijonhufvud, A. 1968. *On Keynesian economics and the economics of Keynes*. Oxford: Oxford University Press.

Malinvaud, E. 1977. *The theory of unemployment reconsidered*. Oxford: Basil Blackwell.

Marschak, T., and R. Selten. 1974. *General equilibrium with price-making firms*. Berlin: Springer-Verlag.

Muellbauer, J., and R. Portes. 1978. Macroeconomic models with quantity rationing. *Economic Journal* 88: 788–821.

Neary, P. 1980. Nontraded goods and the balance of trade in a neo-Keynesian temporary equilibrium. *Quarterly Journal of Economics* 95: 403–430.

Neary, P., and J. Stiglitz. 1983. Towards a reconstruction of Keynesian economics: Expectations and constrained equilibria. *Quarterly Journal of Economics* 98: 199–228.

Negishi, T. 1961. Monopolistic competition and general equilibrium. *Review of Economic Studies* 28: 196–201.

Negishi, T. 1977. Existence of an underemployment equilibrium. In *Equilibrium and disequilibrium in economic theory*, ed. G. Schwödiauer. Boston: D. Reidel Publishing Company.

Negishi, T. 1979. *Microeconomic foundations of Keynesian macroeconomics*. Amsterdam: North-Holland.

Nikaido, H. 1975. *Monopolistic competition and effective demand*. Princeton: Princeton University Press.

Patinkin, D. 1956. *Money, interest and prices*. 2nd ed, 1965. New York: Harper and Row.

Persson, T., and L. Svensson. 1983. Is optimism good in a Keynesian economy? *Economica* 50: 291–300.

Picard, P. 1983. Inflation and growth in a disequilibrium macroeconomic model. *Journal of Economic Theory* 30: 266–295.

Portes, R. 1981. Macroeconomic equilibrium and disequilibrium in centrally planned economies. *Economic Inquiry* 19: 559–578.

Robinson, J. 1933. *The economics of imperfect competition*. 2nd ed, 1969. London: Macmillan.

Schulz, N. 1983. On the global uniqueness of fixprice equilibria. *Econometrica* 51: 47–68.

Silvestre, J. 1982. Fixprice analysis of exchange economies. *Journal of Economic Theory* 26: 28–58.

Silvestre, J. 1983. Fixprice analysis in productive economies. *Journal of Economic Theory* 30: 401–409.

Silvestre, J. 1988. Undominated prices in the three good model. *European Economic Review* 32: 161–178.

Sneessens, H. 1987. Investment and the inflation–unemployment trade off in a macroeconomic rationing model with monopolistic competition. *European Economic Review* 31: 781–815.

Snower, D. 1983. Imperfect competition, unemployment and crowding out. *Oxford Economic Papers* 35: 569–584.

Solow, R., and J. Stiglitz. 1968. Output, employment and wages in the short run. *Quarterly Journal of Economics* 82: 537–560.

Svensson, L. 1986. Sticky goods prices, flexible asset prices, monopolistic competition and monetary policy. *Review of Economic Studies* 53: 385–405.

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

Walras, L. 1874. *Eléments d'économie politique pure*. Lausanne: Corbaz. Trans. W. Jaffe, Elements of pure economics. London: Allen and Unwin, 1954.

Weitzman, M. 1985. The simple macroeconomics of profit sharing. *American Economic Review* 75: 937–953.

Non-competing Groups

Neil de Marchi

The term is due to John Elliot Cairnes, for whom it summarized ‘the limitations imposed by social circumstances on the free competition of labour’ (Cairnes 1974, p. 73). Cairnes stressed acquired skill as the main impediment to occupational mobility, but he was at pains to make clear that there is no uniform nor even necessarily direct relation between skill and wages, nor therefore between skill and cost of production. Cost of production measures the sacrifices of labour and abstinence, but skill is no element of cost, though it will generally serve as an index of the labour and abstinence undertaken to acquire it. The fact is, however, that great skill may issue in a product that sells quite cheaply, and vice versa (p. 85). It is true, Cairnes acknowledged, that rewards within an occupation will reflect skill differences. His concern, however, in this context, was wages and costs *between* occupations. What skill commands between occupations depends on the exercise of monopoly power. Thus non-competing groups meant just that: occupational groups between which wages reflected non-competitive conditions.

Cairnes used the notion of non-competing groups to show the need for modifying the law according to which relative values reflect relative costs. But if cost does not determine value, what does? John Stuart Mill had shown that reciprocal demand between trading partners determines the terms of trade within limits set by comparative costs. Cairnes took the idea of reciprocal demand and applied it to the problem of non-competing groups between occupations in a single country, to create a composite cost-plus-demand law of value for commodities not exchanging under conditions of free competition (p. 99).

This, too, is how the notion has been applied more recently. If, after occupational irksomeness, differences in ability and returns to human capital are fully allowed for but wage differences stubbornly remain, then, it is concluded, the labour market must be segmented artificially. This creates differentials which represent pure surplus or rent. The segmented or 'dual' or internal labour markets thus identified are characterized by restrictions on labour supply or some institutional factor which causes temporary downward inflexibility of wages in the specially protected market. So-called 'dual labour market theory' has been used to help account for urban poverty, minority underemployment and male–female earnings differences. It can also shed light on the functioning of an economy such as that of South Africa.

See Also

▶ Cairnes, John Elliott (1823–1875)

Bibliography

Cairnes, J.E. 1874. *Some leading principles of political economy*. London: Macmillan.

Doeringer, P.B., and M.J. Piore. 1971. *Internal labor markets and manpower analysis*. Lexington: D.C. Heath.

Gordon, D.M. 1972. *Theories of poverty and underemployment*. Lexington: D.C. Heath.

Porter, R.C. 1978. A model of the Southern African-type economy. *American Economic Review* 68(5): 743–755.

Non-convexity

A. Mas-Colell

Since non-convexity is just the negation of convexity, it will be useful to begin by reviewing the justifications for the latter.

Convexity

eOn the importance of the convexity hypothesis, see the entries: ▶ Existence of General Equilibrium, ▶ Convex Programming, ▶ Convexity, ▶ Duality.

The standard convexity hypotheses can be justified in a variety of ways. Three approaches will be reviewed here. The first is relevant mainly (but not exclusively) to consumption, the second to production and the third to both.

Diversification

We assume that the consumption set is always convex. The approach to be considered now has no bearing on this convexity hypothesis.

The classical justification of the convexity of preferences views it as the mathematical expression of a fundamental tendency of economic choice; namely, the propensity to diversify consumption.

Within a traditional cardinalist context (as in, for example, Jevons, Menger and Walras) diversification is the natural consequence of the principle of decreasing marginal utility: successive units of a consumption good yield increasingly smaller amounts of utility. In turn, if decreasing marginal utility is postulated from any origin and for any (simple of composite) commodity what we get, in modern language, is precisely the hypothesis of concavity of the utility function (proof in the differential case: the second derivative matrix of the utility function is negative semi-definite everywhere). Within an ordinalist context, the principle of decreasing marginal utility should be replaced (as was done by Pareto) by the principle of decreasing marginal rate of substitution: keeping utility constant it is increasingly more difficult, i.e. more expensive, to replace units of a consumption good by units of another. Equivalently, indifference hypersurfaces bound convex sets. In other words: preferences are convex.

It should be clear that as an interpretation (but perhaps with less force as a justification), the above applies also to production. Suppose that inputs and outputs are perfectly divisible. Then the convexity hypothesis on the production set simply says that from any initial point at its boundary and for any definition of (simple or composite) input and output commodity, it takes an increasingly large amount of input to produce successive additional units of output.

While the propensity to diversify is plausible enough as a descriptive feature of economic choice (indeed if this were not so much of economics would be seriously out of tune with economic reality) it is by no means a universal principle. A familiar example to illustrate this is the gin and tonic choice situation. One may well like both gin and tonic but hate its mixture (Exercise: do some introspection and come up with a similar example that applies to you, the reader).

The modern theory of choice under risk, i.e., the expected utility theory of von Neumann and Morgenstern (see ► Expected Utility Hypothesis) has provided, in the form of the theory of risk aversion, a powerful reinforcement to the diversification principle. Suppose that preferences over lotteries (with commodity bundles as outcomes and with objective probabilities) are expressible by taking the expectation of a utility function defined on commodity bundles (this is what the Expected Utility Theory yields). Then the concavity hypothesis on this utility function is equivalent, as a matter of the definition of concavity, to the assumption (called risk aversion) that the decision maker would never lose by getting, instead of a risky lottery with commodities or outcomes, the non-risky commodity bundle where the amount of each commodity is precisely the expected amount of that commodity under the given lottery (i.e., the mean of the random variable). To the extent that risk aversion seems more prevalent than its opposite, we thus get additional support for the convexity hypothesis.

Divisibility and Additivity

A production set $Y \subset R^n$ satisfies the *non-increasing returns property* if any feasible technology $y \in Y$ can be scaled down, that is $\alpha y \in Y$ for any $0 \leq \alpha \leq 1$. The condition can be derived from a more basic requirement, namely, the perfect divisibility of all the inputs used in production. Note: the list of inputs should be exhaustive and inclusive of the non-marketed inputs.

A production set $Y \subset R^n$ satisfies the *additivity property* if $y_1 + y_2 \in Y$ whenever $y_1, y_2 \in Y$ or $Y + Y \subset Y$. The economic interpretation of this condition is straightforward: production activities do not interfere with each other. If activities y_1, y_2 are technically feasible, then it is also feasible, say, to set-up two plants producing, respectively, y_1 and y_2 (if $y_1 = y_2$ then this is what is called *free entry*). Note that for this interpretation to make sense we must again have an exhaustive listing of inputs. In fact, it can be argued that additivity is a test for the exhaustiveness of the listing. In this view a lack of additivity is indicative of an input unaccounted for and available in a fixed amount.

The combination of the two properties above implies that Y is convex. Indeed if y_1, y_2 are feasible and $0 \leq \alpha \leq 1$ then by non-increasing returns, $\alpha y_{1,} (1 - \alpha) y_2$ are feasible, and therefore, by additivity, $\alpha y_1 + (1 - \alpha) y_2$ is also feasible. Although we are not now emphasizing this, we should point out that Y is also a cone, i.e., satisfies the constant returns property: if $y \in Y$ then $\alpha y \in Y$ for an $\alpha \geq 0$. To see this note that for an integer $m > \alpha$ we have any $my \in Y$ by additivity and then

$$\alpha y = \frac{\alpha}{m} my \in Y$$

by non-increasing returns.

See Koopmans (1951) for more on this.

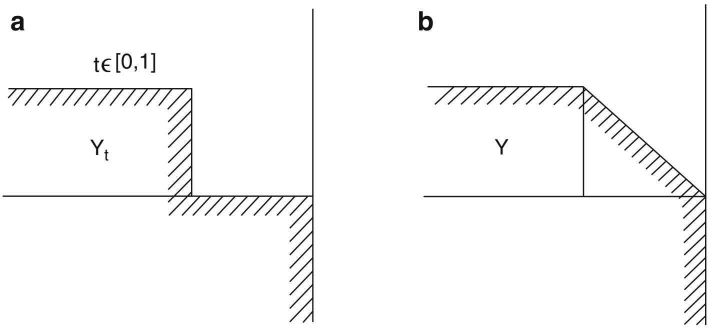
Averaging

In economics we are typically more interested in average than in total magnitudes, e.g., income per capita is a more important concept than total income. It is therefore of great significance that, as we shall now see, the mean behaviour of a collection of economic agents tends to be more regular, more convex-like, than its individual behaviour.

For definiteness the remarks of this subsection will be made in terms of producers. They apply as well to the aggregation of consumers' upper sets (a construct of key importance in welfare economics) or to the aggregation of individual demand correspondences.

Consider first the limit situation where there is literally a continuum of firms. Every firm $t \in [0, 1]$ has a production set $Y_t - R_t^n \subset Y_t$ (free disposal). The dependence of Y_t fulfills the technical condition of measurability. Assume further, and the technical condition of measurability. Assume further, and this is very important, that the Y_t are uniformly bounded above, i.e., there is $z \geq 0$

Non-convexity, Fig. 1



such that $y_t \leq z$ for any $y_t \in Y_t$ and t. Note that the Y_t need not be convex.

The mean (per firm) production set Y is defined in the obvious ways as the collection of mean vectors $\int_0^1 y(t)\mathrm{d}t$ obtained by letting $y(t)$ take values in Y_t. It is denoted by $Y = \int_0^1 Y_t \mathrm{d}t$. It is then a simple consequence of Lyapunov's theorem on the range of a vector measure (see ▶ Lyapunov's Theorem) that Y is convex. Thus even if the individual supply correspondences are not convex valued the aggregate one will be.

The common sense of this result is illustrated in Fig. 1. In it we have a continuum of identical firms. The mean production set is then the convex hull of the common technology.

The limit theory (due to Aumann 1964; Vind 1964) is elegant and conclusive but often one is more interested in obtaining bounds for given, finite situations. In fact, the convexifying effects of averaging were first noted in this context by Farrell (1959) and Rothenberg (1966) and systematically studied by Starr (1969). The key mathematical theorem used by the latter, the Shapley–Folkman theorem (see the entry under that heading), was prompted by the economic application.

The Shapley–Folkman Theorem allows us to assert that every vector in the convex hull of the sum of a finite number of production sets $Y_j \subset R^n$, $j = 1, \ldots, m$ can be obtained as a sum of vectors from the individual convex hulls with at most n of the individual vectors not belonging to the individual sets themselves. Suppose now that the Y_j are uniformly bounded above. It follows that there is a uniform (on j) bound r on the diameters of balls which are contained in the convex hull of Y_j but do not intersect Y_j itself. This r constitutes a measure of the degree of nonconvexity of the family of individual production sets. The Shapley–Folkman theorem implies then that any ball which is contained in the convex hull of $\sum_{j=1}^m Y_j$ but does not intersect $\sum_{j=1}^m Y_t$ itself must have diameter at most lr. Hence, the degree of non-convexity of $\sum_{j=1}^m Y_j$ is bounded independently of the number of firms. So, if m is large the mean production set

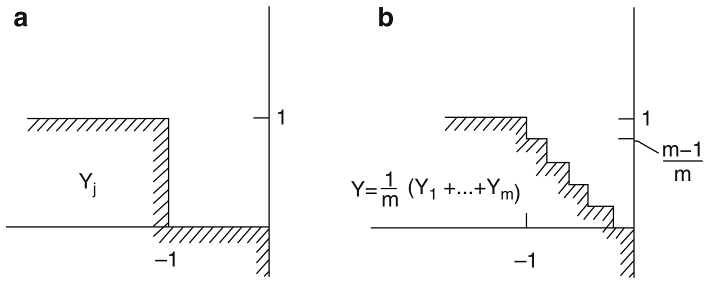
$$Y = \frac{1}{m}\sum_{j=1}^{m} Y_j$$

is almost convex. This is illustrated in Fig. 2. In many cases of economic interest, e.g., if each production set has a smooth boundary, it is possible to do even better: the degree of nonconvexity may actually go to zero. See Mas-Colell (1985) for more on this.

The averaging theory presented so far is entirely modern. The classics, who lacked the concept of supply correspondence, had no inkling of it. They had, however, a very clear conception of the regularizing effects of aggregation. As an example among many we quote from Walras (1954, p. 95, emphasis in the original):

> There is nothing to indicate that the individual demand curves are … *continuous,* in other words that an infinitesimally small increase in p_a produces an infinitesimally small decrease in d_a. On the contrary, these functions are often discontinuous. In the

Non-convexity, Fig. 2



> case of oats, for example, surely our first holder of wheat will not reduce his demand gradually as the price rises, but he will do it in some intermittent way every time he decides to keep one horse less in his stable. His demand curve will, in reality, take the form of a step curve . . . All the other individual demand curves will take the same general form. And yet, the aggregate demand curve can, for all practical purposes, be considered as continuous by virtue of the so-called *law of large numbers.* In fact, whenever a very small increase in price takes place, at least one of the holders of wheat, *out of a large number of them,* will then reach the point of being compelled to keep one horse less, and thus a very small diminution in the total demand for oats will result.

What this says is that if there is enough variation on firms' individual production sets, then, whatever the price system most firms will maximize profits at a single production vector.

Therefore, in the limit, supply jumps are smoothed out and aggregation will yield a supply *function,* i.e., it is as if mean supply was generated from a strictly convex production set. Consider the following example. For every $t \in [0{,}1]$ the production set it $Y_t = \{(0,0), (-1,t)\} - R_t^2$, i.e., one unit of input produces t units of output. The corresponding supply correspondence is

$$f_t(p) = \begin{cases} (0,0) & \text{for } p_1/p_2 > t \\ \{(0,0), (-1,t)\} & \text{for } p_1/p_2 = t \\ (-1,t) & \text{for } p_1/p_2 < t \end{cases}$$

Hence (normalizing to $p_2 = 1$) mean supply is given by the function $F(p_1) = \int f_t(p_1)\mathrm{d}t = \left(p_1 - 1, \frac{1}{2}\left(1 - p_1^2\right)\right)$ for $p_1 \leq 1$ and $F(p) = (0,\ 0)$ for $p_1 > 1$. Figure 3 describes the dispersed family of individual production sets and the corresponding (strictly convex) mean production set.

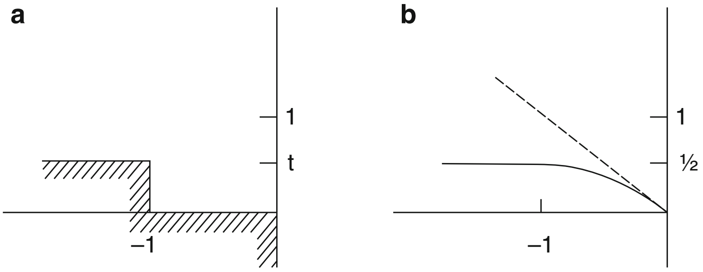
Prompted by a suggestion of Debreu (1972) this 'smoothing by aggregation' problem, which as we have seen can be viewed as an alternative line of attack to the analysis of the convexifying effects of aggregation, has been extensively studied in the last decade. We refer to the excellent survey monograph by Trockel (1984). A conclusion of the research reported in it is that as long as we are interested in the continuity of mean supply and demand then the smoothing intuition can be substantiated by using natural (and weak) concepts of dispersion. Establishing differentiability, however, turns out to be quite a different matter. The theory becomes delicate and powerful mathematical techniques have to be invoked.

Causes of Non-convexities

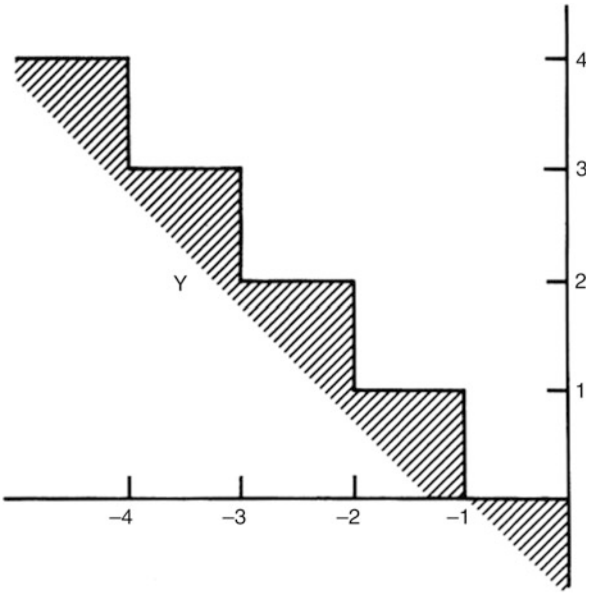
We shall concentrate on the production side. As for consumption, recall the gin and tonic example, or the possibility of risk-loving preferences, or the indivisibilities of many consumption goods. Nonetheless, many of these non-convexities, although individually significant, are small from the aggregate point of view and they may well be averaged out in the manner just mentioned. Of course, this can also happen for many production non-convexities. Thus our interest from now on will be on production non-convexities which matter economy-wide.

In section "Divisibility and Additivity" we saw that non-increasing returns and additivity jointly yield convexity. As indicated there the violation of

Non-convexity, Fig. 3



Non-convexity, Fig. 4



either of those two properties can always be formally traced to, respectively, the indivisibility or the fixity of some input. However, it will be useful now to be rather more concrete.

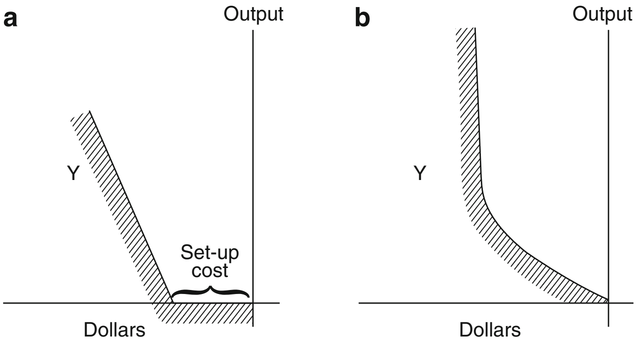
We begin by retaining additivity and examining violations of the non-increasing returns property. Four common instances are:

(a) There is a single input and a single output. The nature of the output, or the input, is such that it can only be produced, or used, in lumps of a fixed size; see Fig. 4.

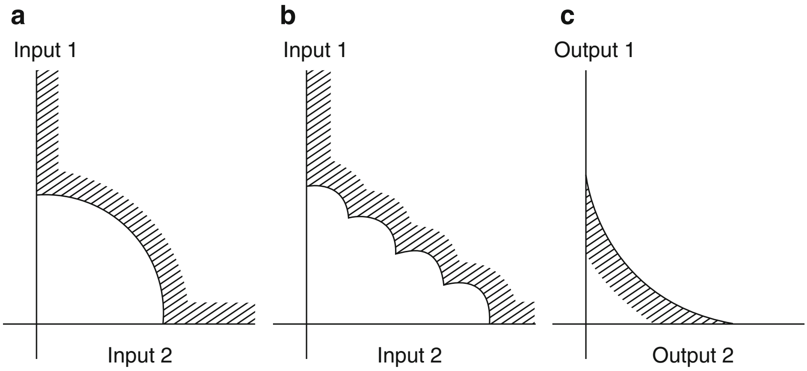
(b) The familiar technology set with set-up cost represented in Fig. 5a (or, in a smoothed out variation, in Fig. 5b). Here the production set is a reduced technology giving the total output optimally obtainable from some total cost or labour input. The non-convexity reflects sizeable indivisibilities in some of the physical inputs required in the production process.

(c) The cause of the increasing returns need not be the indivisibility of a physical input. They could also originate in learning and organizational advantages in the internal structure of production. A classical example is Adam Smith's idea of labour productivity being determined, through specialization and the division of labour, by the extent of the market. Smith's idea can be viewed as a brilliant trick to obtain increasing returns on a scale significantly higher than the individual labourer for a world where labour is the only input and where, therefore, there is no capital good whose indivisibility could be appealed to. In

Non-convexity, Fig. 5



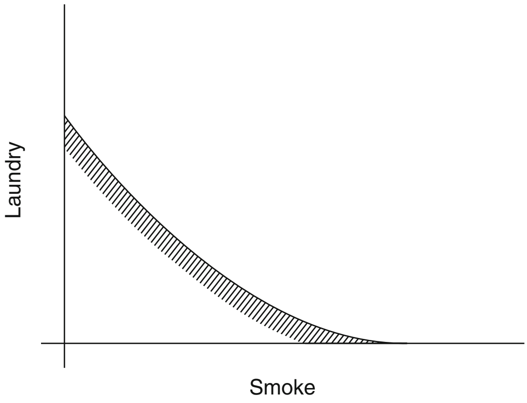
Non-convexity, Fig. 6



Smith's the indivisibilities are present, so to speak, at the level of the performance of individual tasks by individual labourers. Hence, the fewer tasks the latter perform the more productive they will be; see Vassilakis (1986).

(d) Marshallian external economies provide another interesting example. Suppose that the output of an industry is a good proxy for a public positive input (e.g. quality of labour force) to the industry itself. Then the production set of the industry may well be as in Fig. 5b (with free entry this will be the typical shape). We point out that an indivisibility interpretation, while not impossible, would be here rather constrained.

The four previous examples are compatible with additivity of production sets. It is an interesting fact that if increasing returns prevail then the preservation of additivity does not mitigate the non-convexities. Rather the contrary, it only helps to spread them around. For example, if an output can be produced by means of two elementary technologies each of them using a different single input but both of them exhibiting increasing returns then the isoquants of the production function will be as in Fig. 6a (see, e.g., Debreu and Koopmans 1982). Figure 6b represents the situation for a finite number of non-linear elementary activities. Thus we see that a necessary condition for a convex isoquant (an hypothesis very often made in theoretical work) is the availability of an infinite number of elementary activities. Similarly, Fig. 6c represents the production possibility set for two outputs producible (each of them separately) from a single input with increasing returns. Again, it is non-convex. What all this tells us is that while a fully convex world can be supported by a very parsimonious set of microeconomic hypotheses a conveniently 'semiconvex' world is not so easy to justify.



Non-convexity, Fig. 7

Let us now retain non-increasing returns but drop additivity. It is very easy to see how the (negative) interference of two activities can cause non-convexities. The theory of external diseconomies provides classical examples (see Baumol and Oates 1975, or Starret 1972). Suppose that any of two activities (producing, respectively, laundry and smoke producing output) uses labour under constant returns. Then any degree of interference will generate a non-convex production possibility set. See Fig. 7.

Observe that while Figs. 7 and 6c are identical, the underlying reasons for the non-convexity are very different. Here the technology is of constant returns but additivity breaks down while there the technology is of increasing returns and it is additivity that makes the convexity unavoidable. We may also note that both external economies and external diseconomies are sources of non-convexities. But again the reasons are not the same in the two cases.

The Non-convexity Problem

Significant non-convexities create great difficulties both for equilibrium and for welfare theory. We comment on them in turn.

It is obvious, in the first place, that the existence of Walrasian price-taking equilibria is not to be expected. For example, in Fig. 5, only the no production outcome can be sustained by prices. Technically, the convex valuedness and continuity (more precisely: upper hemicontinuity) of supply, required for existence proofs, will fail.

In itself, the above would not be very destructive. It is not clear after all that in a world with large non-convexities the conditions for perfect competition would be met. Walrasian equilibria may not be therefore the most sensible solution concept to look at. The point is, however, that delicate existence problems are present in any of the many, arguably more appropriate, solution concepts proposed in the literature (some will be reviewed in the next subsections). There is a way to see that the difficulty is intrinsic to the non-convex physical environment. Consider a collection $Y_l, \ldots, Y_m \subset R^l$ of production sets and define the feasible set

$$F = \left\{ (y_l, \ldots, y_m) \in \prod_{j=1,\ldots,m} Y_j : \sum_j y_j \geq 0 \right\}.$$

If every Y_j is convex, the F is convex, i.e., it has a simple structure. However, if the Y_j may not be convex then, even if they are otherwise quite nice (e.g., they have smooth boundary and satisfy free disposal), the set F may be far from simple, it may even be formed by several disconnected pieces (e.g., one piece could be the no production point, another a high production region that, so to speak, becomes feasible only due to substantial increasing returns). Directly or indirectly the complexity of the set F bears on the likelihood of existence for any solution concept we may consider.

To obtain, through an equilibrium or an explicitly optimizing process, economic outcomes with good welfare properties (say, Pareto optimality) is also no mean feat in a non-convex world. So much so that most equilibrium approaches simply do not get it. See Calsamiglia (1977) for an impossibility theorem which, in essence, asserts that any decentralized equilibrium notion which guarantees optimality with non-convexities must include as one of its steps the solution of an infinite dimensional programming problem.

The previous remarks should perhaps come as no surprise. The global maximum of an arbitrary function is not characterized by any sort of local conditions. Without some type of structural restriction finding it is a programming problem of intractable complexity. A restriction that proves useful is to limit the permissible non-convexities to those that arise from the indivisibility of explicit inputs or outputs (as in Fig. 4). Then the methods of integer programming can be appealed to. Although those are still complex when compared to convex or linear programming (also, Fig. 4 is misleading as to the higher dimensional possibilities), there is nonetheless an extensive body of technical literature and the field is undergoing rapid progress (e.g. Scarf 1981, 1984). In particular, Scarf (1984) shows that for integer programming problems there is a way to associate to every feasible point a finite system of neighbourhoods in such a way that to test for global optimality it suffices to test every neighbourhood set.

Externalities

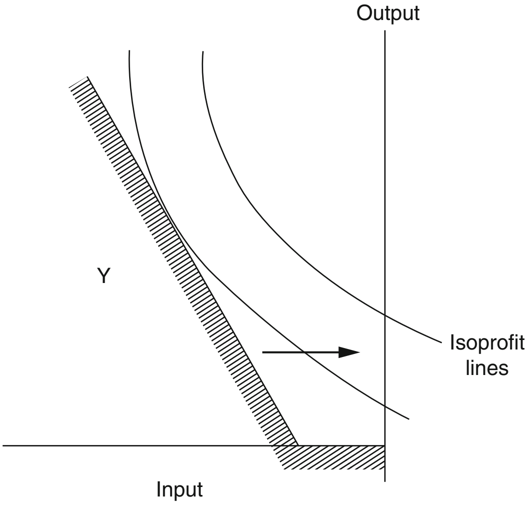
An approach which to a large extent salvages the equilibrium part of Walrasian theory is based on the observation that if all non-convexities in aggregate technologies are external to the single production unit then the decision problem of the individual firm is conventionally looking and, therefore, price taking behaviour is not doomed from the start. The existence of a price taking equilibrium has in fact been proved in considerable generality (see, e.g., Shafer and Sonnenschein 1976; the problem alluded to in section III remains but it can be handled by means of survival hypotheses).

Recently this externality approach has been successfully exploited for the study, by means of dynamic competitive methods, of increasing returns effects in the process of capital accumulation and growth (see Romer 1986).

Because of the presence of externalities the above type of price taking equilibria will typically fail to be Pareto optimal. The other side of the coin is that if external effects are internalized or, simply, priced out, then any Walrasian equilibrium will automatically be Pareto optimal but, because of the non-convexities, it is now the existence of equilibria which will be in serious difficulty (see Starret 1972).

Imperfect Competition

If increasing returns prevail then either the economic equilibrium is very inefficient or individual firms will end up being large. If so, they will be endowed with market power which suggests imperfect competition theory as a proper analytical framework. Interestingly, to this conceptual argument a technical one can be added. The non-linearity of profit functions will increase the likelihood that firms' optimal productions react continuously to market parameters. This is illustrated in Fig. 8 for an output-setting monopolist facing a linear demand function (and maximizing profits in terms of input). It follows that an existence theory for imperfectly competitive equilibria with increasing returns may be available. This is indeed so. It has been developed both for the perceived and the objective demand approach to imperfect competition. The perceived demand case is somewhat easier since the hypotheses of no joint production plus linearity of perceived demand will automatically imply the concavity of profit functions; see Arrow and Hahn (1971), Silvestre (1978) and the survey article by Hart (1985). Altogether, imperfect competition is one of the most promising approaches to increasing returns.



Non-convexity, Fig. 8

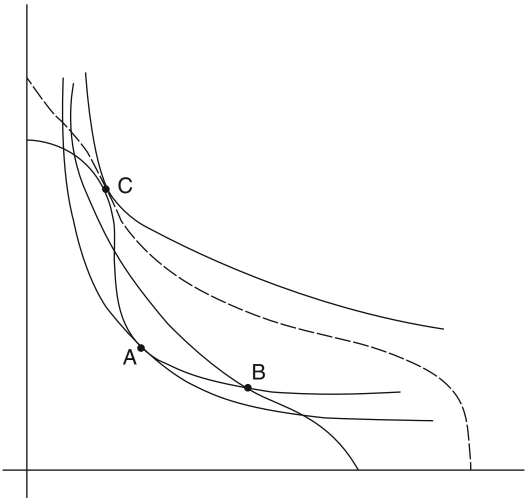
Let us consider a particularly simple example (see Fraysse and Moreau 1981; Dasgupta and Ushio 1981). A certain good can be produced with zero marginal cost but there is a (non-sunk) set-up cost of c. There is free entry and the inverse demand function is $p = 1 - (1/N)Q$, where N is a market size parameter. Such a market will always have a Cournot quantity-setting equilibrium with free entry. The number of active firms will be $\sqrt{N/4c} - 1$ (more precisely, the integer closest from above to this number) while the production per active firm is $2\sqrt{cN}$ and the equilibrium price is $2\sqrt{c/N}$. It is instructive to evaluate the welfare loss. Adopting total surplus as a welfare measure the full optimum would have a single firm producing N at zero price for a total welfare of $(N/2) - c$. In the imperfectly competitive equilibria total welfare would be (approximately)

$$\frac{N}{2} - 3c - \frac{\sqrt{c}}{2}\sqrt{N}.$$

Hence the welfare loss is $2c + (\sqrt{c}/2)\sqrt{N}$ This is of order $\sqrt{N}$ a non-negligible number if N is large (although $\sqrt{N}/N \to 0$ as $N \to \infty$). Is this loss due to the unbounded increasing returns or to the imperfect competition? One way to answer this is to compare it with the situation which is in every way identical except that individual firms have a capacity limit k. Then the welfare loss at the imperfect competition equilibrium can be computed to be of order $k^2/2N$, which is a small number if N is large. Hence increasing returns seem to make quite a difference. Alternatively one could say that the unlimited increasing returns model is inherently much less competitive than the case with bounded non-convexities which, for N large, is almost Walrasian.

Welfare Theory

A Pareto optimal allocation in a non-convex environment satisfies the same first-order necessary conditions as in the convexity case. There must be a price system such that at every production (resp. at every consumption) the price hyperplane must be 'tangent' to the corresponding production set (resp. indifference surfaces). Here tangent means that the firm (resp. the consumer) satisfies the first order necessary conditions for profit (resp. utility) maximization. This is the classical marginal cost pricing principle, so called because for a technology characterized by a single output and a single input it leads to the equality of output price to marginal cost. (Warning: With more than one input cost maximization is not a necessary condition for optimality.) A modern and rigorous analysis of this theory is contained in Bonnisseau and Cornet (1986b). Surprisingly, by using the mathematical techniques of nonsmooth optimization it is possible to relax considerably the differentiability hypotheses.

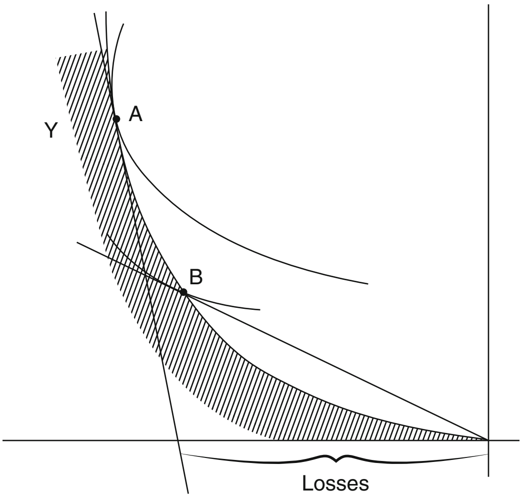


Non-convexity, Fig. 9

A glance at part A in Fig. 9 suffices to see that the first order necessary conditions are not sufficient for optimality. For (local) sufficiency one has to check second order conditions.

Roughly speaking if preferences are convex the second order conditions require that the curvature of the indifference surface by larger than the curvature of the production set, e.g., as in points B and C in Fig. 9. Note that point B is only a local optimum.

It is possible to obtain necessary and sufficient conditions for Pareto optimality by appealing to some form of non-linear prices. Observe, for example, that in Fig. 9 one may separate the production set and the indifference surface at



Non-convexity, Fig. 10

point *C* by a non-linear 'price' surface (dotted line) relative to which the firm maximizes 'profits' and the consumer utility. Note that no such non-linear prices exist for point *B;* see Brown and Heal (1978). Non-linear prices belong to an inherently infinite dimensional price space. Hence the impossibility of reaching a global optimum by using them is not in conflict with the theorem of Calsamiglia mentioned in section "The Non-convexity Problem". For iterative procedures leading to a local optimum see Heal (1973).

Typically if the productions at an optimum are evaluated at the corresponding optimality prices the firms with significant non-convexities will be making losses (marginal cost will be lower than average cost); see point A in Fig. 10. The accounting identities will be taken care of by the lump sum transfers inherent to an optimum (in other words, losses will be covered by receipts from non distortionary taxes). But suppose this is politically infeasible i.e., prices and productions must be such that total profits are non-negative, although they can be limited to be non-positive. Then if we retain the hypothesis that consumers maximize utility given prices (suppose that preferences are convex) Pareto optimality will typically not be reachable. In the one output–one input case, the requirement that profits be zero (i.e., that average and marginal cost by the same) determines the outcome; see point B in Fig. 10. Not so in the multiproduct case. The 'regulatory constraint' of zero profit is compatible with a range of choice of prices and production. This leads to a classical second best problem studied by Boiteux (see Guesnerie 1981, for a modern point of view).

Other Equilibrium Approaches

Imperfect competition is not the only equilibrium approach compatible (to some extent) with non-convexities. A variety of others, more influenced by a planning outlook, have been proposed. Among them are:

(a) Generalized marginal cost pricing equilibrium where firms are assumed to follow the principles described in the previous section, consumers are price takers and distribution rules (including tax subsidies) are given. See Guesnerie (1975), Mantel (1979), Beato (1982) and the recent synthesis by Bonnisseau and Cornet (1986a).

(b) Models where, in contrast to (a), firms do act as profit maximizing price takers but where prices are supplemented by quantity constraints, e.g. perceptions of possible sales. A good example is Dehez and Drèze (1986).

(c) A more abstract approach has been taken by, among others, Dierker et al. (1985), Kamiya (1986), Vohra (1986), and Bonnisseau and Cornet (1986a). Their idea is to analyse the equilibria of systems where firms' behaviour is described by pricing rules (given *a priori*), which specify the prices acceptable at different production decisions; (a) and (b) are included but so are other rules, e.g., average cost pricing.

As could be expected, none of the above approaches yields equilibria with good first best properties (or, for that matter, second best ones; but this has been less studied). This is true even for the notion of marginal cost pricing equilibrium, which is directly inspired by welfare considerations (see Guesnerie 1975; Beato and Mas-Colell 1985). There is, however, an exception: if there is a single production set (i.e., the entire production sector is under a single management) and the curvature of the indifference surfaces is larger than the curvature of the production surface then the marginal cost pricing equilibrium will be Pareto optimal (see Quinzii 1986).

(d) An approach based on (non-linear) Lindhalian prices is pursued in Mas-Colell and Silvestre (1986). The equilibrium is always Pareto optimum (in the one output–one input case it picks the Pareto optima compatible with average cost pricing) but with non-convexities it may not exist (curvature conditions will guarantee existence).

Sustainability

As it is well known there is a close relationship in a convex world between the notion of Walrasian equilibrium and the cooperative game theory concept of the core (see ▶ Cores). With significant non-convexities Walrasian equilibria can easily fail to exist. This is not so clear for the core. In fact the basic intuition of increasing returns seems to suggest that it is difficult for small coalitions to improve their positions by themselves, thus making the core a prime candidate for the analysis of increasing returns economies.

Let $Y \subset R^l$ be a production technology freely available to any agent in the economy. A final allocation of goods is in the core if there is no coalition of agents that can guarantee each of its members a preferred outcome by using only their endowments and the technology *Y*. Note that a core allocation is automatically Pareto optimal. A more general approach would let coalitions have their own technologies; these are the so-called coalition production economies (see, e.g., Oddou 1976). By constructing coalition specific inputs it is possible to view them as a limiting case of the common technology framework.

In the above setting the core has been studied by Scarf (1986). It turns out that the 'basic intuition' described above is not easily substantiated. Indeed, if *Y* is not a convex cone then it is always possible to find a collection of agents yielding an empty core. This is disappointing. There are, however, some special cases for which the core will be non-empty.

(a) There is one output, one input, the technology exhibits decreasing average cost, and consumers own no output.

(b) Consumers derive no utility from input goods and the technology satisfies the property of *distributivity*. By using prices the latter can be described thus: for any efficient production y there is a price system p such that $0 = p \cdot y \geq p \cdot z$ for any $z \in Y$ such that $z^- \leq y^-$ i.e., z should use at most as much input as y.

(c) A particular case of distributive production sets is when there is a single input, average cost decreases radially and the set of output productions attainable from any fixed input is convex (see Sharkey 1979). Recall that this property is not additive. Neither is distributivity.

(d) As with marginal cost pricing relative curvature conditions can also be applied to guarantee a non-empty core (see Quinzii 1986).

There is an intimate connection between the core approach and the sustainability problem in the theory of natural monopoly (see Sharkey 1979; Baumol et al. 1975). Suppose our production set *Y* is additive. This is often described as a natural monopoly situation on the ground that the combined productions of two firms can always be taken care of at least as efficiently by a single firm. The sustainability problem consists in designing a production and compensating (i.e., pricing) system which is immune to (necessarily inefficient) entry. By viewing an entrant as the coalition of its customers the link to core theory becomes clear and it helps to explain the 'paradox' of the existence of unsustainable natural monopolies (i.e., the additivity of *Y* is far from guaranteeing the non-emptiness of the core).

In the theory of natural monopoly a particularly important role is played by the hypothesis that a Walrasian equilibrium exists (e.g., in the one-input, one-output case this says that the demand forthcoming at the minimum average cost is an exact multiple of a minimum efficiency scale). Of course, this implies that the core is non-empty and a sustainable arrangement exists. But more is true. Under weak conditions the Walrasian equilibrium is the only point in the core (a related result, emphasizing the possibility of big players more than non-convexities, is in Shitovitz 1973). Finally, we note that there is a close link between this result and many non-cooperative models of competition 'à la Bertrand'. Indeed, it is often possible to understand the latter as core models in which there are restrictions on which coalitions can form (e.g. they include only one firm) and on the way they can split gains (e.g. only through a uniform price system). The theme that under conditions of free entry (i.e., additivity of the aggregate production set) the existence of a Walrasian equilibrium will imply the non-emptiness and efficiency of the set of non-cooperative equilibria is also common in the latter theory (see Baumol et al. 1982; Grossmann 1981; or Mas-Colell 1985).

See Also

- ▶ Consumption Sets
- ▶ Convex Programming
- ▶ Convexity
- ▶ Cores
- ▶ Duality
- ▶ Existence of General Equilibrium
- ▶ Externalities
- ▶ General Equilibrium
- ▶ Increasing Returns to Scale
- ▶ Lyapunov's Theorem
- ▶ Planning
- ▶ Shapley–Folkman Theorem

Bibliography

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

Baumol, W., and W. Oates. 1975. *The theory of environmental policy.* Engelwood Cliffs: Prentice-Hall.

Baumol, W., J. Panzar, and R. Willig. 1982. *Contestable markets and the theory of industrial structure*. San Diego: Harcourt, Brace, Jovanovich.

Beato, P. 1982. The existence of marginal cost pricing with increasing returns. *Quarterly Journal of Economics* 97: 669–689.

Beato, P., and A. Mas-Colell. 1985. On marginal cost pricing with given tax-subsidy rule. *Journal of Economic Theory* 37(2): 356–365.

Bonnisseau, J.-M., and B. Cornet. 1986a. Existence of equilibria when firms follow bounded losses pricing rules. *Journal of Mathematical Economics,* forthcoming.

Bonnisseau, J.-M., and B. Cornet. 1986b. Valuation equilibrium and Pareto optimum in non-convex economies. *Journal of Mathematical Economics,* forthcoming.

Brown, D., and G. Heal. 1978. Equity, efficiency and increasing returns. *Review of Economic Studies* 46: 571–585.

Calsamiglia, X. 1977. Decentralized resource allocation and increasing returns. *Journal of Economic Theory* 14: 262–285.

Dasgupta, P., and Y. Ushio. 1981. On the rate of convergence of oligopoly equilibria in large markets. *Economic Letters* 8: 13–17.

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

Debreu, G., and T. Koopmans. 1982. Additively decomposed quasiconvex functions. *Mathematical Programming* 24: 1–38.

Dehez, P., and J. Drèze. 1986. *Competitive equilibria with increasing returns*, Core discussion paper, no. 8623. Louvain-la-Neuve: Core.

Dierker, E., R. Guesnerie, and W. Neuefeind. 1985. General equilibrium when some firms follow special pricing rules. *Econometrica* 53: 1369–1393.

Farrell, M. 1959. The convexity assumption in the theory of competitive markets. *Journal of Political Economy* 67: 377–391.

Fraysse, J., and M. Moureau. 1981. Cournot equilibrium in large markets under increasing returns. *Economic Letters* 8(3): 217–220.

Grossman, S. 1981. Nash equilibrium and the industrial organization of markets with large fixed costs. *Econometrica* 49: 1149–1172.

Guesnerie, R. 1975. Pareto optimality in non-convex economies. *Econometrica* 43: 1–29.

Guesnerie, R. 1981. *Modéles de l'économie publique*. Paris: Editions du CNRS.

Hart, O. 1985. Imperfect competition in general equilibrium: An overview of recent work. In *Frontiers of economics*, ed. K. Arrow and S. Honkapohja. Oxford: Blackwell.

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

Kamiya, K. 1986. Existence and uniqueness of equilibria with increasing returns. *Journal of Mathematical Economics.*

Koopmans, T. 1951. Analysis of production as an efficient combination of activities. In *Activity analysis of production and allocation*, ed. T. Koopmans, 33–97. New York: Wiley.

Mantel, R. 1979. Equilibrio con rendimientos crecientes a escala. *Anales de la Asociacion Argentina de Economia Politica* 1: 271–282.

Mas-Colell, A. 1985a. La libre entrada y la eficiencia económica: un análisis de equilibrio parcial. *Revista Española de Economia* 2(1): 135–152.

Mas-Colell, A. 1985b. *The theory of general economic equilibrium: A differentiable approach*. Cambridge: Cambridge University Press.

Mas-Colell, A., and J. Silvestre. 1986. *Cost share equilibria: A Lindahlian approach*, Working paper. Berkeley: MSRI.

Nikaido, H. 1968. *Convex structures and economic theory.* New York: Academic Press.

Oddou, C. 1976. Theorèmes d'existence et d'équivalence pour des économies avec production. *Econometrica* 44: 265–282.

Quinzii, M. 1986. *Rendements croissants et équilibre général*. PhD dissertation, Université Paris II

Rockafeller, T. 1970. *Convex analysis*. Princeton: Princeton University Press.

Romer, P. 1986. Increasing returns and long-run growth. *Journal of Political Economy* 94(5): 1002–1036.

Scarf, H. 1981. Production sets with indivisibilities, Part I: Generalities. *Econometrica* 49: 1–32.

Scarf, H. 1984. *Neighborhood systems for production sets with indivisibilities*, Cowles Foundation discussion paper, no. 728.

Scarf, H. 1986. Notes on the core of a productive economy. Ch. 21. In *Contributions to mathematical economics,* ed. W. Hildenbrand and A. Mas-Colell. Amsterdam: North-Holland.

Shafer, W., and H. Sonnenschein. 1976. Equilibrium, commodity taxation, and lump sum transfers. *International Economic Review* 17: 601–611.

Sharkey, W. 1979. Existence of the core when there are increasing returns. *Econometrica* 47: 869–876.

Shitovitz, B. 1973. Oligopoly in markets with a continuum of traders. *Econometrica* 41: 467–501.

Silvestre, J. 1978. Increasing returns in general non-competitive analysis. *Econometrica* 46: 397–402.

Starr, R. 1969. Quasi-equilibria in markets with non-convex preferences. *Econometrica* 37: 25–38.

Starrett, D. 1972. Fundamental non-convexities in the theory of externalities. *Journal of Economic Theory* 4: 180–199.

Trockel, W. 1984. *Market demand: An analysis of large economies with non-convex preferences*. New York: Springer.

Vassilakis, S. 1986. *Increasing returns and strategic behavior*. PhD thesis, Johns Hopkins University.

Vind, K. 1964. Edgeworth-allocations in an exchange economy with many traders. *International Economic Review* 5(2): 165–177.

Vohra, R. 1986. On the existence of equilibrium with increasing returns: A synthesis. *Journal of Mathematical Economics.*

Walras, L. 1874–7. *Eléments d'économie politique pure*. Definitive edn, 1926, trans. by W. Jaffé as *Elements of pure economics.* Homewood: Irwin, 1954.

Non-cooperative Games

Joseph E. Harrington Jr.

Game theory analyses multi-agent situations in which the payoff to an agent is dependent not only upon his own actions but also on the actions of others. Zero-sum games assume that the payoffs to the players always sum to zero. In that case, the interests of the players are diametrically opposed. In non-zero-sum games, there is typically room for cooperation as well as conflict.

The normal or strategic form characterizes a game by three elements. First, the set of players, $N = \{1, 2,\ldots, n\}$, who will be making decisions. Second, the set of strategies, S_i, available to player $i \forall i \in N$ where a strategy is a rule which tells a player how to behave over the entire course of the game. A strategy often takes the form of a function which maps information sets (that is, a description of where a player is at each stage in the game) into the set of possible actions. Thus, an action is a realization of a strategy. Finally, the normal form specifies the payoff function, $V_{\mathrm{i}}(\cdot)$of player $i \forall i \in N$. A payoff function is a composition of a player's von Neumann–Morgenstern utility function over outcomes and the outcome function which determines the outcome of the game for a given set of strategies chosen. The normal form of a particular game is presented in Fig. 1. The set N is $\{1, 2\}$ while $\mathrm{S}_1 = \{\alpha, \beta, \gamma\}$ and $S_2 = \{a, b, c\}$. The payoff to player 1 (2), for a given pair of strategies, is the first (second) number in the box.

A game is classified as either cooperative or non-cooperative, a distinction which rests not on the behaviour observed but rather on the institutional structure. A cooperative game assumes the existence of an institution which can make any agreement among players binding. In a non-cooperative game, no such institution exists. The only agreements in a non-cooperative game that are meaningful are those which are *self-enforcing*. That is, it is in the best interest of each player to go along with the agreement,

		1		
		α	β	γ
	a	2,2	1,6	0,1
2	b	6,1	5,5	1,2
	c	1,0	2,1	4,4

Non-cooperative Games, Fig. 1

given that the other players plan to do so. In analysing the pricing behaviour of firms in an oligopolistic industry, a non-cooperative game is generally appropriate since, in most countries, cartel agreements are prohibited by law. Therefore, firms do not have access to legal institutions for enforcing contracts and making agreements binding.

Let us examine the game in Fig. 1 under the assumptions that it is non-cooperative and that players are allowed preplay communications. After discussing how they each plan to behave, players 1 and 2 will simultaneously make a decision as to which strategy to play. After the strategies are chosen, the payoffs will be distributed. It is straightforward to show that the class of self-enforcing agreements for this game is {(*α*, *a*), (*γ*, *c)*}. Consider the agreement that player 1 chooses α and player 2 chooses *a.* By choosing *a,* player 2 maximizes his payoff under the assumption that player 1 goes along and plays α Similarly, player 1 finds it optimal to choose α if he believes player 2 will go along with the agreement. Thus, (*α, a)* is a self-enforcing agreement.

To understand the cost imposed by the restriction that an agreement must be self-enforcing, consider the agreement (*β, b*) Since it yields payoffs which are Paretosuperior to both (*α, a)* and (*γ, c*) the two players obviously have an incentive to try and achieve those strategy choices. However, even if they came to the agreement (*β, b*) it would be ineffectual. If player 1 truly believe that player 2 would honour the agreement and play *b,* player 1 would be better off reneging and choosing α instead. Because agreements cannot be made binding, the two players are then forced to settle on a Pareto-inferior outcome.

A solution concept for non-cooperative games which encompasses the notion of self-enforcing agreements is *Nash equilibrium.* Originally formulated by Nash (1950, 1951), the concept finds its roots in the work of Cournot (1838).

Definition An *n*-tuple of strategies, $(s_1^*, \ldots, s_n^*)$ is a Nash equilibrium if

$$V_i\left(s_1^*,\ldots,s_n^*\right) \geq V_i\left(s_1^*,\ldots,s_{i-1}^*,s_i,s_{i+1}^*,\ldots,s_n^*\right) \times\forall s_i \in S_i, \forall i \in N. \tag{1}$$

A profile of strategies forms a Nash equilibrium if each player's strategy is a best reply to the strategies of the other *n*-1 players. The appeal of Nash equilibrium as a solution concept rests on two pillars. First is the stability inherent in a Nash equilibrium since no player has an incentive to change his strategy. Second is the very large class of games for which it can be proved that a Nash equilibrium exists. To substantiate this last remark, we will first need to introduce an additional concept – the mixed strategy. A mixed strategy takes the form of a probability distribution over the set of pure strategies S_i (for example, for the game in Fig. 1, a mixed strategy for player 1 could be to choose α with probability 0.4 and β with probability 0.6). A pure strategy is thus a special case of a mixed strategy in which unit mass is placed on the pure strategy. A game is said to be finite if the strategy set S_i is finite $\forall i \in N$

Theorem (Nash 1950, 1951): In any finite non-cooperative game $\langle N, \{S_i\}_{i \in N}, \{V_i\}_{i \in N} \rangle$, there exists a Nash equilibrium in mixed strategies.

The ease of existence of Nash equilibria also brings forth the major drawback to the

		1	
		α	β
	a	1,1	1,0
2	b	0,1	1,0
	c	1,0	2,2

Non-cooperative Games, Fig. 2

concept – the lack of uniqueness. It has also been observed that when multiple Nash equilibria exist, some of them can be quite unreasonable. For the two-player game in Fig. 2, there exist two pure-strategy Nash equilibria, $\{(\alpha, a), (\beta, c)\}$. However, (α, a) is rather unreasonable as it entails player 1 using the strategy α which is weakly dominated by β (That is, by choosing β instead of α he would never be worse off and could end up better off.) In attempting to define what is meant by reasonable and to achieve a unique solution, work by Selten (1975), Myerson (1978), Kreps and Wilson (1982), Kalai and Samet (1984) and others has developed solution concepts that are more restrictive than Nash equilibrium.

Due to the difference in institutional structures, the issues analysed under a non-cooperative game setting tend to be quite different from those dealth with in cooperative games. Since all agreements can be made binding in cooperative games, much of the analysis is concerned with determining which point in the Pareto-efficient set the players will settle on. Issues of importance are then coalition formation and the division of gains among coalition members. In contrast, in a non-cooperative game, at issue is whether players can even reach the Pareto-efficient frontier; in the game in Fig. 1, they do not. However, it has been observed in both experimental and real world situations (e.g., see Axelrod 1984) that when the game is repeated players are indeed able to achieve Pareto efficiency in a non-cooperative game like in Fig. 1. An important issue in non-cooperative games is then to understand the role of repetition in allowing players to overcome the inability to cooperate.

Let us suppose the one-shot game in Fig. 1 is repeated T times, where $T \geqslant 2$, and the players are fully aware of the repetition. A strategy for player i now takes the form of a sequence of functions, $\{G_i^t\}_{t=1}^T$, where G_i^t maps the history of play over $\{1, \ldots, t-1\}$ into the set $\{\alpha, \beta, \gamma\}\{(a, b, c)\}$ if $i = 1(2)$. Assume that the payoff to a player is the (undiscounted) sum of the single-period payoffs.

Let g_i^t denote the observed action of player i in period t. Consider the following pair of strategies:

$$G_1^1 = \beta$$

$$\boldsymbol{G}_1^t = \begin{cases} \beta \text{ if } \boldsymbol{g}_1^\tau = -\beta, \boldsymbol{g}_2^\tau = \boldsymbol{b}, \tau = 1, \ldots, \boldsymbol{t} - 1, \\ \qquad\qquad 2 \le \boldsymbol{t} \le \boldsymbol{T} - 1 \\ \gamma \text{ if } \boldsymbol{g}_1^\tau = \beta, \boldsymbol{g}_2^\tau = \boldsymbol{b}, \tau = 1, \ldots, \boldsymbol{T} - 1, \boldsymbol{t} = \boldsymbol{T} \\ \alpha \text{ otherwise;} \end{cases} \tag{2}$$

$$G_2^1 = \boldsymbol{b}$$

$$\boldsymbol{G}_2^t = \begin{cases} \boldsymbol{b} \textit{ if } \boldsymbol{g}_1^\tau = \beta, \boldsymbol{g}_2^\tau = \boldsymbol{b}, \tau = 1, \ldots, \boldsymbol{t} - 1, \\ \qquad\qquad 2 \le \boldsymbol{t} \le \boldsymbol{T} - 1 \\ \boldsymbol{c} \textit{ if } \boldsymbol{g}_1^\tau = \beta, \boldsymbol{g}_2^\tau = \boldsymbol{b}, \tau = 1, \ldots, \boldsymbol{T} - 1, \boldsymbol{t} = \boldsymbol{T} \\ \mathbf{a} \text{ otherwise} \end{cases} \tag{3}$$

The strategy of player 1 says that he will start off by playing β and will continue to do so as long as (β, b) has been observed in all previous periods. If (β, b) was observed for all $t \in \{1, \ldots, T-1\}$ player 1 will choose γ in the final play. However, if the path ever deviates from (β, b) for any $t \leqslant T-1$, he will choose α for the remainder of the game. The strategy of player 2 is similarly defined.

If the two players pursue these strategies, the path of play will be (β, b) for $t \in \{1, \ldots, T-1\}$ and (γ, c) for period T. Each player will earn a total payoff of $5\ T - 1$. The key issue, however, is whether these strategies form a Nash equilibrium. Given that player 1 pursues $\{G_1^t\}_{t=1}^T$, can player

2 earn a payoff higher than 5 $T - 1$ by choosing a strategy different from $\{G_2^t\}_{t=1}^T$? If not, then player 2's strategy in (3) is optimal. The strategy in (3) calls for player 2 to cooperate over $\{1, \ldots, T-1\}$ in the sense of not maximizing his single-period payoff. The alternative strategy is to choose *a* rather than *b* for some $t \leq T - 1$ and earn 6 rather than 5 in that period. Since the gain from cheating is only in that period, it is best to cheat at the last moment so as to maximize the time of cooperation. The best alternative strategy for player 2 is then to choose *b* over $\{1, \ldots, T-2\}$ and cheat in period $T-1$ in playing *a*. The resulting payoff is $5(T-2) + 6 + 2 = 5\ T - 2$. Since this is less than $5\ T - 1$ then $\{G_2^t\}_{t=1}^T$ is a best reply to $\{G_1^t\}_{t=1}^T$. Similarly, one can show that this is true for player 1 as well and therefore the two strategies from a Nash equilibrium.

Repetition on the one-shot game has allowed players to earn an average payoff of 5 – (1/*T*) compared with 4 or 2 in the one-shot game. Furthermore, as the horizon tends to infinity, the average payoff converges to the Pareto-efficient solution. Repetition expands the set of self-enforcing agreements by allowing players to be penalized in the future for cheating on an agreement. The penalty here is that the game moves to the Pareto-inferior single-period Nash equilibrium of (α, *a*) Because it is a Nash equilibrium, this threat is credible. Cooperation is rewarded by settling at the preferred solution (γ, *c*) in the final period. Note that cooperation cannot be maintained over the entire horizon since, in the final period, it is just like the one-shot game. Thus, the players must settle at either (α, *a*) or (γ, *c*) Development of cooperative behaviour in the finite horizon setting is due to work by Benoit and Krishna (1985), Friedman (1985), and Moreaux (1985). However, the original work for the infinite horizon game goes back to Friedman (1971).

When players are allowed preplay communication, there is a very strong basis for requiring a solution to be a Nash equilibrium because such equilibria are self-enforcing. However, when players cannot communicate, Nash equilibrium loses some of its appeal as a solution concept. (Actually, if there are multiple Nash equilibria, this is also true for games with preplay communication as players may fail to come to an agreement.) Work by Bernheim (1984) and Pearce (1984) suggests that there can be profiles of strategies which are reasonable for players to choose yet which do not constitute a Nash equilibrium.

Let us start with the basic premise that each player holds a subjective probability distribution over the strategies *and* beliefs of other players. Furthermore, impose the axiom that 'rationality is common knowledge'. That is, it is common knowledge that each player acts to maximize his payoff subject to his subjective beliefs. A set of beliefs is said to be *consistent* if it is not in violation of the 'rationality is common knowledge' axiom. In particular, you do not expect another player to pursue a nonoptimal strategy. A strategy is *rationalizable* if there exists a set of consistent beliefs for which that strategy is optimal.

To understand rationalizability as a solution concept, consider the game in Fig. 3. The unique pure-strategy Nash equilibrium is (*β, b*). It is easy to show that every Nash equilibrium strategy is rationalizable. *β* is optimal for player 1 if he believes player 2 will choose *b*. This belief is consistent if 1 believes that 2 believes that 1 will choose *β* so that *b* is a best reply for player 2. Similarly, the belief of player 1 that 2 believes that 1 will choose *β* is consistent if 1 believes that 2 believes that 1 believes that 2 will choose *b* and so forth. Thus, Nash equilibria are always rationalizable. However, one can show that γ is also a rationalizable strategy even though it is not part of a Nash equilibrium. γ is optimal for 1 if he believes 2 will choose *a*. That belief is consistent if 1 believes that 2 believes that 1 will play α so that *a* is a best reply. Now that belief is consistent if 1 believes that 2 believes that 1 believes that 2 will choose *c* so that α is a best reply. Finally, if 1 believes that 2 believes that 1 believes that 2 believes that 1 will play γ then *c* is a best reply and we have a cycle of (γ – *a* – α – *c*). By repeating this cycle we have a consistent set of beliefs which makes γ rationalizable. Actually, all the strategies in that cycle can be rationalized by a set of beliefs generated by that cycle. Thus, each strategy in this game is consistent with *some* basic premise concerning rational behaviour.

			1	
		α	β	γ
	a	0,4	1,1	3,2
2	b	1,1	2,2	0,0
	c	2,3	0,0	1,4

Non-cooperative Games, Fig. 3

In this light, we gain a better idea of what the Nash equilibrium concept actually demands. It is not only a restrictions on strategies but also on beliefs. It requires that strategies be best responses to some set of conjectures and that these conjectures about other players' strategies be fulfilled in equilibrium. In a game without preplay communication, such a restriction on beliefs is by no means natural. On the other hand, rationalizability opens up a much wider set of possible outcomes and thus makes it difficult to come to a conclusion concerning behaviour. Since players themselves are faced with the same problem, they may resort to Nash equilibria as a focal point, as defined by Schelling (1960). On this basis, Nash equilibrium regains some of its appeal as a solution concept for non-cooperative games.

See Also

- ► Cooperative Games
- ► Game Theory
- ► Nash Equilibrium

Bibliography

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

Benoit, J.-P., and V. Krishna. 1985. Finitely repeated games. *Econometrica* 53: 905–922.

Bernheim, B.D. 1984. Rationalizable strategic behavior. *Econometrica* 52: 1007–1028.

Cournot, A.A. 1838. *Researches into the mathematical principles of the theory of wealth*. Trans. from French, New York: Macmillan, 1897.

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

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

Friedman, J.W. 1985. Cooperative equilibria in finite horizon noncooperative supergames. *Journal of Economic Theory* 35: 390–398.

Kalai, E., and D. Samet. 1984. Persistent equilibria in strategic games. *International Journal of Games Theory* 13: 129–145.

Kreps, D.M., and R. Wilson. 1982. Sequential equilibria. *Econometrica* 50: 863–894.

Moreaux, M. 1985. Perfect Nash equilibrium in finite repeated games and uniqueness of Nash equilibrium in the constituent game. *Economics Letters* 17: 317–320.

Myerson, R.B. 1978. Refinements of the Nash equilibrium concept. *International Journal of Games Theory* 7: 73–80.

Nash Jr., J.F. 1950. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences of the United States of America* 36: 48–49.

Nash Jr., J.F. 1951. Non-cooperative games. *Annals of Mathematics* 54: 286–295.

Pearce, D.G. 1984. Rationalizable strategic behavior and the problem of perfection. *Econometrica* 52: 1029–1050.

Schelling, T.C. 1960. *The strategy of conflict*. Cambridge, MA: Harvard University Press.

Selten, R. 1975. Reexamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Games Theory* 4: 25–55.

Voro'ev, N.N. 1977. *Game theory*. New York: Springer.

Non-cooperative Games (Equilibrium Existence)

Philip J. Reny

Abstract

This article provides a brief overview of equilibrium existence results for continuous and discontinuous non-cooperative games.

Keywords

Auctions; Bertrand price-competition models; Convexity; Cournot oligopoly models; Discontinuous games; Endogenous sharing rules; Equilibrium; Equilibrium existence; Finite-action games; Fixed point theorems; Infinite-action games; Mixed strategy Nash equilibria; Nash equilibrium; Non-cooperative games; Pure strategy Nash equilibria; Quasi-concavity; Quasi-convexity; Spatial economics; Strategic form games

JEL Classifications

C7

Introduction

Nash equilibrium is *the* central notion of rational behavior in non-cooperative game theory (see Osborne and Rubinstein, 1994, for a discussion of Nash equilibrium, including motivation and input). Our purpose here is to discuss various conditions under which a strategic form game possesses at least one Nash equilibrium.

Strategic settings arising in economics are often naturally modelled as games with infinite strategy spaces. For example, models of price and spatial competition (Bertrand, 1883; Hotelling, 1929), quantity competition (Cournot, 1838), auctions (Milgrom and Weber, 1982), patent races (Fudenberg et al., 1983), and so on, typically allow players to choose any one of a continuum of actions. The analytic convenience of the continuum from both an equilibrium characterization and a comparative statics point of view is perhaps the central reason for the prevalence and usefulness of infinite-action games. Because of this, our treatment will permit both finite-action and infinite-action games.

Games with possibly infinite strategy spaces can be divided into two categories: those with continuous payoffs and those with discontinuous payoffs. Cournot oligopoly models and Bertrand price-competition models with differentiated products, as well as all finite-action games, are important examples of continuous games, while Bertrand price-competition with homogeneous products, auctions, and Hotelling spatial competition are important examples in which payoffs are discontinuous. Equilibrium existence results for both continuous and discontinuous games will be reviewed here. We begin with some notation.

A strategic form game, $G = (S_i, u_i)_{i=1}^N$, consists of a positive finite number, N, of players, and for each player $i \in \{1, \ldots, N\}$, a non-empty set of pure strategies, S_i, and a payoff function $u_i : S \to \mathbb{R}$, where $S = \times_{i=1}^N S_i$. The notations s_{-i} and S_{-i} have their conventional meanings: $s_{\text{-}i} = (s_1, .., s_{i\text{-}1}, s_{i+1}, .., s_N)$ and $S_{-i} = \times_{j \neq i} S_j$. Throughout, we assume that each S_i is a subset of some metric space and that, if any finite number of sets are each endowed with a topology, then the product of those sets is endowed with the product topology.

Continuous Games

Pure Strategy Nash Equilibria

Pure strategy equilibria are more basic than their mixed strategy counterparts for at least two reasons. First, pure strategies do not require the players to possess preferences over lotteries. Second, mixed strategy equilibrium existence results often follow as corollaries of the pure strategy results. It is therefore natural to consider first the case of pure strategies.

Definition $s^* \in S$ is a pure strategy Nash equilibrium of $G = (S_i, u_i)_{i=1}^N$ if for every player i, $u_i(s^*) \geq u_i\left(s_i, s^*_{-1}\right)$ for every $s_i \in S_i$.

An important and very useful result is the following.

Theorem 1 If each S_i is a non-empty, compact, convex subset of a metric space, and each $u_i(s_1, .. , s_N)$ is continuous in $(s_1, .. , s_N)$ and quasi-concave in s_i, then $G = (S_i, u_i)_{i=1}^N$ possesses at least one pure strategy Nash equilibrium.

Proof For each player i, and each $s_{\text{-}i} \in S_{\text{-}i}$, let $B_i(s_{\text{-}i})$ denote the set of maximizers in S_i of $u_i(\cdot; s_{\text{-}i})$. The continuity of u_i and the compactness of S_i ensure that $B_i(s_{-i})$ is non-empty and also ensure, given the compactness of S_{-i}, that the correspondence, $B_i : S_{-i} \twoheadrightarrow S_i$ is upper hemi-continuous. The quasi-concavity of u_i in s_i implies that $B_i(s_{-i})$ is convex. Consequently, each B_i is upper hemi-continuous, non empty-valued and convex-valued. All three of these properties are therefore inherited by the correspondence $B : S \twoheadrightarrow S$ defined by $B(s) = \times_{i=1}^{N} B_i(s_{-i})$ for each $s \in S$. Consequently, we may apply Glicksberg's (1952) fixed point theorem to B and conclude that there exists $\hat{s} \in S$ such that $\hat{s} \in B(\hat{s})$. This $\hat{s}$ is therefore a pure strategy Nash equilibrium. Q.E.D.

Remark Theorem 1 remains valid when 'metric space' is replaced by 'locally convex Hausdorff topological vector space'. See Glicksberg (1952).

Remark The convexity property of strategy sets and the quasi-concavity of payoffs in own action cannot be dispensed with. For example, strategy sets are not convex in matching pennies, and, even though the continuity and compactness assumptions hold there, no pure strategy equilibrium exists. On the other hand, in the two-person zero-sum game in which both players' compact convex pure strategy set is $[-1,1]$ and player 1's payoff function is $u_1(s_1, s_2)=|s_1 + s_2|$, all of the assumptions of Theorem 1 hold except the quasi-concavity of u_1 in s_1. But this is enough to preclude the existence of a pure strategy equilibrium because in any such equilibrium player 2's payoff would have to be zero (given s_1, 2 can choose $s_2 = - s_1$) and 1's payoff would have to be positive (given s_2, 1 can choose $s_1 \neq - s_2$).

Remark More general results for continuous games can be found in Debreu (1952) and Schafer and Sonnenschein (1975). Existence results for games with strategic complements on lattices can be found in Milgrom and Roberts (1990) and Vives (1990).

Mixed Strategy Nash Equilibria

A *mixed strategy* for player i is a probability measure, m_i, over S_i. If S_i is finite, then $m_i(s_i)$ denotes the probability assigned to s_i A S_i by the mixed strategy m_i, and i's set of mixed strategies is the compact convex subset of Euclidean space $M_i = \left\{ m_i \in [0,1]^{\#S_i} : \sum_{si \in S_i} m_i(s_i) = 1 \right\}$.

In general, we shall not require S_i to be finite. Rather, we shall suppose only that it is a subset of some metric space. In this more general case, a mixed strategy for player i is a (regular, countably additive) probability measure, m_i, over the Borel subsets of S_i; for any Borel subset A of S_i, $m_i(A)$ denotes the probability assigned to A by the mixed strategy m_i. Player i's set of such mixed strategies, M_i, is then convex. Further, if S_i is compact, the convex set M_i is compact in the weak-* topology (see, for example, Billingsley, 1968).

Extend $u_i : S \to \mathbb{R}$ to $M = \times_{i=1}^{N} M_i$ by an expected utility calculation (hence, the $u_i(s)$ are assumed to be von Neumann–Morgenstern utilities). That is, define $u_i(m_1, \ldots, m_N) = \int_{S_1} \cdots \int_{S_N} u_i(s_1, \ldots, s_N) dm_1 \ldots dm_N$ for all $m = (m_1, \ldots, m_N) \in M$. (This is an extension because we view S as a subset of M; each s A S is identified with the m A M that assigns probability one to s.) Finally, let $\overline{G} = (M_i, u_i)_{i=1}^{N}$ denote the *mixed extension* of $G = (S_i, u_i)_{i=1}^{N}$.

Definition $m^* \in M$ is a mixed strategy Nash equilibrium of $G = (S_i, u_i)_{i=1}^{N}$ if m^* is a pure strategy Nash equilibrium of the mixed extension, $\overline{G}$, of G. That is, if for every player i, $u_i(m^*) \geq u_i(m_i, m_{-i}^*)$ for every $m_i \in M_i$.

Because $u_i(m_i, m_{-i})$ is linear and therefore quasi-concave, in $m_i \in M_i$ for each $m_{-i} \in M_{-i}$, and because continuity of $u_i(\cdot)$ on S implies continuity of $u_i(\cdot)$ on M (in the weak-* topology), Theorem 1 applied to the mixed extension of G yields the following basic mixed strategy Nash equilibrium existence result:

Corollary 1 If each S_i is a non-empty compact subset of a metric space, and each $u_i(s)$ is continuous in $s \in S$, then $G = (S_i, u_i)_{i=1}^{N}$ possesses at least one mixed strategy Nash equilibrium, $m^* \in M$.

Remark Note that Corollary 1 does not require $u_i(s_i, s_{-i})$ to be quasi-concave in S_i, nor does it require the S_i to be convex.

Remark Corollary 1 yields von Neumann's (1928) classic result for two-person zero-sum games as well as Nash's (1950, 1951) seminal result for finite games as special cases. To obtain Nash's result, note that if each $\mathrm{S_i}$ is finite, then each u_i is continuous on S in the discrete metric. Hence, the corollary applies and we conclude that every finite game possesses at least one mixed strategy Nash equilibrium.

Remark To see how Theorem 1 can be applied to obtain the existence of mixed strategy equilibria in Bayesian games, see Milgrom and Weber (1985).

Remark See Glicksberg (1952) for a generalization to non-metrizable strategy spaces.

Discontinuous Games

The basic challenge one must overcome in extending equilibrium existence results from continuous games to discontinuous games is the failure of the best reply correspondence to satisfy the properties required for application of a fixed point theorem. For example, in auction or Bertrand price-competition settings, discontinuities in payoffs sometimes preclude the existence of best replies. The best reply correspondence then fails to be non-empty valued, and Glicksberg's theorem, for example, cannot be applied.

A natural technique for overcoming such difficulties is to approximate the infinite strategy spaces by a sequence of finer and finer *finite* approximations. Each of the approximating finite games is guaranteed to possess a mixed strategy equilibrium (by Corollary 1) and the resulting sequence of equilibria is guaranteed, by compactness, to possess at least one limit point. Under appropriate assumptions, the limit point is a Nash equilibrium of the original game. This technique has been cleverly employed in Dasgupta and Maskin's (1986) pioneering work, and also by Simon (1987). However, while this finite approximation technique can yield results on the existence of *mixed* strategy Nash equilibria, it is unable to produce equally general existence results for pure strategy Nash equilibria. The reason, of course, is that the approximating games, being finite, are guaranteed to possess mixed strategy, but not necessarily pure strategy, Nash equilibria. Consequently, the sequence of equilibria, and so also the limit point, cannot be guaranteed to be pure.

One might be tempted to conclude that, unlike the continuous game case where the mixed strategy result is a special case of the pure strategy result, discontinuous games require a separate treatment of pure and mixed strategy equilibria. But such a conclusion would be premature. A connection between pure and mixed strategy equilibrium existence results similar to that for continuous games can be obtained for discontinuous games by considering a different kind of approximation. Rather than approximating the infinite strategy spaces by a sequence of finite approximations, one can instead approximate the discontinuous payoff functions by a sequence of continuous payoff functions. This payoff-approximation technique is employed in Reny (1999), whose main result we now proceed to describe. All of the definitions, notation, and conventions of the previous sections remain in effect. In particular, each S_i is a subset of some metric space. (This is for simplicity of presentation only. The results to follow hold in non metrizable settings as well. See Reny, 1999.)

Better-Reply Security

Definition Player i can *secure* a payoff of $\alpha \in \mathbb{R}$ at $s \in S$ if there exists $\bar{s}_i \in S_i$, such that $u_i\left(\bar{s}_i, s'_{-i}\right) \geq \alpha_\mathrm{a}$ for all s'_{-i}close enough to $s_{\text{-}i}$.

Thus, a payoff can be secured by i at s if i has a strategy that guarantees at least that payoff even if the other players deviate slightly from s.

A pair $(s, u) \in S \times \mathbb{R}^N$ is in the closure of the graph of the vector payoff function if $u \in \mathbb{R}^N$ is the limit of the vector of player payoffs for some sequence of strategies converging to s. That is, if $u = \lim_n(u_1(s^n), \ldots, u_N(s^n))$ for some $s^n \to s$

Definition A game $G = (S_i, u_i)_{i=1}^N$ is *better-reply secure* if whenever (s^*, u^*) is in the closure of the graph of its vector payoff function and s^* is not a Nash equilibrium, some player i can secure a payoff strictly above u_i^* at s^*.

All games with continuous payoff functions are better-reply secure. This is because if (s^*,u^*) is in the closure of the graph of the vector payoff function of a continuous game, we must have $u^* = (u_1(s^*), \ldots, u_N(s^*))$. Also, if s^* is not a Nash equilibrium then some player i has a strategy $\overline{s}_i$ such that $u_i(\overline{s}_i, s_{-i}^*) > u_i(s^*)$, and continuity ensures that this inequality is maintained even if the others deviate slightly i from s^*. Consequently, player i can secure a payoff strictly above $u_i^* = u_i(s^*)$.

The import of better-reply security is that it is also satisfied in many discontinuous games. For example, Bertrand's price-competition game, many auction games, and many games of timing are better-reply secure.

Pure Strategy Nash Equilibria

The following theorem provides a pure strategy Nash equilibrium existence result for discontinuous games.

Theorem 2 (Reny, 1999). If each S_i is a non-empty, compact, convex subset of a metric space, and each $u_i(s_1, \ldots, s_N)$ is quasi-concave in s_i, then $G = (S_i, u_i)_{i=1}^N$ possesses at least one pure strategy Nash equilibrium if in addition G is better-reply secure.

Remark Theorem 1 is a special case of Theorem 2 because every continuous game is better-reply secure.

Remark A classic result due to Sion (1958) states that every two-person zero-sum game with compact strategy spaces in which player 1's payoff is upper-semi-continuous and quasi-concave in his own strategy, and lower-semi-continuous and quasi-convex in the opponent's strategy, has a value and each player has an optimal pure strategy. (Sion does not actually prove the existence of optimal strategies, but this follows rather easily from his compactness assumptions and his result that the game has a value, that is, that infsup = supinf.) It is not difficult to show that Sion's result is a special case of Theorem 2.

Remark A related result that weakens quasi-concavity but adds conditions to the sum of the players' payoffs can be found in Baye, Tian and Zhou (1993). Dasgupta and Maskin (1986) provide two interesting pure strategy equilibrium existence results, both of which require each player's payoff function to upper semi-continuous in the vector of all players' strategies.

Mixed Strategy Nash Equilibria

One easily obtains from Theorem 2 a mixed strategy equilibrium existence result (the analogue of Corollary 1) by treating each M_i as if it were player i's pure strategy set and by applying the definition of better-reply security to the mixed extension $\overline{G} = (M_i, u_i)$. This observation yields the following result.

Corollary 2 (Reny, 1999). If each $\mathrm{S_i}$ is a non-empty, compact, convex subset of a metric space, then $G = (S_i, u_i)_{i=1}^N$ possesses at least one mixed strategy Nash equilibrium if in addition its mixed extension, $\overline{G} = (M_i, u_i)$, is better-reply secure.

Remark Better-reply security of G neither implies nor is implied by better-reply security of $\overline{G}$. (See Reny, 1999, for sufficient conditions for better-reply security.)

Remark Corollary 1 is a special case of Corollary 2 because continuity of each $u_i(s)$ in $s \in S$ implies (weak-*) continuity of $u_i(m)$ in $m \in M$, which implies that the mixed extension, $\overline{G}$, is better-reply secure.

Remark Corollary 2 has as special cases the mixed strategy equilibrium existence results of Dasgupta and Maskin (1986), Simon (1987) and Robson (1994).

Remark Theorem 2 can similarly be used to obtain a result on the existence of mixed strategy equilibria in discontinuous Bayesian games by following Milgrom and Weber's (1985) seminal distributional strategy approach. One simply replaces Milgrom and Weber's payoff continuity assumption with the assumption that the Bayesian game is better-reply secure in distributional strategies. An example of this technique is provided in the next subsection.

An Application to Auctions

Auctions are an important class of economic games in which payoffs are discontinuous. Furthermore, when bidders are asymmetric, in general one cannot prove existence of equilibrium by construction, as in the symmetric case.

Consequently, an existence theorem applicable to discontinuous games is called for. Let us very briefly sketch how Theorem 2 can be applied in this case.

Consider a first-price single-object auction with N bidders. Each bidder i receives a private value $v_i \in [0, 1]$ prior to submitting a sealed bid, $b_i \geq 0$. Bidder i's value is drawn independently according to the continuous and positive density f_i. The highest bidder wins the object and pays his bid. Ties are broken randomly and equiprobably. Losers pay nothing.

Because payoffs are not quasi-concave in own bids, one cannot appeal directly to Theorem 2 to establish the existence of an equilibrium in pure strategy bidding functions. On the other hand, it is not difficult to show that all mixed strategy equilibria are pure and non-decreasing. Hence, to obtain an existence result for pure strategies, it suffices to show that there is an equilibrium in mixed, or equivalently in distributional, strategies. (In this context, a distributional strategy for bidder i is a joint probability distribution over his values and bids with the property that the marginal density over his values is f_i ; see Milgrom and Weber, 1985.)

Because the set of distributional strategies for each bidder is a non-empty compact convex metric space and each bidder's payoff is linear in his own distributional strategy, Theorem 2 can be applied so long as a first-price auction game in distributional strategies is better-reply secure. Better-reply security can be shown to hold by using the facts that payoff discontinuities occur only when there are ties in bids and that bidders can always break a tie in their favour by increasing their bid slightly. Consequently, a Nash equilibrium in distributional strategies exists and, as mentioned above, this equilibrium is pure and non-decreasing.

Endogenous Sharing Rules

Discontinuities in payoffs sometimes arise endogenously. For example, consider a political game in which candidates first choose a policy from the interval [0,1] and each voter among a continuum then decides for whom to vote. Voters vote for the candidate whose policy they most prefer, and if there is more than one such candidate it is conventional to assume that voters randomize equiprobably over them. The behaviour of voters in the second stage can induce discontinuities in the payoffs of the candidates in the first stage since a candidate can discontinuously gain or lose a positive fraction of votes by choosing a policy that, instead of being identical to another candidate's policy, is just slightly different from it.

Simon and Zame (1990) suggest an elegant way to handle such discontinuities. In particular, for the political game example above, they would not insist that voters, when indifferent, randomize equiprobably. Indeed, applying subgame perfection to the two-stage game would permit voters to randomize in any manner whatsoever over those candidates whose policies they most prefer. With this in mind, if s is a joint pure strategy for the N candidates specifying a location for each, let us denote by $U(s)$ the resulting *set* of payoff vectors for the N candidates when all best replies of the voters are considered. If no voter is indifferent, then $U(s)$ contains a single payoff vector. On the other hand, if some voters are indifferent (as would be the case if two or more candidates chose the same location) and $U(s)$ is not a singleton, then distinct payoff vectors in $U(s)$ correspond to different ways the indifferent voters can randomize between the candidates among whom they are indifferent.

The significance of the correspondence $U(\cdot)$ is this. Suppose that we are able to select, for each s, a payoff vector $u(s) \in U(s)$ in such a way that some joint mixed strategy m^* for the N candidates is a Nash equilibrium of the induced policy-choice game between them when their vector payoff function is $u(\cdot)$. Then m^* together with the voter behaviour that is implicit in the definition of $u(s)$ for each s, constitutes a subgame perfect equilibrium of the original two-stage game. Thus, solving the original problem with potentially endogenous discontinuities boils down to obtaining an appropriate selection from $U(\,\cdot\,)$. Simon and Zame (1990) provide a general result concerning the existence of such selections, which they refer to as 'endogenous sharing rules'. This method therefore provides an additional tool for obtaining equilibrium existence when discontinuities are present. Simon and Zame's main result is as follows.

Theorem 3 (Simon and Zame, 1990). Suppose that each S_i is a compact subset of a metric space and that U: $\mathrm{S}^2 \twoheadrightarrow \mathbb{R}^N$ is a bounded, upper hemi-continuous, non-empty-valued, convex-valued correspondence. Then for each player i, there is a measurable payoff function, $u_i : S \rightarrow \mathbb{R}$, *such that* $(u_1(s), \ldots, u_N(s)) \in U(s)$ for every $s \in S$ and such that the game $(S_i, u_i)_{i=1}^N$ possesses at least one mixed strategy Nash equilibrium.

Remark Theorem 3 applies to the political game example above because for any policy choice s of the N candidates, the resulting set of payoff vectors $U(s)$ is convex, a fact that follows from the presence of a continuum of voters. It can also be shown that, as a correspondence, $U(\,\cdot\,)$ is upper hemi-continuous.

Remark In the context of Bayesian games, an even more subtle endogenous-sharing rule result can be found in Jackson et al. (2002). This result, too, can be very helpful in dealing with discontinuous games. Indeed, Jackson and Swinkels (2005) have shown how it can be used to obtain equilibrium existence results in a variety of auction settings, including double auctions.

See Also

▶ Auctions (theory)
▶ Epistemic game theory: incomplete information
▶ Fixed point theorems
▶ Mathematical methods in political economy
▶ Spatial economics
▶ Strategic and extensive form games

Bibliography

Baye, M., G. Tian, and J. Zhou. 1993. Characterizations of the existence of equilibria in games with discontinuous and non-quasiconcave payoffs. *Review of Economic Studies* 60: 935–948.

Bertrand, J. 1883. Théorie mathématique de la richesse sociale. *Journal des Savants* 67: 499–508.

Billingsley, P. 1968. *Convergence of probability measures*. New York: John Wiley and Sons.

Cournot, A. 1838. In *Researches into the mathematical principles of the theory of wealth*, ed. N. Bacon, 1897. New York: Macmillan.

Dasgupta, P., and E. Maskin. 1986. The existence of equilibrium in discontinuous economic games, I: Theory. *Review of Economic Studies* 53: 1–26.

Debreu, G. 1952. A social equilibrium existence theorem. *Proceedings of the National Academy of Sciences* 38: 386–393.

Fudenberg, D., R. Gilbert, J. Stiglitz, and J. Tirole. 1983. Preemption, leapfrogging, and competition in patent races. *European Economic Review* 22: 3–31.

Glicksberg, I. 1952. A further generalization of the Kakutani fixed point theorem. *Proceedings of the American Mathematical Society* 3: 170–174.

Hotelling, H. 1929. The stability of competition. *Economic Journal* 39: 41–57.

Jackson, M., and J. Swinkels. 2005. Existence of equilibrium in single and double private value auctions. *Econometrica* 73: 93–139.

Jackson, M., L. Simon, J. Swinkels, and W. Zame. 2002. Communication and equilibrium in discontinuous games of incomplete information. *Econometrica* 70: 1711–1740.

Milgrom, P., and J. Roberts. 1990. Rationalizability, learning, and equilibrium in games with strategic complementarities. *Econometrica* 58: 1255–1277.

Milgrom, P., and R. Weber. 1982. A theory of auctions and competitive bidding. *Econometrica* 50: 1089–1122.

Milgrom, P., and R. Weber. 1985. Distributional strategies for games with incomplete information. *Mathematics of Operations Research* 10: 619–632.

Nash, J. 1950. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences* 36: 48–49.

Nash, J. 1951. Non-cooperative games. *Annals of Mathematics* 54: 286–295.

Osborne, M., and A. Rubinstein. 1994. *A course in game theory*. Cambridge, MA: MIT Press.

Reny, P. 1999. On the existence of pure and mixed strategy Nash equilibria in discontinuous games. *Econometrica* 67: 1029–1056.

Robson, A. 1994. An 'informationally robust' equilibrium in two-person nonzero-sum games. *Games and Economic Behavior* 2: 233–245.

Schafer, W., and H. Sonnenschein. 1975. Equilibrium in Abstract economies without ordered preferences. *Journal of Mathematical Economics* 2: 345–348.
Simon, L. 1987. Games with discontinuous payoffs. *Review of Economic Studies* 54: 569–597.
Simon, L., and W. Zame. 1990. Discontinuous games and endogenous sharing rules. *Econometrica* 58: 861–872.
Sion, M. 1958. On general minimax theorems. *Pacific Journal of Mathematics* 8: 171–176.
Vives, X. 1990. Nash equilibrium with strategic complementarities. *Journal of Mathematical Economics* 19: 305–321.
von Neumann, J. 1928. Zur Theorie der Gesellshaftspiele. *Mathematische Annalen* 100, 295–320. Trans. S. Bargmann [On the theory of games of strategy]. In *Contributions to the theory of games*, vol. 4, ed. R. Luce and A. Tucker, Princeton: Princeton University Press, 1959.

Non-expected Utility Theory

Mark J. Machina

Abstract

Beginning with the work of Allais and Edwards in the early 1950s and continuing through the present, psychologists and economists have uncovered a growing body of evidence that individuals do not necessarily conform to many of the key assumptions or predictions of the expected utility model of choice under uncertainty, and seem to depart from this model in systematic and predictable ways. This has led to the development of alternative models of preferences over objectively or subjectively uncertain prospects, which seek to accommodate these systematic departures from the expected utility model while retaining as much of its analytical power as possible.

Keywords

Allais Paradox; Allais, M.; Ambiguity aversion; Asset demand theory; Choquet expected utility model; Common consequence effect; Common ratio effect; Comparative statics; Ellsberg Paradox; Expected utility hypothesis; First-order stochastic dominance preference; Independence Axiom; Insurance; Non-expected utility theory; Objective vs. subjective uncertainty; Regret theory; Risk; Risk aversion; Stochastic dominance; Transitivity; Uncertainty; von Neumann–Morgenstern utility function

JEL Classifications

D1; D8

Although the expected utility model has long been the standard theory of individual choice under objective and subjective uncertainty, experimental work by both psychologists and economists has uncovered systematic departures from the expected utility hypothesis, which has led to the development of alternative models of preferences over uncertain prospects.

The Expected Utility Model

In one of the simplest settings of choice under economic uncertainty, the objects of choice consist of finite-outcome *objective lotteries* of the form $\mathbf{P} = (x_1, p_1;\ldots; x_n, p_n)$, yielding a monetary payoff of x_i with probability p_i, where $p_1 + \ldots + p_n = 1$. In such a case, the expected utility model of risk preferences assumes (or posits axioms sufficient to imply) that the individual ranks these prospects on the basis of an *expected utility preference function* of the form

$$\begin{aligned} V_{EU}(\mathbf{P}) &\equiv V_{EU}(x_1, p_1, \ldots, x_n, p_n) \\ &\equiv U(x_1)\cdot p_1 + \ldots + U(x_n)\cdot p_n \end{aligned}$$

in the standard economic sense that the individual prefers lottery $\mathbf{P}^* = \left(x_1^*, p_1^*;\ldots; x_n^*, p_{n^*}^*\right)$ over lottery $\mathbf{P} = (x_1, p_1;\ldots;;x_n, p_n)$ if and only if $V_{EU}(\mathbf{P}^*) > V_{EU}(\mathbf{P})$, and is indifferent between them if and only if $V_{EU}(\mathbf{P}^*) = V_{EU}(\mathbf{P})$. $U(\cdot)$ is termed the individual's *von Neumann–Morgenstern utility function* (von Neumann and Morgenstern 1944, 1947, 1953) and its various mathematical properties serve to characterize various features of the individual's attitudes toward risk, for example:

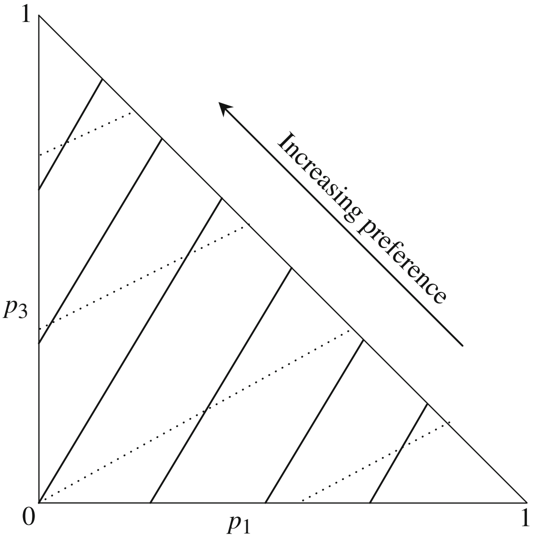
- $V_{EU}(\,\cdot\,)$ exhibits *first-order stochastic dominance preference* (a preference for shifting probability from lower to higher outcome values) if and only if $U(x)$ is an increasing function of x.
- $V_{EU}(\,\cdot\,)$ exhibits *risk aversion* (an aversion to all mean-preserving increases in risk) if and only if $U(x)$ is a concave function of x.
- $V^*_{EU}(\cdot)$ is *at least as risk averse as* $V_{EU}(\,\cdot\,)$ (in several equivalent senses) if and only if its utility function $U^*(\,\cdot\,)$ is a concave transformation of $U(\,\cdot\,)$ (that is, if and only if $U^*(x) \equiv \rho(U(x))$ for some increasing concave function $\rho(\,\cdot\,)$).

As shown by Bernoulli (1738), Arrow (1965), Pratt (1964), Friedman and Savage (1948), Markowitz (1952) and others, this model admits of a tremendous flexibility in representing attitudes towards risk, and can be applied to many types of economic decisions and markets.

But in spite of its flexibility, the expected utility model has testable implications which hold regardless of the shape of the utility function $U(\,\cdot\,)$, since they follow from the *linearity in the probabilities* property of the preference function $V_{EU}(\,\cdot\,)$. These implications can be best expressed by the concept of an $\alpha : (1-\alpha)$ *probability mixture* of two lotteries $\mathbf{P} = (x_1, p_1;...; x_n, p_n)$ and $\mathbf{P}^* = \left(x^*_1, p^*_1;...; x^*_n, p^*_{n^*}\right)$, which is defined as the single-stage lottery $\alpha \cdot \mathbf{P} + (1-\alpha)\cdot \mathbf{P}^* = (x_1, \alpha \cdot p_1;...; x_n, \alpha \cdot p_n; x^*_1, (1-\alpha)\cdot p^*_1;...; x^*_{n^*}, (1-\alpha)\cdot p^*_{n^*})$. The mixture $\alpha \cdot \mathbf{P} + (1-\alpha)\cdot \mathbf{P}^*$ can be thought of as a coin flip yielding lotteries P and P* with probabilities $\alpha : (1-\alpha)$, where the uncertainty in the coin and in the subsequent lottery is resolved simultaneously. Linearity in the probabilities is equivalent to the following property, which serves as the key foundational axiom of the expected utility model (Marschak 1950):

Independence Axiom If lottery **P*** is preferred (indifferent) to lottery **P**, then the probability mixture $\alpha \cdot \mathbf{P}^* + (1-\alpha)\cdot \mathbf{P}^{**}$ is preferred (indifferent) to $\alpha \cdot \mathbf{P} + (1-\alpha)\cdot \mathbf{P}^{**}$ for every lottery **P**** and every mixture probability $\alpha \in (0, 1]$.

This axiom can be interpreted as saying 'given an $\alpha : (1-\alpha)$ coin, the individual's preferences for receiving **P*** versus **P** in the event of a head should not depend upon the prize **P**** that would be received in the event of a tail, nor upon the probability α of landing heads (so long as this probability is positive)'. The strong normative appeal of this axiom has contributed to the widespread adoption of the expected utility model.



Non-expected Utility Theory, Fig. 1 Expected utility indifference curves in the probability triangle

The property of linearity in the probabilities, as well as the senses in which it has been found to be empirically violated, can be illustrated in the special case of preferences over all lotteries $\mathbf{P} = (\bar{x}_1, p_1; \bar{x}_2, p_2; \bar{x}_3, p_3)$ over a fixed set of outcome values $\bar{x}_1 < \bar{x}_2 < \bar{x}_3$. Since we must have $p_2 = 1 - p_1 - p_3$, each such lottery can be completely summarized by its pair of probabilities (p_1, p_3), as plotted in the 'probability triangle' of Fig. 1. Since upward movements in the diagram (increasing p_3 for fixed p_1) represent shifting probability from outcome x_2 up to x_3, and leftward movements represent shifting probability from x_1 up to x_2, such movements constitute first-order stochastically dominating shifts and will thus always be preferred. Expected utility indifference curves (loci of constant expected utility) are given by the formula

$$U(\bar{x}_1)\cdot p_1 + U(\bar{x}_2)\cdot[1 - p_1 - p_3] + U(\bar{x}_3)\cdot p_3 = \text{constant}$$

and are thus seen to be parallel straight lines of slope $[U(\bar{x}_2) - U(\bar{x}_1)]/[U(\bar{x}_3) - U(\bar{x}_2)]$, as indicated by the solid lines in the figure. The dotted lines in Fig. 1 are loci of constant *expected value*, given by the formula $\bar{x}_1 \cdot p_1 + \bar{x}_2 \cdot [1 - p_1 - p_3] + \bar{x}_3 \cdot p_3 = \text{constant}$, with slope $[\bar{x}_2 - \bar{x}_1]/[\bar{x}_3 - \bar{x}_2]$. Since north-east movements along the constant expected value lines shift probability from x_2 down to x_1 and up to x_3 in a manner that preserves the mean of the distribution, they represent simple increases in risk (Rothschild and Stiglitz 1970, 1971). When $U(\cdot)$ is concave (that is, risk averse), its indifference curves will have a steeper slope than these constant expected value lines, and such increases in risk move the individual from more to less preferred indifference curves, as illustrated in the figure. It is straightforward to show that the indifference curves of any expected utility maximizer with a more risk-averse (that is, more concave) utility function $U^*(\cdot)$ will be steeper than those generated by $U(\cdot)$.

Systematic Violations of the Expected Utility Hypothesis

In spite of its normative appeal, researchers have uncovered several types of widespread systematic violations of the expected utility model and its underlying assumptions. These can be categorized into (*a*) violations of the Independence Axiom (such as the common consequence and common ratio effects), (*b*) violations of the hypothesis of probabilistic beliefs (such as the Ellsberg Paradox) and (*c*) violations of the model's underlying assumptions of descriptive and procedural invariance (such as reference-point and response-mode effects).

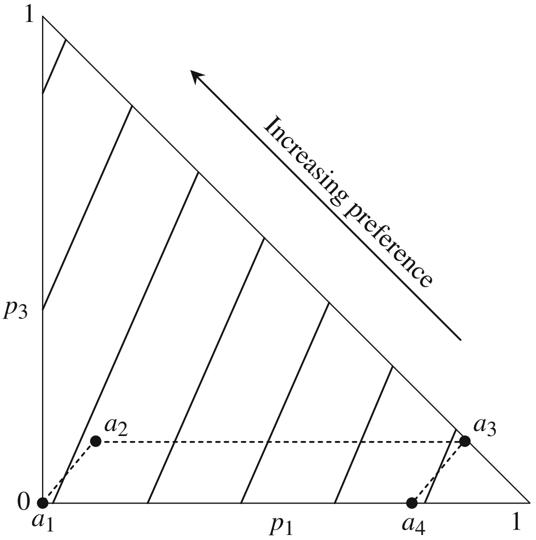
Violations of the Independence Axiom

The best-known violation of the Independence Axiom is the so-called *Allais Paradox*, in which individuals are asked to rank the lotteries in each of the following pairs, where \$1 M = \$1; 000, 000:

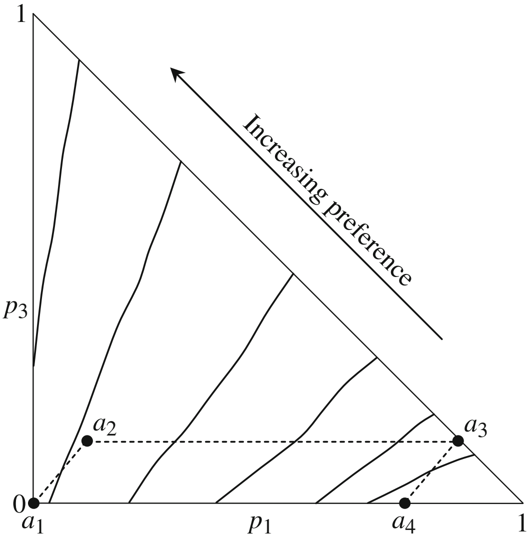
$$a_1 : \begin{cases} 1.00 \text{ chance of \$1M} \end{cases} \quad \text{versus} \quad a_2 : \begin{cases} .10 \text{ chance of \$5M} \\ .89 \text{ chance of \$1M} \\ .01 \text{ chance of \$0} \end{cases}$$

$$a_3 : \begin{cases} .10 \text{ chance of \$5M} \\ .90 \text{ chance of \$0} \end{cases} \quad \text{versus} \quad a_4. : \begin{cases} .11 \text{ chance of \$1M} \\ .89 \text{ chance of \$0} \end{cases}$$

Researchers such as Allais (1953), Morrison (1967), Raiffa (1968), Slovic and Tversky (1974) and others have found that the modal if not majority preference of subjects is for a_1 over a_2 in the first pair of choices and for a_3 over a_4 in the second pair. However, such preferences violate expected utility, since the first ranking implies the inequality $U(\$1\text{ M}) > .10 \cdot U(\$5\text{ M}) + .89 \cdot U(\$1\text{ M}) + .01 \cdot U(\$0)$ whereas the second implies the inconsistent inequality $..10 \cdot U(\$5\text{ M}) + .90 \cdot U(\$0) > .11 \cdot U(\$1\text{ M}) + .89 \cdot U(\$0)$. By setting $\bar{x}_1 = \$0, \bar{x}_2 = \$1\text{M}$ and $\bar{x}_3 = \$5\text{M}$, the lotteries a_1, a_2, a_3 and a_4 are seen to form a parallelogram when plotted in the probability triangle (Fig. 2), which explains why the parallel straight line indifference curves of an expected utility maximizer must either prefer a_1 and a_4 (as illustrated for the relatively steep indifference curves of the figure) or else prefer a_2 and a_3 (for relatively flat indifference curves). Fig. 3 illustrates *non-expected utility indifference curves* which *fan out*, and are seen to exhibit the typical Allais Paradox rankings of a_1 over a_2 and a_3 over a_4. Although the Allais Paradox was originally dismissed as an isolated example, subsequent experimental work by psychologists, economists and others have uncovered a similar pattern of violations over a range of probability and payoff values, and the Allais Paradox is now seen to be a special case of a widely observed phenomenon known as the *common consequence effect.* This effect involves pairs of prospects (probability mixtures) of the form:



Non-expected Utility Theory, Fig. 2 Expected utility indifference curves and the Allais Paradox choices



Non-expected Utility Theory, Fig. 3 Allais Paradox choices and indifference curves which 'fan out'

$$b_1 : \begin{cases} \alpha & \text{chance of} \quad x \\ 1-\alpha & \text{chance of} \quad \mathbf{P}^{**} \end{cases} \quad \text{versus}$$

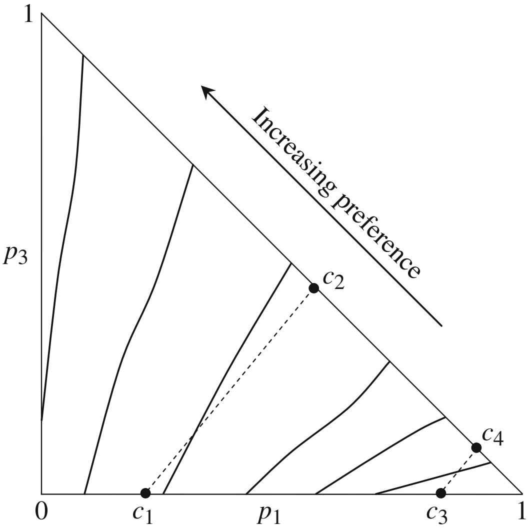
$$b_2 : \begin{cases} \alpha & \text{chance of} \quad \mathbf{P} \\ 1-\alpha & \text{chance of} \quad \mathbf{P}^{**} \end{cases}$$

$$b_3 : \begin{cases} \alpha & \text{chance of} \quad x \\ 1-\alpha & \text{chance of} \quad \mathbf{P}^{*} \end{cases} \quad \text{versus}$$

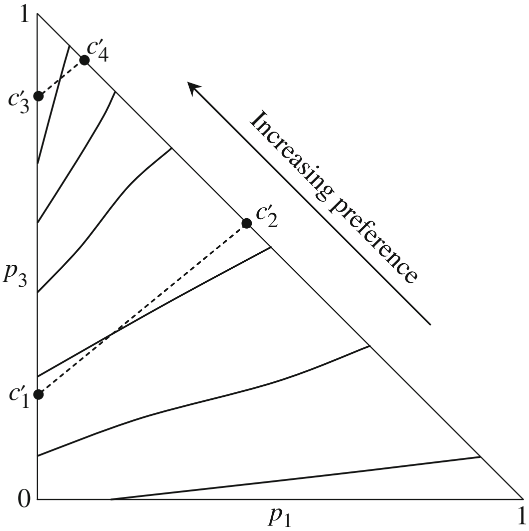
$$b_4 : \begin{cases} \alpha & \text{chance of} \quad \mathbf{P} \\ 1-\alpha & \text{chance of} \quad \mathbf{P}^{*} \end{cases}$$

where the lottery **P** involves outcomes both greater and less than the amount x, and **P**** first order stochastically dominates **P*** (in Allais's example, $x = \$1\text{M}$, $\mathbf{P} = (\$5\text{M}, 10 = 11, \$0, 1 = 11)$, $\mathbf{P}^* = (\$0, 1)$, $\mathbf{P}^{**} = (\$1\text{M}, 1)$ and $\alpha = .11$). Although the Independence Axiom clearly implies choices of either b_1 and b_3 (if x is preferred to **P**) or else b_2 and b_4 (if **P** is preferred to x), researchers have found a tendency for subjects to choose b_1 in the first pair and b_4 in the second. When the distributions **P**, **P*** and **P**** are each over a common outcome set $\{\bar{x}_1\ ,\ \bar{x}_2\ ,\ \bar{x}_3\}$ with $\bar{x}_2\ = x$, the prospects $\{b_1, b_2, b_3, b_4\}$ again form a parallelogram in the (p_1, p_3) triangle, and a choice of b_1 and b_4 again implies indifference curves which fan out.

The intuition behind this phenomenon can be described in terms of the above 'coin-flip' scenario. According to the Independence Axiom, one's preferences over what would occur in the event of a head ought not depend upon what would occur in the event of a tail. However, they *may well* depend upon what would otherwise happen (as Bell 1985, p. 1, notes, 'winning the top prize of $10,000 in a lottery may leave one much happier than receiving $10,000 as the lowest prize in a lottery'). The common consequence effect states that the *better off* individuals would be in the event of a tail (in the sense of stochastic dominance), the *more risk averse* their preferences over what they would receive in the event of a head. That is, if the distribution **P**** in the pair $\{b_1, b_2\}$ involves very high outcomes, one may prefer not to bear further risk in the unlucky event that one doesn't receive it, and hence opt for the sure outcome x over the risky distribution **P** (that is, choose b_1 over b_2). But, if **P*** in $\{b_3, b_4\}$ involves very low outcomes, one might be more willing to bear risk in the lucky event that one doesn't receive it, and prefer going for the lottery P rather than the sure outcome x (choose b_4 over b_3).



Non-expected Utility Theory, Fig. 4 Common ratio effect and fanning out indifference curves



Non-expected Utility Theory, Fig. 5 Common ratio effect for losses and fanning out indifference curves

A second type of systematic violation of linearity in the probabilities, also noted by Allais and subsequently termed the *common ratio effect*, involves prospects of the form:

$$c_1 : \begin{cases} p & \text{chance of} \quad \$X \\ 1-p & \text{chance of} \quad \$0 \end{cases} \quad \text{versus}$$

$$c_2 : \begin{cases} q & \text{chance of} \quad \$Y \\ 1-q & \text{chance of} \quad \$0 \end{cases}$$

$$c_3 : \begin{cases} \alpha \cdot p & \text{chance of} \quad \$X \\ 1-\alpha \cdot p & \text{chance of} \quad \$0 \end{cases} \quad \text{versus}$$

$$c_4 : \begin{cases} \alpha \cdot q & \text{chance of} \quad \$Y \\ 1-\alpha \cdot q & \text{chance of} \quad \$0 \end{cases}$$

where $p > q$, $0 < X < Y$and $\alpha \in (0, 1)$. (The term 'common ratio effect' comes from the common value of prob(\$$X$)/prob(\$$Y$) in the upper and lower pairs.) Setting $\{\bar{x}_1, \bar{x}_2, \bar{x}_3\} = \{\$0, \$X, \$Y\}$ and plotting these prospects in the probability triangle as in Fig. 4, the line segments $\overline{c_1c_2}$ and $\overline{c_3c_4}$ are seen to be parallel, so that the expected utility model again predicts choices of c_1 and c_3 (if the indifference curves are relatively steep) or else c_2 and c_4 (if they are flat). However, experimental studies by MacCrimmon (1968), Tversky (1975), MacCrimmon and Larsson (1979), Kahneman and Tversky (1979), Hagen (1979), Chew and Waller (1986) and others have found a systematic tendency for choices to depart from these predictions in the direction of preferring c_1 over c_2 and c_4 over c_3, which again suggests that indifference curves fan out, as in the figure. For example, Kahneman and Tversky (1979) found that, while 86 per cent of their subjects preferred a .90 chance of winning \$3,000 to a .45 chance of \$6,000, 73 per cent preferred a .001 chance of \$6,000 to a .002 chance of \$3,000. Kahneman and Tversky (1979) observed that, when the positive outcomes \$3000 and \$6000 in the above gambles are replaced by *losses* of these magnitudes, to obtain the lotteries c'_1, c'_2, c'_3 and c'_4, preferences typically 'reflect,' to prefer c'_2over c'_1 and c'_3 over c'_4. Setting $\bar{x}_1 = -\$6000$, $\bar{x}_2 = -\$3000$ and $\bar{x}_3 = -\$0$ (to preserve the ordering $\bar{x}_1 < \bar{x}_2 < \bar{x}_3$) and plotting as in Fig. 5, such preferences again suggest that indifference curves in the probability triangle fan out. Battalio et al. (1985) found that laboratory rats choosing among gambles involving substantial variations in their daily food intake also exhibited this pattern of choices.

One criticism of this evidence has been that individuals whose initial choices violated the

Independence Axiom in the above manners would typically ‘correct’ themselves once the nature of their violations was revealed by an application of the above type of coin-flip argument. Thus, while even Leonard Savage chose a_1 and a_3 when first presented with such choices by Allais, he concluded upon reflection that these preferences were in error (Savage 1954, pp. 101–3). Although Moskowitz found that allowing subjects to discuss opposing written arguments led to a decrease in the proportion of violations, 73 per cent of the initial fanning-out type choices remained unchanged after the discussions (1974, pp. 232–7, Table 6). When written arguments were presented but no discussion was allowed, there was a 93 per cent persistency rate of such choices (1974, p. 234, Tables 4 and 6). In experiments where subjects who responded to Allais-type problems were then presented with written arguments both for *and against* the expected utility position, neither MacCrimmon (1968), Moskowitz (1974) nor Slovic and Tversky (1974) found predominant net swings toward the expected utility choices.

Further descriptions of these and other violations of the Independence Axiom can be found in Camerer (1989), Machina (1983, 1987), Starmer (2000), Sugden (1986) and Weber and Camerer (1987).

Non-existence of Probabilistic Beliefs

Although the expected utility model was first formulated in terms of preferences over *objective lotteries* $\mathbf{P} = (x_1, p_1;\ldots; x_n, p_n)$ with pre-specified probabilities, it has also been applied to preferences over *subjective acts* $f(\cdot) = [x_1 \text{ on } E_1;\ldots; x_n \text{ on } E_n]$, where the uncertainty is represented by a set $\{E_1;\ldots; E_n\}$ of mutually exclusive and exhaustive *events* (such as the alternative outcomes of a horse race) (Savage 1954). As long as an individual possesses well-defined *subjective probabilities* $\mu(E_1);\ldots;\mu(E_n)$ over these events, their *subjective expected utility* preference function takes the form

$$W_{SEU}(f(\cdot)) \equiv W_{SEU}(x_1 \text{ on } E_1;\ldots;x_n \text{ on } E_n)$$
$$\equiv \mathrm{U}(x_1)\cdot\mu(E_1)+\ldots+U(x_n)\cdot\mu(E_n).$$

However, researchers have found that individuals may not possess such well-defined subjective probabilities, in even the simplest of cases. The best-known example of this is the *Ellsberg Paradox* (Ellsberg 1961), in which the individual must draw a ball from an urn that contains 30 red balls, and 60 black or yellow balls in an unknown proportion, and is offered the following bets based on the colour of the drawn ball:

	30 balls	60 balls	
	Red	Black	Yellow
$f_1(\cdot)$	\$100	\$0	\$0
$f_2(\cdot)$	\$0	\$100	\$0
$f_3(\cdot)$	\$100	\$0	\$100
$f_4(\cdot)$	\$0	\$100	\$100

Most individuals exhibit a preference for $f_1(\cdot)$ over $f_2(\cdot)$ and $f_4(\cdot)$ over $f_3(\cdot)$. When asked, they explain that the chance of winning under $f_2(\cdot)$ could be anywhere from 0 to 2/3 whereas under $f_1(\cdot)$ it is known to be exactly 1/3, and they prefer the bet that offers the known probability. Similarly, the chance of winning under $f_3(\cdot)$ could be anywhere from 1/3 to 1 whereas under $f_4(\cdot)$ it is known to be exactly 2/3, so the latter is preferred. However, such preferences are inconsistent with any assignment of subjective probabilities μ(red), μ(black), μ(yellow) to the three events. If the individual were to be choosing on the basis of such probabilistic beliefs, the choice of $f_1(\cdot)$ over $f_2(\cdot)$ would ‘reveal’ that μ(red) > μ(black), but the choice of $f_4(\cdot)$ over $f_3(\cdot)$ would reveal that μ(red) < μ(black). A preference for gambles based on probabilistic partitions such as {red, black ∪ yellow} over gambles based on subjective partitions such as {black, red ∪ yellow} is termed *ambiguity aversion*.

In an even more basic example, Ellsberg presented subjects with a pair of urns, the first containing 50 red balls and 50 black balls, and the second with 100 red and black balls in an unknown proportion. When asked, a majority of subjects strictly preferred to stake a prize on drawing red from the first urn over drawing red from the second urn, and strictly preferred staking the prize on drawing black from the first urn over drawing black from the second. It is clear that there can exist no subjective probabilities p:(1 − p) of red:black in the second urn, including 1/2:1/2, which can simultaneously generate both of these strict preferences. Similar behaviour in this and related problems has been observed by Raiffa (1961), Becker and Brownson (1964), MacCrimmon (1965), Slovic and Tversky (1974) and MacCrimmon and Larsson (1979).

Violations of Descriptive and Procedural Invariance

Researchers have also uncovered several systematic violations of the standard economic assumptions of stability of preferences and invariance with respect to problem description in choices over risky prospects. In particular, psychologists have found that alternative means of representing or *framing* probabilistically equivalent choice problems lead to systematic differences in choice. Early examples of this were reported by Slovic (1969), who found that offering a gain or loss contingent on the joint occurrence of four independent events with probability *p* elicited different responses than offering it on the occurrence of a single event with probability p_4 (all probabilities were stated explicitly). In comparison with the single-event case, making a gain contingent on the joint occurrence of events was found to make it more attractive, and making a loss contingent on the joint occurrence of events made it more unattractive.

One class of framing effects exploits the phenomenon of a *reference point*. According to economic theory, the variable which enters an individual's von Neumann–Morgenstern utility function should be total (that is, final) wealth, and gambles phrased in terms of gains and losses should be combined with current wealth and re-expressed as distributions over final wealth levels before being evaluated. However, risk attitudes towards gains and losses tend to be more stable than can be explained by a fixed utility function over final wealth, and utility functions might be best defined in terms of changes from the *reference point* of current wealth. In his discussion of this phenomenon, Markowitz (1952, p. 155) suggested that certain circumstances may cause the individual's reference point to temporarily deviate from current wealth. If these circumstances include the manner in which a problem is verbally described, then differing risk attitudes towards gains and losses from the reference point can lead to different choices, depending upon the exact description of an otherwise identical problem. A simple example of this, from Kahneman and Tversky (1979, p. 273), involves the following two choices:

In addition to whatever you own, you have been given 1,000 (Israeli pounds). You are now asked to choose between a 1/2:1/2 chance of a gain of 1,000 or 0 or a sure chance of a gain of 500.

and

In addition to whatever you own, you have been given 2,000. You are now asked to choose between a 1/2:1/2 chance of a loss of 1,000 or 0 or a sure loss of 500.

These two problems involve identical distributions over final wealth. But, when put to two different groups of subjects, 84 per cent chose the sure gain in the first problem but 69 per cent chose the 1/2:1/2 gamble in the second.

In another class of examples, not based on reference point effects, Moskowitz (1974), Keller (1985) and Carlin (1990) found that the proportion of subjects choosing in conformity with the Independence Axiom in examples like the Allais Paradox was significantly affected by whether the problems were described in the standard matrix form, decision tree form, roulette wheels, or as minimally structured written statements. Interestingly, the form judged the 'clearest representation' by the majority of Moskowitz's subjects (the tree form) led to the lowest degree of consistency with the Independence Axiom, the highest proportion of Allais-type (fanning out) choices, and the highest persistency rate of these choices Moskowitz (1974, pp. 234, 237–8).

In other studies, Schoemaker and Kunreuther (1979), Hershey and Schoemaker (1980), Kahneman and Tversky (1982, 1984), and Slovic et al. (1977) found that subjects' choices in otherwise identical problems depended upon whether they were phrased as decisions whether or not to gamble as opposed to whether or not to insure, whether statistical information for different therapies was presented in terms of cumulative survival probabilities or cumulative mortality probabilities, and so on (see the references in Tversky and Kahneman 1981).

Whereas framing effects involve alternative *descriptions* of an otherwise identical choice problem, alternative *response formats* have also been found to lead to different choices, leading to what have been termed *response-mode effects*. For example, under expected utility, an individual's von Neumann–Morgenstern utility function can be assessed or elicited in a number of different manners, which typically involve a sequence of pre-specified lotteries $\mathbf{P}_1$, $\mathbf{P}_2$, $\mathbf{P}_3$, .. .; and ask for (*a*) the individual's certainty equivalent $CE(\mathbf{P}_i)$ of each lottery P_i, (*b*) the *gain equivalent* G_i that would make the gamble (G_i,1/2, \$0,1/2) indifferent to $\mathbf{P}_i$, or (*c*) the *probability equivalent* $\wp_i$ that would make the gamble (\$1000, $\wp_i$, \$0,1 − $\wp_i$) indifferent to P_i. Although such procedures should generate equivalent assessed utility functions, they have been found to yield systematically different ones (for example, Hershey et al. 1982; Hershey and Schoemaker 1985).

In a separate finding now known as the *preference reversal phenomenon*, subjects were first presented with a number of pairs of lotteries and asked to make one choice out of each pair. Each pair of lotteries took the following form:

$$p-\text{bet}:\begin{cases} p & \text{chance of} \quad \$X \\ 1-p & \text{chance of} \quad \$0 \end{cases} \quad \text{versus}$$

$$\$-\text{bet}:\begin{cases} q & \text{chance of} \quad \$Y \\ 1-q & \text{chance of} \quad \$0 \end{cases}$$

where $0 < X < Y$ and $p > q$. The terms '*p*-bet' and '\$-bet' derive from the greater probability of winning in the first bet, and greater possible gain in the second bet. Subjects were next asked for their certainty equivalents of each of these bets, via a number of standard elicitation techniques. Standard theory predicts that, for each such pair, the prospect selected in the direct choice problem would also be assigned the higher certainty equivalent. However, subjects exhibit a systematic departure from this prediction in the direction of choosing the *p*-bet in a direct choice, but assigning a higher certainty equivalent to the \$-bet (Lichtenstein and Slovic 1971). Although this finding initially generated widespread scepticism, it has been replicated by both psychologists and economists in a variety of settings involving real-money gambles, patrons in a Las Vegas casino, group decisions and experimental market trading. By expressing the implied preferences as '\$-bet $\sim CE$ (\$-bet) $\succ CE(p\text{-bet}) \sim p$-bet $\succ$ \$-bet', some economists have categorized this phenomenon as a violation of transitivity and tried to model it as such (see the 'regret theory' model below). However, most psychologists and economists now view it as a response-mode effect: specifically, that the psychological processes of valuation (which generates certainty equivalents) and direct choice are differentially influenced by the probabilities and payoffs involved in a lottery, and that under certain conditions this can lead to choices and valuations which 'reveal' opposite preference rankings over a pair of gambles.

Non-expected Utility Models of Risk Preferences

Non-expected Utility Functional Forms

Researchers have responded to departures from linearity in the probabilities in two manners. The first consists of replacing the expected utility form $V_{EU}\,(\mathbf{P}) = U\,(x_1)\cdot\, p_1 + \ldots + U\,(x_n)\cdot\, p_n$ by some more general form for the preference function $V\,(\mathbf{P}) = V\,(x_1,\, p_1;\ldots;\, x_n,\, p_n)$. Several such forms have been proposed (for the Rank Dependent, Dual and Ordinal Independence forms, the payoffs must be labelled so that $x_1 \le \ldots \le x_n$, and $G(\,\cdot\,)$ must satisfy $G(0) = 0$ and $G(1) = 1$):

Most of these forms have been formally axiomatized, and, under the appropriate monotonicity and/or curvature assumptions on their constituent functions $\upsilon(\,\cdot\,)$, $G(\,\cdot\,)$, and so on, most are capable of exhibiting first-order stochastic dominance preference, risk aversion, and the above types of systematic violations of the Independence Axiom. Researchers such as Konrad and

Non-expected Utility Theory, Table 1

Prospect theory	$\sum_{i=1}^{n} \upsilon(x_i) \cdot \pi(p_i)$	Edwards (1955, 1962), Kahneman and Tversky (1979)
Subjectively weighted utility	$\sum_{i=1}^{n} \upsilon(x_i) \cdot \pi(p_i) / \sum_{i=1}^{n} \pi(p_i)$	Karmarkar (1978, 1979)
Rank-dependent expected utility	$\sum_{i=1}^{n} \upsilon(x_i) \cdot \left[G\left(\sum_{j=1}^{i} p_j\right) - G\left(\sum_{j=1}^{i-1} p_j\right)\right]$	Quiggin (1982)
Dual expected utility	$\sum_{i=1}^{n} x_i \cdot \left[G\left(\sum_{j=1}^{i} p_j\right) - G\left(\sum_{j=1}^{i-1} p_j\right)\right]$	Yaari (1987)
Ordinal independence	$\sum_{i=1}^{n} h\left(x_i, \sum_{j=1}^{i} p_j\right) \cdot \left[G\left(\sum_{j=1}^{i} p_j\right) - G\left(\sum_{j=1}^{i-1} p_j\right)\right]$	Segal (1984), Green and Jullien (1988)
Moments of utility	$M\left(\sum_{i=1}^{n} \upsilon(x_i) \cdot p_i \sum_{i=1}^{n} \tau(x_i)^2 \cdot p_i, \ldots\right)$	Múnera and de Neufville (1983), Hagen (1979)
Weighted utility	$\sum_{i=1}^{n} \upsilon(x_i) \cdot p_i / \sum_{i=1}^{n} \tau(x_i) \cdot p_i$	Chew (1983)
Optimism–pessimism	$\sum_{i=1}^{n} \upsilon(x_i) \cdot \mathrm{g}(p_i, x_1, \ldots, x_n)$	Hey (1984)
Quadratic in the probabilities	$\sum_{i=1}^{n} \sum_{j=1}^{n} K(x_i, x_j) \cdot p_i \cdot p_j$	Chew et al. (1991)
Regret theory	$\sum_{i=1}^{n} \sum_{j=1}^{n*} R\left(x_i, x_j^*\right) \cdot p_i \cdot p_j^*$	Loomes and Sugden (1982)

Skaperdas (1993), Schlesinger (1997) and Gollier (2000) have used these forms to revisit many of the applications previously modelled by expected utility theory, such as asset and insurance demand, in order to determine which expected-utility-based results are, and which are not, robust to departures from linearity in the probabilities, and which additional properties of risk-taking behaviour can be modelled.

Although the form $\sum_{i=1}^{n} \upsilon(x_i) \cdot \pi(p_i)$ was the earliest non-expected utility model to be proposed, it was largely abandoned when it was realized that, whenever the weighting function $\pi(\cdot)$ was nonlinear, the generic inequalities $\pi(p_i)+\pi(p_j) \neq \pi(p_i+p_j)$ and $\pi(p_1)+\ldots+\pi(p_n) \neq 1$ implied discontinuities in the payoffs and inconsistency with first-order stochastic dominance preference. Both problems were corrected by adopting weights $\left[G\left(\sum_{j=1}^{i} p_j\right) - G\left(\sum_{j=1}^{i-1} p_j\right)\right]$ based on the *cumulative* probability values $p_1, p_1 + p_2, p_1 + p_2 + p_3, \ldots$, to obtain the Rank-Dependent form. Under the above-mentioned restrictions on this form, these weights necessarily sum to unity, and the Rank-Dependent form has emerged as the most widely adopted model in both theoretical and applied analyses. The Dual Expected Utility and Ordinal Independence forms are based on similar weighting formulas.

Unlike the other models, the regret theory form dispenses with the assumption of a preference function over lotteries, and instead derives choice from the psychological notions of *rejoice* and *regret* – that is, the reaction to receiving outcome x when an alternative decision would have led to outcome x^*. The primitive of this model is a *regret:rejoice function* $R(x, x^*)$ which is positive if x is preferred to x^*, negative if x^* is preferred to x, zero if they are indifferent, and satisfies the skew-symmetry condition $R(x, x^*) \equiv -R(x^*, x)$. In a choice between lotteries $\mathbf{P} = (x_1, p_1; \ldots; x_n, p_n)$ and $\mathbf{P}^* = \left(x_1^*, p_1^*; \ldots; x_{n^*}^*, p_{n^*}^*\right)$ which are realized independently, the individual's *expected rejoice* from choosing **P** over **P*** is given by $\sum_{i=1}^{n} \sum_{j=1}^{i^*} R\left(x_i, x_j^*\right) \cdot p_i \cdot p_j^*$, and the individual is predicted to choose **P** if this value is positive, **P*** if it is negative, and be indifferent if it is zero (various proposals for extending this approach beyond pairwise choice have been offered). Since this model specifies choice in pairwise comparisons rather than preference levels of individual lotteries, it allows choice to be intransitive, so the individual might select **P** over **P***, **P*** over **P****, and **P**** over **P**. Though some have argued that such cycles allow for the phenomenon of 'money pumps', it has allowed the model to serve as a proposed solution to the Preference Reversal Phenomenon.

Generalized Expected Utility Analysis

An alternative approach to non-expected utility preferences does not rely upon any specific functional form, but links properties of attitudes toward risk directly to the probability derivatives of a general 'smooth' preference function V (P)= V $(x_1, p_1;...; x_n, p_n)$. Such analysis reveals that the basic analytics of the expected utility model are in fact quite robust to general smooth departures from linearity in the probabilities. This approach is based on the observations that for the expected utility function V_{EU} $(x_1, p_1;...; x_n, p_n) \equiv U(x_1)\cdot p_1 + ... + U(x_n)\cdot p_n$, the value $U(x_i)$ can be interpreted as the coefficient of p_i, and that many theorems involving a linear function's *coefficients* continue to hold when generalized to a nonlinear function's *derivatives*. By adopting the notation $U(x; \mathbf{P}) \equiv \partial V(\mathbf{P})/\partial \text{prob}(x)$ and the term 'local utility function' for the function $U(\cdot\,;\mathbf{P})$, standard expected utility characterizations such as those listed at the beginning of this article can be generalized to any smooth non-expected utility preference function V(P) in the following manners (Machina 1982):

- $V(\cdot)$ exhibits global *first order stochastic dominance preference* if and only if, at each lottery **P**, its local utility function $U(x; \mathbf{P})$ is an increasing function of x.
- $V(\cdot)$ exhibits global *risk aversion* (aversion to small or large mean-preserving increases in risk) if and only if, at each lottery **P**, its local utility function $U(x; \mathbf{P})$ is a concave function of x.
- $V^*(\cdot)$ is globally *at least as risk averse as* $V(\cdot)$ if and only if, at each lottery **P**, $V^*(\cdot)$'s local utility function $U^*(x; \mathbf{P})$ is a concave transformation of $V(\cdot)$'s local utility function U (x; **P**).

Similar generalizations of expected utility results and characterizations can be obtained for general comparative statics analysis, the theory of asset demand, and the demand for insurance. With regard to the Allais Paradox and other observed violations of the Independence Axiom, it can be shown that the indifference curves of a smooth preference function $V(\cdot)$ will fan out in the probability triangle if and only if $U(x; \mathbf{P}^*)$ is a concave transformation of $U(x; \mathbf{P})$ whenever $\mathbf{P}^*$ first-order stochastically dominates **P**. This analytical approach has been extended to larger classes of preference functionals and distributions by Chew et al. (1987), Karni (1987, 1989) and Wang (1993), formally axiomatized by Allen (1987), and applied to the analysis of choices under uncertainty by Chew et al. (1988), Chew and Nishimura (1992), Dekel (1989), Green and Jullien (1988), Machina (1984, 1989, 1995) and others.

Non-expected Utility Preferences Under Subjective Uncertainty

Recent years have seen a growing interest in models of choice under subjective uncertainty, with efforts to represent and analyse departures from both expected utility risk preferences and probabilistic beliefs. A non-expected utility preference function $W(f(\cdot)) \equiv W(x_1 \text{ on } E_1;...; x_n \text{ on } E_n)$ over subjective acts $f(\cdot) = [x_1 \text{ on } E_1;...; x_n \text{ on } E_n]$ is said to be *probabilistically sophisticated* if it takes the form $W(f(\cdot)) \equiv V(x_1, \mu(E_1);...; x_n, \mu(E_n))$ for some subjective probability measure $\mu(\cdot)$ over the space of events and some non-expected utility preference function $V(\mathbf{P}) = V(x_1, p_1;...; x_n, p_n)$. Such preferences have been axiomatized in a manner similar to Savage's (1954) axiomatization of the subjective expected utility form $W_{SEU}(f(\cdot)) \equiv U(x_1)\cdot \mu(E_1) + ... + U(x_n)\cdot \mu(E_n)$ (Machina and Schmeidler 1992). Although such preferences can be consistent with Allais-type departures from linearity in (subjective) probabilities, they are not consistent with Ellsberg-type departures from probabilistic beliefs.

Efforts to accommodate the Ellsberg Paradox and the general phenomenon of ambiguity aversion have led to the development of several non-probabilistically sophisticated models of preferences over subjective acts (see the analysis of Epstein 1999, as well as the surveys of Camerer and Weber 1992, and Kelsey and Quiggin 1992). One such model, the *maximin expected utility* form, replaces the unique probability measure $\mu(\cdot)$ of the subjective expected utility model by a finite or infinite family M of such measures, to obtain the preference function

$$W_{maximin}(x_1 \text{ on } E_1;\ldots;x_n \text{ on } E_n) \\ \equiv \min_{\mu(\cdot)\in\mathscr{M}} [U(x_1)\cdot\mu(E_1)+\ldots+U(x_n)\cdot\mu(E_n)]$$

When applied to the Ellsberg Paradox, the family of subjective probability measures M = {(μ(red), μ(black), μ(yellow)} = (1/3, γ, 2/3 − γ)|γ ∈[0, 2/3]} will yield the typical Ellsberg-type choices of $f_1(\cdot)$ over $f_2(\cdot)$ and $f_4(\cdot)$ over $f_3(\cdot)$ (Gilboa and Schmeidler 1989).

Another important model for the representation and analysis of ambiguity averse preferences, based on the Rank Dependent form under objective uncertainty, is the *Choquet expected utility* form:

$$W_{Choquet}(x_1 \text{ on } E_1;\ldots;x_n \text{ on } E_n) \\ \equiv \sum_{i=1}^{n} U(x_i)\cdot\left[C\left(\cup_{j=1}^{i}E_j\right)-C\left(\cup_{j=1}^{i-1}E_j\right)\right]$$

where for each act $f(\cdot) = [x_1 \text{ on } E_1;\ldots; x_n \text{ on } E_n]$, the payoffs must be labelled so that $x_1 \leq \ldots \leq x_n$, and $C(\cdot)$ is a *nonadditive* measure over the space of events which satisfies $C(\varnothing) = 0$ and $C\left(\cup_{i=1}^{n}E_j\right) = 1$ (Gilboa 1987; Schmeidler 1989). This model has been axiomatized in a manner similar to the subjective expected utility model, and with proper assumptions on the shape of the utility function $U(\cdot)$ and the nonadditive measure $C(\cdot)$ it is capable of demonstrating ambiguity aversion as well as a wide variety of observed properties of risk preferences.

The technique of generalized expected utility analysis under objective uncertainty has also been adopted to the analysis of general non-expected utility/non-probabilistically sophisticated preference functions $W(f(\cdot)) \equiv W(x_1 \text{ on } E_1;\ldots; x_n \text{ on } E_n)$ over subjective acts. So long as such a function is 'smooth in the events' it will possess a 'local expected utility function' (which may be state-dependent) and a 'local probability measure' at each act $f(\cdot)$, and classical results involving expected utility risk preferences and probabilistic beliefs can typically be generalized in the manner described above (Machina 2005).

See Also

▶ Allais Paradox
▶ Expected Utility Hypothesis
▶ Preference Reversals
▶ Prospect Theory
▶ Risk
▶ Risk Aversion
▶ Savage's Subjective Expected Utility Model
▶ Uncertainty

Bibliography

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

Allen, B. 1987. Smooth preferences and the local expected utility hypothesis. *Journal of Economic Theory* 41: 340–355.

Arrow, K. 1965. *Aspects of the theory of risk bearing*. Yrjö Jahnsson Säätiö: Helsinki.

Battalio, R., J. Kagel, and D. Macdonald. 1985. Animals' choices over uncertain outcomes. *American Economic Review* 75: 597–613.

Becker, S., and F. Brownson. 1964. What price ambiguity? Or the role of ambiguity in decision-making. *Journal of Political Economy* 72: 62–73.

Bell, D. 1985. Disappointment in decision making under uncertainty. *Operations Research* 33: 1–27.

Bernoulli, D. 1738. Specimen theoriae novae de mensura sortis. *Commentarii academiae scientiarum imperialis petropolitanae*. Trans. as Exposition of a new theory on the measurement of risk. *Econometrica* 22(1954): 23–36.

Camerer, C. 1989. An experimental test of several generalized utility theories. *Journal of Risk and Uncertainty* 2: 61–104.

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

Carlin, F. 1990. Is the Allais Paradox robust to a seemingly trivial change of frame? *Economics Letters* 34: 241–244.

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

Chew, S., L. Epstein, and U. Segal. 1991. Mixture symmetry and quadratic utility. *Econometrica* 59: 139–163.

Chew, S., L. Epstein, and I. Zilcha. 1988. A correspondence theorem between expected utility and smooth utility. *Journal of Economic Theory* 46: 186–193.

Chew, S., E. Karni, and Z. Safra. 1987. Risk aversion in the theory of expected utility with rank dependent probabilities. *Journal of Economic Theory* 42: 370–381.

Chew, S., and N. Nishimura. 1992. Differentiability, comparative statics, and non-expected utility preferences. *Journal of Economic Theory* 56: 294–312.

Chew, S., and W. Waller. 1986. Empirical tests of weighted utility theory. *Journal of Mathematical Psychology* 30: 55–72.

Dekel, E. 1989. Asset demands without the independence axiom. *Econometrica* 57: 163–169.

Edwards, W. 1955. The prediction of decisions among bets. *Journal of Experimental Psychology* 50: 201–214.

Edwards, W. 1962. Subjective probabilities inferred from decisions. *Psychological Review* 69: 109–135.

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

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

Friedman, M., and L. Savage. 1948. The utility analysis of choices involving risk. *Journal of Political Economy* 56: 279–304.

Gilboa, I. 1987. Expected utility with purely subjective non-additive probabilities. *Journal of Mathematical Economics* 16: 65–88.

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

Gollier, C. 2000. Optimal insurance design: What can we do without expected utility. In *Handbook of insurance*, ed. G. Dionne. Boston: Kluwer Academic Publishers.

Green, J., and B. Jullien. 1988. Ordinal independence in non-linear utility theory. *Journal of Risk and Uncertainty* 1: 355–387.

Hagen, O. 1979. Towards a positive theory of preferences under risk. In *Expected utility hypotheses and the Allais Paradox*, ed. M. Allais and O. Hagen. Dordrecht: Reidel.

Hershey, J., H. Kunreuther, and P. Schoemaker. 1982. Sources of Bias in assessment procedures for utility functions. *Management Science* 28: 936–954.

Hershey, J., and P. Schoemaker. 1980. Risk-taking and problem context in the domain of losses – An expected utility analysis. *Journal of Risk and Insurance* 47: 111–132.

Hershey, J., and P. Schoemaker. 1985. Probability versus certainty equivalence methods in utility measurement: Are they equivalent? *Management Science* 31: 1213–1231.

Hey, J. 1984. The economics of optimism and pessimism: A definition and some applications. *Kyklos* 37: 181–205.

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

Kahneman, D., and A. Tversky. 1982. The psychology of preferences. *Scientific American* 246: 160–173.

Kahneman, D., and A. Tversky. 1984. Choices, values and frames. *American Psychologist* 39: 341–350.

Karmarkar, U. 1978. Subjectively weighted utility: A descriptive extension of the expected utility model. *Organizational Behavior and Human Performance* 21: 61–72.

Karmarkar, U. 1979. Subjectively weighted utility and the Allais Paradox. *Organizational Behavior and Human Performance* 24: 67–72.

Karni, E. 1987. Generalized expected utility analysis of risk aversion with state- dependent preferences. *International Economic Review* 28: 229–240.

Karni, E. 1989. Generalized expected utility analysis of multivariate risk aversion. *International Economic Review* 30: 297–305.

Keller, L. 1985. The effects of problem representation on the sure-thing and substitution principles. *Management Science* 31: 738–751.

Kelsey, D., and J. Quiggin. 1992. Theories of choice under ignorance and uncertainty. *Journal of Economic Surveys* 6: 133–153.

Konrad, K., and S. Skaperdas. 1993. Self-insurance and self-protection: A nonexpected utility analysis. *Geneva Papers on Risk and Insurance Theory* 18: 131–146.

Lichtenstein, S., and P. Slovic. 1971. Reversals of preferences between bids and choices in gambling decisions. *Journal of Experimental Psychology* 89: 46–55.

Loomes, G., and R. Sugden. 1982. Regret Theory: An alternative theory of rational choice under uncertainty. *Economic Journal* 92: 805–824.

MacCrimmon, K. 1965. *An experimental study of the decision making behavior of business executives*. Doctoral dissertation, University of California, Los Angeles.

MacCrimmon, K. 1968. Descriptive and normative implications of the decision-theory postulates. In *Risk and uncertainty: Proceedings of a conference held by the International Economic Association*, ed. K. Borch and J. Mossin. London: Macmillan.

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

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

Machina, M. 1983. Generalized expected utility analysis and the nature of observed violations of the independence axiom. In *Foundations of utility and risk theory with applications*, ed. B. Stigum and F. Wenstøp. Dordrecht: Reidel.

Machina, M. 1984. Temporal risk and the nature of induced preferences. *Journal of Economic Theory* 33: 199–231.

Machina, M. 1987. Choice under uncertainty: Problems solved and unsolved. *Journal of Economic Perspectives* 1(1): 121–154.

Machina, M. 1989. Comparative statics and non-expected utility preferences. *Journal of Economic Theory* 47: 393–405.

Machina, M. 1995. Non-expected utility and the robustness of the classical insurance paradigm. *Geneva Papers on Risk and Insurance Theory* 20: 9–50.

Machina, M. 2005. ‘Expected utility/subjective probability’ analysis without the sure-thing principle or probabilistic sophistication. *Economic Theory* 26: 1–62.

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

Markowitz, H. 1952. The utility of wealth. *Journal of Political Economy* 60: 151–158.

Marschak, J. 1950. Rational behavior, uncertain prospects, and measurable utility. *Econometrica* 18: 111–141.

Morrison, D. 1967. On the consistency of preferences in Allais’ paradox. *Behavioral Science* 12: 373–383.

Moskowitz, H. 1974. Effects of problem representation and feedback on rational behavior in Allais and Morlat-type problems. *Decision Sciences* 5: 225–242.

Múnera, H., and R. de Neufville. 1983. A decision analysis model when the substitution principle is not acceptable. In *Foundations of utility and risk theory with applications*, ed. B. Stigum and F. Wenstøp. Dordrecht: Reidel.

Pratt, J. 1964. Risk aversion in the small and in the large. *Econometrica* 32: 122–136.

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

Raiffa, H. 1961. Risk, ambiguity, and the Savage axioms: Comment. *Quarterly Journal of Economics* 75: 690–694.

Raiffa, H. 1968. *Decision analysis: Introductory lectures on choices under uncertainty.* Reading: Addison-Wesley.

Rothschild, M., and J. Stiglitz. 1970. Increasing risk: I a definition. *Journal of Economic Theory* 2: 225–243.

Rothschild, M., and J. Stiglitz. 1971. Increasing risk: II its economic consequences. *Journal of Economic Theory* 3: 66–84.

Savage, L. 1954. *The foundations of statistics*. New York: John Wiley and Sons Revised edn, New York: Dover Publications, 1972.

Schlesinger, H. 1997. Insurance demand without the expected utility paradigm. *Journal of Risk and Insurance* 64: 19–39.

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

Schoemaker, P., and H. Kunreuther. 1979. An experimental study of insurance decisions. *Journal of Risk and Insurance* 46: 603–618.

Segal, U. 1984. *Nonlinear decision weights with the independence axiom*. Working Paper No. 353, Department of Economics, University of California, Los Angeles.

Slovic, P. 1969. Manipulating the attractiveness of a gamble without changing its expected value. *Journal of Experimental Psychology* 79: 139–145.

Slovic, P., B. Fischhoff, and S. Lichtenstein. 1977. Behavioral decision theory. *Annual Review of Psychology* 28: 1–39.

Slovic, P., and A. Tversky. 1974. Who accepts Savage’s axiom? *Behavioral Science* 19: 368–373.

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

Stigum, B., and F. Wenstøp. *Foundations of utility and risk theory with applications*. Dordrecht: Reidel.

Sugden, R. 1986. New developments in the theory of choice under uncertainty. *Bulletin of Economic Research* 38: 1–24.

Tversky, A. 1975. A critique of expected utility theory: Descriptive and normative considerations. *Erkenntnis* 9: 163–173.

Tversky, A., and D. Kahneman. 1981. The framing of decisions and the psychology of choice. *Science* 211: 453–458.

von Neumann, J., and O. Morgenstern. 1944. *Theory of games and economic behavior*. Princeton: Princeton University Press 2nd edn, 1947; 3rd edn, 1953.

Wang, T. 1993. Lp-Fréchet differentiable preference and ‘local utility’ analysis. *Journal of Economic Theory* 61: 139–159.

Weber, M., and C. Camerer. 1987. Recent developments in modeling preferences under risk. *OR Spektrum* 9: 129–151.

Yaari, M. 1987. The dual theory of choice under risk. *Econometrica* 55: 95–115.

Non-governmental Organizations

Caren Grown

Abstract

This article defines the term ‘non-governmental organizations’ (NGOs) and describes how they operate. It reviews the growth of the NGO sector since the 1980s, examines the reasons why NGOs have proliferated, reviews evidence on NGO impact, and summarizes how economists have modelled and tested hypotheses about the role of NGOs in development assistance.

Keywords

Charitable giving; Contract theory; Development assistance; Foreign aid; Grameen Bank; Humanitarian and relief work; Imperfect information; Incomplete contracts; International development; International donors; Non-governmental organizations; Non-profit organizations; Participation of the poor; Poverty alleviation; Principal agent; Private voluntary organizations; Public goods; Volunteerism

JEL Classifications

L31

The term ‘non-governmental organization’ came into currency in 1945 when the United Nations Charter distinguished between participation rights for intergovernmental specialized agencies and international organizations (Willetts 2002). Non-governmental organizations (NGOs) form that subset of non-profit organizations working in development assistance, international disaster relief, poverty alleviation, and human rights in developing countries (see non-profit organizations). In the literature and in practice, the term ‘NGO’ is often used interchangeably with ‘private voluntary organization’, a term used to refer to organizations based in the United States engaged in overseas provision of services (Anheier and Salamon 1998).

As the NGO sector has grown, so too has the number of definitions, classifications, and taxonomies (Vakil 1997). According to Bebbington (2004, p. 729), ‘discussions of NGOs continue to be plagued by the vexed and ultimately unanswerable question of “what is an NGO” and haunted by endless typologies. While some of these clarify functional differences, they are less helpful in an explanatory sense – why NGOs emerge, why they do what they do and where, and why certain ideas underlie their actions.’ Despite the lack of a uniform definition, most commentators agree that NGOs can be characterized as private, autonomously managed, value-based organizations that depend, in whole or in part, on charitable donations and voluntary service. Although the sector has become increasingly professionalized since the mid-1980s, principles of altruism and volunteerism remain key defining characteristics.

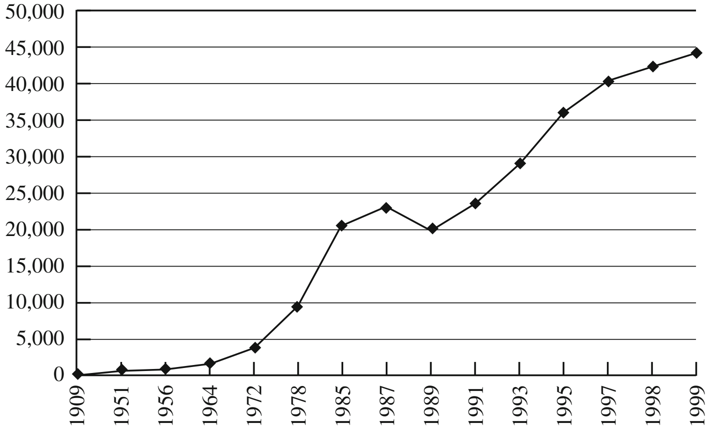
The lack of a uniform definition reflects the heterogeneity of NGOs around the world. They can be structured as large global federated entities, small community based organizations, local or national cooperatives, or large national or international membership organizations. They can carry out a range of functions, from advocacy on behalf of vulnerable or other groups, to direct service (such as providing credit, education and health), research, organizing and public education, humanitarian and relief operations, and peacekeeping operations. Their geographic reach may be in a local community or an entire country, or they may operate across many countries. They are not part of the public sector nor are they dependent on the political process, but in various countries some may seek to influence the formal political process. In many countries, they are exempt from taxes on corporate income. Some NGOs receive funding in the form of grants and contracts from governments and private foundations, others from membership dues and individual contributions, and still others from fees for goods or services. Some receive funding from all these sources.

Examples of organizations that fall in the broad category of NGOs include:

- Centro Mujeres, a small community health organization with 12 staff in 2004 dedicated to fostering the empowerment and well-being of women and adolescents in La Paz, Mexico;
- the Bangladesh Rural Advancement Committee (BRAC), a nationwide organization dedicated to poverty alleviation in Bangladesh with branches in 65,000 villages and more than 97,000 employees in 2006;
- Amnesty International, a global organization with over 1.8 million members in over 150 countries in 2006 and local chapters that undertake research and campaigns to protect human rights and prevent abuses.

Given their heterogeneity of purpose, form and function, NGOs are a multidisciplinary topic, and studies on this topic tend to be published in multidisciplinary journals such as *World Development, the Journal of International Development and Third World Quarterly*. By contrast, the non-profit sector has its own specialized journals.

Non-governmental Organizations, Fig. 1 Total number of NGOs worldwide by year, 1909–1999 (*Source*: Agg (2006))



Research on NGOs is far more common in disciplines other than economics; it is an active field in international relations and development studies. Much of the literature is descriptive, relying on historical analysis or contemporary case studies of single countries, single sectors, or single organizations (Bebbington 2004; Edwards and Hulme 1996). There is surprisingly little survey based research on NGOs in developing countries, especially Africa (Barr et al. 2005). The broad literature explores the growth, evolution, and impact of NGOs in development and relief work in different contexts, NGO relationships with states and donors (and firms in a few instances), and community-based action and social change (Lewis and Opoku-Mensah 2006). NGOs are frequently cast in a favourable light. It is quite common to read articles about the potential of NGOs to transform the development process as opposed to articles about corruption or project failure.

By contrast, the economics literature has tended to develop a narrow range of theoretical models and to take a more critical view of NGOs. Theoretical models explore imperfect information, contracting problems, and accountability in developing countries, using the broad descriptive literature to provide support. With a few exceptions, empirical work by economists has focused mostly on NGOs that provide micro-finance services (Morduch 1999; Pitt and Khandker 1998). The exceptions include Barr et al. (2005), who conducted a survey to document the funding sources and examine monitoring and oversight procedures of NGOs in Uganda; Gauri and Galef (2005), who analysed data from a nationally representative survey of NGOs in Bangladesh; Gauri and Fruttero (2003), who used the Bangladesh Household Income and Expenditure Survey to examine location decisions of NGO programs; and Leonard (1998), who analysed data on health care providers in Cameroon.

Growth of the NGO Sector

Although statistics are hard to come by and what is covered in the numbers can be unclear, Fig. 1 shows spectacular growth of the NGO sector since the 1980s.

The NGO sector has also proliferated in various countries. A recent survey by Gauri and Galef (2005) shows that Bangladesh has one of the largest and most sophisticated NGO sectors in the developing world: over 90 per cent of villages in the country had at least one NGO in 2000 (Gauri and Fruterro 2003) and foreign assistance channelled through NGOs has been above ten per cent since 1993 (Gauri and Galef 2005). As Phinney (2002) notes, ‘In some villages in Bangladesh, you can send your child to an NGO school, have a vasectomy arranged by an NGO health worker, sell your milk to an NGO dairy, and talk on an NGO phone. And, there’s usually a choice of NGO banks.’ International NGOs were responsible for the creation of the NGO community in Bangladesh, although they have withdrawn in recent years and now play a secondary role to local NGOs (Stiles 2002). Price (1999) has noted a significant concentration of NGOs in Latin America, although they are unevenly distributed across countries. In Uganda, Barr et al. (2005) identified 3,500 NGOs registered with the government. Ghanaian NGOs provide 40 per cent of clinical care needs, 27 per cent of hospital beds, and 35 per cent of outpatient services, and in Tanzania NGOs provide half of all hospitals and beds and receive half of all curative visits (Leonard and Leonard 2004). Few national surveys have been undertaken to identify NGO prevalence and incidence in other African countries.

What Explains the Rise of NGOs?

Development studies scholars (geographers, political scientists, anthropologists) argue that NGO involvement in public projects in developing countries has grown in response to budgetary stringency and public sector cutbacks often imposed by macroeconomic stabilization policies (Bebbington and Farrington 1993; Edwards and Hulme 1996). Economists take a less political position, arguing that NGOs are a response to the undersupply of public goods. Bebbington (1997) provides empirical support, noting that Latin American states shifted away from direct implementation of development initiatives in the 1980s and increasingly subcontracted or financed programmes implemented by non-state institutions. In Bolivia, NGOs manage national parks, reserves, and protected areas. In Chile, since the mid-1980s, governments have subcontracted extension services to the private sector; beginning in the 1990s, NGOs and farmers' organizations can also bid for these contracts.

Political scientists and others also highlight the changing preferences of international funders to direct money through private channels due to increasing donor frustration with the public sector because of corruption, inefficiency, and poor results in reducing poverty (Clark 1991; Edwards and Hulme 1996). Empirical evidence suggests that donors have played a key role in the proliferation of international and national NGOs. According to Woods (2003, p. 9), 'resources channeled through NGOs in all OECD member countries rose from 0.2 per cent of the total bilateral ODA [official development assistance] of members of the Development Assistance Committee in 1970 to 17.0 per cent in 1996, to reach, in absolute terms an amount equal to twice the total 1996 ODA of the United Kingdom, the DAC's sixth largest donor by volume.' OECD Development Assistance Committee (DAC) figures show that net grants by NGOs rose from five per cent of total net flows in 2000 to eight per cent of flows in 2004 (OECD 2005, Table 2).

The data have many limitations, and these numbers are likely to be an underestimate. There are complex reporting requirements that are interpreted differently by different governments. For example, donors must choose between designating a disbursement as 'emergency and distress relief' or a grant to an NGO (Agg 2006). Nor do the data include US funds channeled through NGOs. Nonetheless, the OECD data are the only aid data collected over time and from all donor governments.

Meyer (1995) concurs that NGOs arise in part because of donor dissatisfaction with the level of public goods in developing countries, so donors turn to NGOs, which are seen to have some comparative advantages over governments. A number of contributions to the World Development special issue on NGOs in 1987 claim that NGOs have better information on the needs of poor people than do governments; have lower transaction costs; are more flexible than governments and better able to respond to crises such as drought or floods. Because they are part of dense networks with close ties to the community, they are also better at fostering community participation and responding to local needs (Bebbington 2004). Bebbington (2004), for instance, documents how NGOs use methodologies and actions that strengthen capacity and involve poor people in project activities in Latin America. Finally, NGOs are seen to promote new ideas and practices (Scott and Hopkins 1999; Meyer 1995).

What is the Evidence on NGO Impact?

There is little systematic empirical evidence to either support or refute the notion that NGOs are more cost-effective than governments. Some country case studies find that large NGOs working in some sectors do provide some services more cost-effectively than governments (Hasan 1993; AFK/NOVIB 1993; Riddell and Robinson 1992), while others find little difference between governments and NGOs (Tendler 1982, 1989).

Similarly, the evidence on whether NGOs are better at reaching the poorest is also mixed (Fowler 2000; Edwards and Hulme 1996; Arellano-Lopez and Petras 1994; Riddell and Robinson 1992; Tendler 1982). Most NGOs reach the poor, but not necessarily the poorest (UNRISD 2000). An analysis of NGO activity in Bangladesh found that NGO assistance reached those in the second wealth quintile but not those in the poorest (Gauri and Galef 2005). Even the most well-known NGO in Bangladesh, the Grameen Bank, was found to reach less than 20 per cent of landless households in the country (Farrington et al. 1993).

There is greater empirical support for the notion that NGOs have pioneered and used instruments that emphasize the participation of the poor in poverty and development projects (Clark 1995; Bratton 1990). Kilby (2006) finds that formal participation measures and 'downward' accountability practices (for example, to members, clients, other beneficiaries) are correlated with empowerment outcomes in India. Bebbington and Farrington (1993) observe that NGOs that emphasized project methodologies and actions that promote participation have increased the impacts of agricultural development projects.

A number of studies document NGO innovations in various sectors of service delivery, for example in financial services for the poor (Hulme and Mosely 1996), in the creation of debt-for-nature swaps (Meyer 1995), in agriculture technology development (Bebbington and Farrington 1993), and in oral rehydration therapy (Howes and Sattar 1992).

The literature also highlights concerns about the effects of donor financing on NGOs. Edwards and Hulme (1996) argue that increasing reliance on donor funding weakens key attributes that make NGOs attractive to donors in the first place. It can reduce advocacy efforts on behalf of poor and vulnerable groups, negatively affect NGO institutional development, weaken their legitimacy as independent actors, distort their accountability away from internal constituencies to donors and patrons, and lead to an overemphasis on short-term outputs. Fyvie and Ager (1999) argue that donor requirements constrain NGO capacity for innovation. Bebbington (1997, 2005) shows that donor funding of three poverty-oriented rural development NGOs in Peru has over time had several of these effects.

Concerns have also been raised about the nature of NGO, government and donor relations. Scholars have uncovered a range of relationships between NGOs and government, from strongly adversarial to tight partnerships, and between NGOs and donors, from dependent recipient to co-financers/implementers of projects. They have also identified the institutional, economic and political factors that condition which types of relationship emerge and are sustained in different contexts (Bebbington 1997; Nelson 2006; Atack 1999; Anheier and Salamon 1998; Ahmad 2006). By contrast, the economics literature has focused largely on the relationship between NGOs and donors and the conditions for different types of partnerships.

NGO Role in Development Assistance: Theoretical Models

The theoretical economics literature has yet to reflect the diversity of NGO types, the varied impacts of NGO projects, and the multiplicity of NGO, donor, and government relations. Most of this literature focuses on how NGOs compensate for the undersupply of government-provided public goods or are a device to overcome imperfect information and incomplete contracts. Economists have applied principal–agent models with NGOs to the African health-care sector and foreign assistance chains.

Scott and Hopkins (1999) identify the organizational comparative advantage of NGOs and develop a model that explains the circumstances under which they emerge and dominate other types of firms/entities. NGOs predominate in environments where public goods are undersupplied to citizens whose demand for that good exceeds demand of the median voter. The authors argue that the potential superiority of NGOs derives from an institutional environment that selectively attracts altruists who have a lower reservation wage than egotists, and who have the ability to develop efficient technologies for converting the effort of their staffs into local outputs highly valued by the target group of beneficiaries.

The technical superiority of NGOs stems from the way NGOs operate – their interaction with local communities, which enables them to articulate and aggregate local demands. Additionally, NGOs recruit field staff from among beneficiaries and target groups, which facilitates communication and assists in creation of trust between beneficiaries and the target agency. As donors get to know the field and seek the most efficient organizations, NGOs would generally dominate when they have the same or better development technologies than public agencies and wages are similar in both sectors. They may dominate even when wages paid by public agencies are higher, if NGO technology is superior and warm-glow effects are strong enough to outweigh the wage differential.

Besley and Ghatak (2001) develop a model of NGO involvement in public goods provision and enumerate several propositions based on observations from the case study literature. Pure NGO involvement will be more prevalent in projects where the marginal cost of public funds is high and/or the public sector is relatively less efficient in input provision. In activities where performance is hard to measure, NGOs are perceived to be committed to high quality or serve some groups better than others due to their religious or ideological orientation. NGO involvement in supplying services is less dominant in types of projects that are infrastructure-intensive and in countries where the government manages infrastructure well. Decentralization initiatives have often resulted in increased NGO involvement, in part because resource constraints are more severe. NGO provision will also be more prevalent in projects where the NGO cares more about the beneficiaries. Support for this proposition is provided by the World Development Report 1997 (World Bank 1997), which described how governments typically prefer NGOs for delivery of social services while preferring for-profit contractors for the management of infrastructure, such as road maintenance in Brazil.

The models of Leonard (2002) and Leonard and Leonard (2004) address imperfect information and incomplete contracts in the health-care sector in Africa. In sectors where goods or services are characterized by asymmetric information, such as in health care, mechanisms other than prices are needed for the market to function well. In Africa, NGOs are one mechanism to solve the asymmetric information problem. Leonard (2002) and Leonard and Leonard (2004) show they have a stock of attributes which, when combined with the institutional environment in Africa, make them more successful than governments with similar values in providing quality services and reducing the transaction costs of asymmetric information. (There are few private providers in Africa so the relevant comparison is between government and non-governmental services.)

Finally, Azam and Laffont (2003) apply contract theory to shed light on the aid relationship between a donor and recipient country, where consumption of the poor is assumed to be an international public good. The authors model the intricacies of coordinating the efforts of government and NGOs in the fight against poverty. When aid is introduced, several possibilities emerge. Most importantly, free riding problems arise in the provision of aid to the poor when there are several providers. Yontcheva (2003) models a dynamic game between a principal (donor) and agent (NGO), where the model's objective is to identify the long-term determinants of the principal's choice of whether to delegate a project to an NGO and to verify the impact of the allocation on the principal's payoff and the agent's effort.

Conclusion

NGOs are a burgeoning field of cross-disciplinary study. Economists can learn from this voluminous literature both to enrich their models and to contribute theoretical and empirical rigour to a rather messy descriptive literature. They can develop richer theorizations of NGOs roles, relationships and power vis-à-vis governments and donors. They can also work with other social scientists to gather better data on the range of NGO motivations, roles and impacts in various country contexts. This information can help fill an important gap in understanding the dynamic and growing NGO sector in developing countries.

See Also

- ▶ Non-profit organizations
- ▶ Poverty alleviation programmes

Bibliography

Agg, C. 2006. *Trends in government support for non-governmental organizations: Is the 'golden age' of the NGO behind us?* Programme Paper No. 23. Geneva: UNRISD.

Ahmad, M.M. 2006. The 'partnership' between international NGOs (nongovernmental organisations) and local NGOs in Bangladesh. *Journal of International Development* 18: 629–638.

AKFC/NOVIB (Aga Khan Foundation Canada/Netherlands Organization for International Development Cooperation). 1993. *Going to scale: The BRAC experience 1972–1992 and beyond*. The Hague: AKFC/NOVIB.

Anheier, H., and L.M. Salamon. 1998. *The nonprofit sector in the developing world: A comparative analysis*. Manchester/New York: Manchester University Press.

Arellano-Lopez, S., and J.F. Petras. 1994. Non-governmental organizations and poverty alleviation in Bolivia. *Development and Change* 25: 555–568.

Atack, I. 1999. Four criteria of development NGO legitimacy. *World Development* 27: 855–864.

Azam, J., and J. Laffont. 2003. Contracting for aid. *Journal of Development Economics* 70: 25–58.

Barr, A., M. Fafchamps, and T. Owens. 2005. The governance of non-governmental organizations in Uganda. *World Development* 33: 657–679.

Bebbington, A. 1997. New states, new NGOs? Crises and transitions among rural development NGOs in the Andean region. *World Development* 25: 1755–1765.

Bebbington, A. 2004. NGOs and Uneven development: Geographies of development intervention. *Progress in Human Geography* 28: 725–745.

Bebbington, A. 2005. Donor–NGO relations and representations of livelihood in nongovernmental aid chains. *World Development* 33: 937–950.

Bebbington, A., and J. Farrington. 1993. Governments, NGOs and agricultural development: Perspectives on changing inter-organisational relationships. *Journal of Development Studies* 29: 199–219.

Besley, T., and M. Ghatak. 2001. Government versus private ownership of public goods. *Quarterly Journal of Economics* 116: 1343–1372.

Bratton, M. 1990. NGOs in Africa: Can they influence public policy? *Development and Change* 21: 87–188.

Clark, J. 1991. *Democratizing development: The role of voluntary organizations*. West Hartford: Kumarian Press.

Clark, J. 1995. The state, popular participation, and the voluntary sector. *World Development* 23: 593–601.

Collier, P., and D. Dollar. 2004. Development effectiveness: What have we learnt? *Economic Journal* 114: F244–F271.

Edwards, M., and D. Hulme. 1996. Too close for comfort? The impact of official aid on nongovernmental organizations. *World Development* 24: 961–973.

Farrington, J., and D. Lewis, eds. (with S. Satish and A. Miclat-Teves). 1993. *NGOs and the State in Asia: Rethinking roles in sustainable agricultural development*. London: Routledge.

Fowler, A. 2000. NGO futures: Beyond aid: NGDO values and the fourth position. *Third World Quarterly* 21: 589–603.

Fyvie, C., and A. Ager. 1999. NGOs and innovation: Organizational characteristics and constraints in development assistance work in The Gambia. *World Development* 27: 1383–1395.

Gauri, V., and A. Fruttero. 2003. *Location decisions and nongovernmental organization motivation: evidence from rural Bangladesh*, Policy Research Working Paper No. 3176. Washington, DC: World Bank.

Gauri, V., and J. Galef. 2005. NGOs in Bangladesh: Activities, resources, and governance. *World Development* 33: 2045–2065.

Hasan, A. 1993. *Scaling-up the OPP's low-cost sanitation programme*. Karacji: OPP-RTI.

Howes, M., and M. Sattar. 1992. Bigger and better? Scaling-up strategies pursued by BRAC 1972–1991. In *Making a difference: NGOs and development in a changing world*, ed. M. Edwards and D. Hulme. London: Earthscan.

Hulme, D., and P. Mosely. 1996. *Finance against poverty*. London: Routledge.

Kilby, P. 2006. Accountability for empowerment: Dilemmas facing non-governmental organizations. *World Development* 34: 951–963.

Leonard, K.L. 1998. *Institutional structure of health care in rural Cameroun: Structural estimation of production in teams with unobservable effort*, Discussion Paper No. 9798–16, Department of Economics, Columbia University.

Leonard, K.L. 2002. When both states and markets fail: Asymmetric information and the role of NGOs in African health care. *International Review of Law and Economics* 22: 61–80.

Leonard, K.L., and D.K. Leonard. 2004. The political economy of improving health care for the poor in rural Africa: Institutional solutions to the principal–agent problem. *Journal of Development Studies* 40: 50–77.

Lewis, D., and P. Opoku-Mensah. 2006. Moving forward research agendas on international NGOs: theory, agency and context. *Journal of International Development* 18: 665–675.

Meyer, C.A. 1995. Opportunism and NGOs: Entrepreneurship and green North–South transfers. *World Development* 23: 1277–1289.

Morduch, J. 1999. The microfinance promise. *Journal of Economic Literature* 37: 1569–1615.

Nelson, P. 2006. Policy arena, the varied and conditional integration of NGOs in the aid system: NGOs and the World Bank. *Journal of International Development* 18: 701–713.

OECD (Organisation for Economic Co-operation and Development) (ed.). 2005. *Geographical distribution of financial flows to aid recipients 2000–2004*. Paris: OECD/DAC.

Phinney, R. 2002. A model NGO? *Global Policy Forum*. Online. Available at http://www.globalpolicy.org/ngos/fund/2002/1205model.htm. Accessed 31 Mar 2007.

Pitt, M., and S. Khandker. 1998. The impact of group-based credit programs on poor households in Bangladesh: Does the gender of participants matter? *Journal of Political Economy* 106: 958–996.

Price, M.D. 1999. Non-governmental organizations on the geopolitical frontline. In *Reordering the world. Geopolitical perspectives on the 21st century*, ed. G. Demko and W.B. Wood. Boulder: Westview Press.

Riddell, R., and M. Robinson. 1992. *The impact of NGO poverty alleviation projects: Results of the case study evaluations*, Working Paper No. 68. London: Overseas Development Institute.

Scott, R., and C.D. Hopkins. 1999. *The economics of non-governmental organizations*, Development Economics Discussion Paper No. 15, London School of Economics.

Stiles, K. 2002. International support for NGOs in Bangladesh: Some unintended consequences. *World Development* 30: 835–846.

Tendler, J. 1982. *Turning private voluntary agencies into development agencies: questions for evaluation*, Evaluation Discussion Paper No. 10. Washington DC: USAID.

Tendler, J. 1989. Whatever happened to poverty alleviation? *World Development* 17: 1033–1044.

UNRISD (United Nations Research Institute for Social Development). 2000. *Visible hands: Taking responsibility for social development*. Geneva: UNRISD.

Vakil, A. 1997. Confronting the classification problem: Toward a taxonomy of NGOs. *World Development* 25: 2057–2070.

Weisbrod, B.A. 1972. Towards a theory of the voluntary nonprofit sector in a three-sector economy. In *Altruism, morality, and economic theory*, ed. E.S. Phelps. New York: Russell Sage Foundation.

Willetts, P. 2002. What is a non-governmental organization? Article 1.44.3.7, UNESCO Encyclopedia of Life Support Systems. Online. Available at http://www.staff.city.ac.uk/p.willetts/CS-NTWKS/NGO-ART.HTM. Accessed 7 Mar 2007.

Wood, G.D. 1997. States without citizens: The problem of the franchise state. In *NGOs, states and donors: Too close for comfort?* ed. M. Edwards and D. Hulme. London: Macmillan.

Woods, A. 2003. *Facts about European NGOs active in international development*. Paris: Development Centre, OECD.

World Bank. 1997. *World development report 1997: State in a changing world*. New York: Oxford University Press.

World Bank. 2001. *Attacking poverty. World development report, 2000–2001*. Washington, DC: World Bank.

Yontcheva, B. 2003. *Hierarchy and authority in a dynamic perspective: a model applied to donor financing of NGO proposals*, Working Paper No. 03/157, International Monetary Fund.

Non-linear Methods in Econometrics

A. Ronald Gallant

Economic theory guides empirical research primarily by suggesting which variables ought to enter a relationship. But as to the functional form that this relationship ought to take, it only gives general information such as stating that certain first and second partial derivatives of a relationship must be positive or such as ruling out certain functional forms. In some applications, notably consumer demand systems, the theory rules out models that are linear in the parameters such as $y = \sum x_i\beta_i + e$ and thus provides a natural impetus to the development of statistical methods for models that are non-linear in the parameters such as

$$y = \left(\sum x_i\beta_i\right) / \left(\sum x_i\gamma_i - 1\right) + e.$$

A more subtle but more profound influence in the same direction is exerted by the converse aspect of suggesting what variables ought to enter a relationship, that is variables not suggested ought not be present. Thus, when searching for a model that explains data better than an existing model, one will prefer a more complicated model involving only the suggested variables to a model of equal complexity in additional variables. One will inevitably fit models to data that are nonlinear in the parameters during the search.

It is not surprising, therefore, that the subject of nonlinear statistical models developed primarily within econometrics once advances in computing technology and probability theory occurred that would permit it. What is surprising is the rapidity of the development and the speed at which the frontier of econometric research has passed beyond the study of nonlinear statistical models to the natural focus of study, a focus that takes the view that it is best to think of a model as being a point in a function space. Since, as indicated above, the most that economic theory can really say is that a model is a point in a function space, it would seem that the model ought to be studied as such. The process of moving from linear statistical models, through nonlinear models, to the new frontier has taken about fifteen years. Here we shall give an accounting of the statistical aspects of the process. A more detailed development of the subject that follows approximately the same lines as this survey and includes discussion of computations and applications is Gallant (1987).

Prior to 1969, there were scattered papers on nonlinear models with Hartley (1961, 1964) being the most notable contributor. A paper by Jennrich (1969) sparked research in nonlinear statistical models by econometricians. It considered the univariate, nonlinear explicit model

$$y_t = f\left(x_t, \theta^0\right) + e_t \quad t = 1, 2, \ldots, n$$

where y_t is a univariate response, x_t is a k-vector of explanatory variables, θ° is a p-vector of unknown parameters to be estimated, and e_t is an additive error assumed to be independently and identically distributed with mean zero and unknown variance σ^2. In the paper, sufficient conditions were obtained such that the least squares estimator, that is, the estimator $\widehat{\theta}$ that minimizes

$$s_n(\theta) = (1/n)\sum_{t=1}^{n}[y_t - f(x_t, \theta)]^2,$$

over a parameter space Θ, is consistent and asymptotically normally distributed. What had blocked development of an asymptotic theory along conventional lines was the fact that the random variables y_t are not independently and identically distributed. Jennrich showed that the key to overcoming this technical difficulty was a uniform strong law of large numbers that holds if Θ is compact and a central limit theorem that holds for independently but not identically distributed random variables. The compactness assumption is somewhat restrictive but a paper by Malinvaud (1970) showed how the compactness assumption can be circumvented if need be.

These papers set the stage. Over the next ten years there followed a stream of papers extending econometric methods for linear models with a regression structure – ancillary explanatory variables and independent errors – to the analogous nonlinear model. Practical applications proceeded apace. Examples include papers on the asymptotic theory of estimation and inference for multivariate nonlinear models and for nonlinear simultaneous equations models.

From the long-run perspective, the most important outcome of this activity for econometric theory was a reasonable set of conditions such that a triangular array of data generated according to an implicit, nonlinear, simultaneous equations model (the most general model that need be considered) given by

$$q\left(y_t, x_t, \gamma_n^0\right) = e_t, \;\; t = 1, 2, \ldots, n, \quad n = 1, 2, \ldots \mathrm{g}$$

will obey a uniform strong law of large numbers

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} \left|(1/n)\sum_{i=1}^{n}[g(y_t, x_t, \theta) - \mathcal{E}g(y_t, x_t, \theta)]\right| = 0 \text{ a.s.}$$

and will follow a continuously convergent central limit theorem

$$\mathscr{I}_n^{-1/2}(1/\sqrt{n})\sum_{t=1}^{n}\left[g\left(y_t,x_t,\theta_n^0\right)-\mathcal{E}g\left(y_t,x_t,\theta_n^0\right)\right]$$
$$\xrightarrow{L} N(0,I).$$

Above, y_t is an M-variate response, x_t is a k-vector of explanatory variables, γ_n^0 is a sequence of (possibly infinite dimensional) parameters that converges with respect to some metric, $\{e_t\}$ is a sequence of independent, M-variate errors, θ_n^0 is a convergent sequence of p-vectors from a compact set $_{\Theta}$, and $I_n^{-1/2}$ is the Cholesky factorization of the inverse of

$$I_n = \mathrm{Var}\left[(1/\sqrt{n})\sum_{t=1}^{n} g\left(y_t,x_t,\theta_n^0\right)\right]\cdot$$

The dependence of the parameter γ_n^0 on the sample size n is a technical expedient that allows one to deduce a non-null, asymptotic distribution of test statistics. Other than that application, one usually presumes that there is no drift, which is to say that γ_n^0 is equal to some value γ^0 for all n.

With these results in hand, a unified treatment of nonlinear statistical models became possible. It was accomplished in a paper by Burguete et al. (1982). The unifying concept was that estimators $\widehat{\theta}_n$ are solutions to an optimization problem: minimize $S_n(\theta)$ subject to θ in Θ. From this concept, an asymptotic theory of estimation and inference follows by mimicking the standard methods of proof used in maximum likelihood theory but replacing the classical strong law and central limit theorem with those above.

The types of sample objective functions $S_n(\theta)$ that arise in econometrics can be divided into two groups. The first group is least mean distance estimators which have the form

$$s_n(\theta) = (1/n)\sum_{t=1}^{n} s(y_t,x_t,\widehat{\tau}_n,\theta)$$

where $\widehat{\tau}_n$ is a (possibly matrix valued) estimator of nuisance parameters and $s(y, x, \tau, \theta)$ is real valued. The leading example of this type of estimator is multivariate least squares (seemingly unrelated regressions) where data are presumed to be generated according to the explicit, multivariate model

$$y_t = f\left(x_t,\theta^0\right) + e_t,$$

with

$$s_n(\theta) = (1/n)\times\sum_{t=1}^{n}[y_t - f(x_t,\theta)]'\widehat{\tau}_n^{-1}[y_t - f(x_t,\theta)];$$

$\widehat{\tau}_n$ is some estimate of var(e_t), the errors are assumed to be independently and identically distributed. Other examples are maximum likelihood estimators, M-estimators and iteratively rescaled M-estimators for nonlinear (univariate or multivariate) explicit models, and maximum likelihood estimators for nonlinear, simultaneous systems.

The second group is method of moments estimators which have the form

$$S_n(\theta) = (1/2)m_n'(\theta)\widehat{D}_n m_n(\theta)$$
$$m_n(\theta) = (1/n)\sum_{t=1}^{n} m(y_t,x_t,\widehat{\tau}_n,\theta)$$

where $\widehat{\tau}_n$ is an estimator of nuisance parameters and $\widehat{D}_n$ is some matrix valued function of $\widehat{\tau}_n$. The leading example of this type of estimator is the three-stage least-squares estimator where data are presumed to be generated according to the implicit, simultaneous equations model

$$q\left(y_t,x_t,\theta^0\right) = e_t,$$

and

$$m_n(\theta)(1/n)\sum_{t=1} q(y_t,x_t,\theta)\bigotimes Z(x_t)$$

In the expressions above, q, y_t, e_t, are M-vectors, x_t is a k-vector, $Z(x)$ is some (possibly nonlinear) vector-valued function of the explanatory variables x_t, usually low order monomials in the components of x, and $\widehat{D}_n$ is some estimator of Var[$\sqrt{n}m$ (θ^0). Other examples are the twostage least-squares estimator, scale invariant M-estimators, and the Hartley and Booker (1965) estimator.

As regards estimation, the result which follows from the unifying concept are that $\widehat{D}_n$ is estimating that value θ_n^0 which minimizes a function of the form $\overline{s_n}(\theta,\gamma_n^0)$ in the sense that $\sqrt{n}\left(\widehat{\theta}_n - \theta_n^0\right)$ is asymptotically normally distributed. In the case of least mean distance estimators, this function is computed as

$$\overline{s}_n(\theta,\gamma_n^0) = \mathcal{E}(1/n)\sum_{t=1}^{n} s(y_t,x_t,\tau_n^0,\theta)$$

where $\{\tau_n^0\}$ is some sequence for which $\sqrt{n}(\hat{\tau}_n - \tau_n^0)$ is bounded in probability and in the case of method of moments estimators this function is computed as

$$\overline{s_n}(\theta,\gamma_n^0) = (1/2)\overline{m}_n'(\theta,\gamma_n^0)D(\tau_n^0)\overline{m}_n(\theta,\gamma_n^0)$$

$$\overline{m}_n(\theta,\gamma_n^0) = \mathcal{E}(1/n)\sum_{t=1}^{n} m(y_t,x_t,\tau_n^0,\theta)$$

The expectation $\mathcal{E}(.)$ in the expression above is computed according to the model

$$q(y_t,x_t,\gamma_n^0) = e_t, \quad t = 1,2,\ldots,n, \quad n = 1,2,\ldots$$

that actually generates the data, which may be different from the model that was presumed to hold for the purpose of defining the estimation procedure. As a consequence, θ_n^0 will depend on the (possibly infinite dimensional) parameter vector γ_n^0 and, hence, will depend on n. In general, there will be a dependence on n even if γ_n^0 does not drift because the function $\overline{s_n}(\theta,\gamma)$ that defines θ_n^0 depends on n none the less. We will have $\theta_n^0 = \theta^0$ for all n when the presumed model and the actual model coincided and there is no drift.

Nearly every scientist regards a model as an approximation to nature not a description of nature. Thus, the importance of the result above derives not from the fact that it gives an asymptotic approximation to the sampling distribution of an estimator when the presumed model generates the data but from the fact that it gives an approximation when it does not. This provides a scientist with the tools with which to assess the adequacy of the approximation under alternative states of nature.

Above is the statement that $\sqrt{n}\left(\widehat{\theta}_n - \theta_n^0\right)$ is asymptotically normally distributed. More precisely,

$$\mathscr{I}_n^{-1/2}\mathscr{J}_n\sqrt{n}\left(\widehat{\theta}_n - \theta_n^0\right)\xrightarrow{L} N(0,I)$$

In the case of least mean distance estimators, $\mathscr{I}_n$ is computed as

$$\mathscr{I}_n = \mathrm{Var}\left[(1/\sqrt{n})\sum_{t=1}^{n}(\partial/\partial\theta)s(y_t,x_t,\tau_n^0,\theta_n^0)\right]$$

and in the case of method of moments estimators as

$$\begin{aligned}\mathscr{I}_n = \quad & [(\partial/\partial\theta')\overline{m}_n(\theta_n^0,\gamma_n^0)]'D(\tau_n^0)S_nD(\tau_n^0)\\ & \times[(\partial/\partial\theta')\overline{m}_n(\theta_n^0,\gamma_n^0)],\end{aligned}$$

$$S_n = \mathrm{var}\left[(1/\sqrt{n})\sum_{t=1}^{n} m(y_1,x_1,\tau_n^0,\theta_n^0)\right]$$

In either case

$$\mathscr{I}_n = (\partial^2/\partial\theta\partial\theta')\overline{s}_n(\theta_n^0,\gamma_n^0)$$

All computations above are carried out using the actual model to define the expectation and variance operator, not the presumed model. An estimator $\widehat{\mathscr{I}}_n$ is obtained using the obvious sample analogs of $\mathscr{I}_n$ in each instance; for example,

$$\begin{aligned}\widehat{\mathscr{I}}_n = \quad & \left[(\partial/\partial\theta')m_n\left(\widehat{\theta}_n\right)\right]'D(\widehat{\tau}_n)\widehat{S}_nD(\widehat{\tau}_n)\\ & \times\left[(\partial/\partial\theta')m_n\left(\widehat{\theta}_n\right)\right],\end{aligned}$$

$$\begin{aligned}\widehat{S}_n &= (1/n)\\ &\quad\times\sum_{t=1}^{n}\left[m\left(y_t,x_t,\widehat{\tau}_n,\widehat{\theta}_n\right)\right]\left[m\left(y_t,x_t,\widehat{\tau}_n,\widehat{\theta}_n\right)\right]'\end{aligned}$$

In either case $\widehat{\mathscr{I}}_n = (\partial^2/\partial\theta\partial\theta')s_n\left(\widehat{\theta}_n\right)$.

For testing the hypothesis

$$H : h\left(\theta^0\right) = 0 \quad \text{against } A : h\left(\theta^0\right) \neq 0$$

where h is a q-vector one has a Wald test statistic

$$W = n\widehat{h}'\left(\widehat{H}\widehat{V}\widehat{H}'\right)^{-1}\widehat{h},$$

a ‘likelihood ratio’ test statistic

$$L = 2n\left[s_n\left(\widetilde{\theta}_n\right) - s_n\left(\widehat{\theta}_n\right)\right],$$

and a Lagrange multiplier test statistic

$$R = n\left[(\partial/\partial\theta)s_n\left(\widetilde{\theta}_n\right)\right]'\tilde{\mathscr{I}}^{-1}\tilde{H}'\left(\tilde{H}\tilde{V}\tilde{H}'\right)^{-1}\tilde{H}\tilde{\mathscr{I}}^{-1}\left[(\partial/\partial\theta)s_n\left(\widetilde{\theta}_n\right)\right]$$

where $\widetilde{\theta}_n$ minimizes S$n(\theta)$ subject to $h(\theta) = 0$, $\widehat{h} = h\left(\widehat{\theta}_n\right)$, $\widehat{H} = (\partial/\partial\theta')h\left(\widehat{\theta}_n\right)$, $\widehat{V} = \tilde{I}^{-1}\tilde{I}\tilde{I}^{-1}$, and the ~ denotes the same quantities evaluated at $\widetilde{\theta}_n$ instead of at $\widehat{\theta}_n$.

The Wald test can be computed from knowledge of the unconstrained estimate alone; the Lagrange multiplier test from knowledge of the constrained estimate alone. Often one of these will be much easier to compute than the other, thus dictating a choice of test statistics. The ‘likelihood ratio’ test requires knowledge of both and requires, in addition, that $I_n = J_n$ when the presumed model generates the data. It is the preferred statistic when available.

In each instance, one rejects the null hypothesis when the statistic exceeds the upper α 100 percentage point of the chi-square distribution with q degrees of freedom. If the presumed model generates the data then each statistic is asymptotically distributed as a non-central chi-square random variable with q degrees of freedom; if some other model generates the data then each statistic is asymptotically distributed as the ratio of quadratic forms in normal random variables.

As seen from the results summarized above, twelve years after the seminal papers by Jennrich and Malinvaud the literature on nonlinear models with a regression structure was fairly mature. The literature on dynamic models was not. Dynamic models are those where time indexes the observations, where lagged dependent variables, y_{t-1}, y_{t-2} are permitted as explanatory variables amongst the components of x_t, and where errors may be serially correlated. However, some progress in accommodating models with serially correlated errors was made during this period.

The literature failed to accommodate fully dynamic models in the sense that no general, theoretical developments specifically demonstrated that nonlinear models with lagged dependent variables as explanatory variables were included within their scope. This was due to the use of stationary stochastic processes and martingales (which are essentially linear concepts) as the underpinnings of the theory; for instance, a nonlinear transformation of a martingale is in general itself not a martingale. The exceptions to this failure were a monograph by Bierens (1981) and a paper by White and Domowitz (1984). White and Domowitz relied on mixing conditions and a notion of asymptotic martingales due to McLeish (1975, 1977) – notions that will withstand nonlinear transformation – and pointed the way to a general asymptotic theory similar to that outlined above, which was accomplished in a monograph by Gallant and White (1987). The results are the same as those outlined above, with two exceptions.

The first is that the theory cannot accommodate a drift in the traditional fashion where the parameter γ_n^0 tends to a point γ^0 fast enough that $\surd nh\left(\theta_n^0\right)$ is bounded. Rather, γ_n^0 must be held fixed for all n with drift accomplished by formulating the null hypothesis as $H : h\left(\theta_n^0\right) = h_n^0$ and bounding $\sqrt{n}\left[h\left(\theta_n^0\right) - h_n^0\right]$. This is irrelevant as a practical matter because the formulas that one uses to approximate power do not change.

The second exception is that estimating the variance of a sum becomes much more troublesome. For instance, to estimate

$$S_n = \mathrm{Var}\left[(1/\surd n)\sum_{i=1}^{n} m\left(y_t, x_t, \tau_n^0, \theta_n^0\right)\right]$$

one uses

$$\widehat{S}_n = \sum_{\tau=-l(n)}^{l(n)} w[\tau/l(n)]\widehat{S}_{n\tau}$$

where ln(n) is the integer nearest $n^{1/5}$ and

$$w(x) = \begin{cases} 1 - 6|x|^2 + 6|x|^3 & 0 \le |X| \le 1/2 \\ 2\left(1 - |x|^3\right) & 1/2 \le |x| \le 1 \end{cases}$$

$$\widehat{S}_{n\tau} = \begin{cases} (1/n)\displaystyle\sum_{t=1+\tau}^{n} m\left(yt,\ xt,\ \widehat{\tau}_n,\ \widehat{\theta}_n\right) & \\ \times m'\left(y_{t-\tau},\ x_{t-\tau},\ \widehat{\tau}_n,\ \widehat{\theta}_n\right) & \tau \ge 0 \\ \left(\widehat{S}_{n,-\tau}\right)' & \tau < 0 \end{cases}$$

The progress in nonlinear models has indeed been rapid. The developments just described provide an essentially complete asymptotic theory for nonlinear models in as much generality as is likely ever to be useful. There will be refinements over the years but, in the broad sense, the frontier has moved on.

See Also

- ▶ Estimation
- ▶ Least Squares
- ▶ Regression and Correlation Analysis

Bibliography

Bierens, H.J. 1981. Robust methods and asymptotic theory. In *Lecture notes in economics and mathematical systems,* 192. Berlin: Springer-Verlag.

Burguete, J.F., A.R. Gallant, and G. Souza. 1982. On unification of the asymptotic theory of nonlinear econometric models. *Econometric Reviews* 1: 151–190.

Gallant, A.R. 1987. *Nonlinear statistical models*. New York: Wiley.

Gallant, A.R., and H.L. White. 1987. *Consistency and asymptotic normality for parametric estimation with dependent observations*. Oxford: Basil Blackwell.

Hartley, H.O. 1961. The modified Gauss–Newton method for the fitting of nonlinear regression functions by least squares. *Technometrics* 3: 269–280.

Hartley, H.O. 1964. Exact confidence regions for the parameters in nonlinear regression laws. *Biometrika* 51: 347–353.

Hartley, H.O., and A. Booker. 1965. Nonlinear least squares estimation. *Annals of Mathematical Statistics* 36: 638–650.

Jennrich, R.I. 1969. Asymptotic properties of nonlinear least squares estimators. *Annals of Mathematical Statistics* 40: 633–643.

Malinvaud, E. 1970. The consistency of nonlinear regressions. *Annals of Mathematical Statistics* 41: 956–969.

McLeish, D.L. 1975. A maximal inequality and dependent strong laws. *Annals of Probability* 3: 829–839.

McLeish, D.L. 1977. On the invariance principle for nonstationary mixingales. *Annals of Probability* 5: 616–621.

White, H.L., and I. Domowitz. 1984. Nonlinear regression with dependent observations. *Econometrica* 52: 143–162.

Non-linear Panel Data Models

Ekaterini Kyriazidou

Keywords

Conditional maximum likelihood; Conditioning variables; Fixed effects;heteroskedasticity; Incidental parameters; Individual effects; Logit models; Longitudinal data; Nonlinear panel data models; Quadrature procedures; Random effects; Random variables; Tobit models

JEL Classifications

C23; C33

Panel or longitudinal data are becoming increasingly popular in applied work as they offer a number of advantages over pure cross-sectional or pure time-series data. They allow researchers to model *unobserved heterogeneity* at the level of the observational unit, where the latter may be an individual, a household, a firm or a country. This article describes several estimation methods that are available for nonlinear panel data models, that is, models which are nonlinear in the parameters of interest and which include models that arise frequently in applied work, such as discrete choice models and limited dependent variable models, among others.

Introduction

Panel or longitudinal data are becoming increasingly popular in applied work as they offer a number of advantages over pure cross-sectional or pure time-series data. A particularly useful feature is that they allow researchers to model *unobserved heterogeneity* at the level of the observational unit, where the latter may be an individual, a household, a firm or a country. Standard practice in the econometric literature is to model this heterogeneity as an individual-specific effect which enters additively in the model, typically assumed to be linear, that captures the statistical relationship between the dependent and the independent variables. The presence of these *individual effects* may cause problems in estimation. In particular in short panels, that is, in panels where the time-series dimension is of smaller order than the cross-sectional dimension, their estimation in conjunction with the other parameters of interest usually yields inconsistent estimators for both. (Notable exceptions are the static linear and the Poisson count panel data models, where estimation of the individual effects along with the finite dimensional coefficient vector yields consistent estimators of the latter.) This is the well-known *incidental parameters* problem (Neyman and Scott 1948). In linear regression models, this problem may be dealt with by taking transformations of the model, such as first differences or differences from time averages ('within transformation'), which remove the individual effect from the equation under consideration. However they do not apply to nonlinear econometric models, that is, models which are nonlinear in the parameters of interest and which include models that arise frequently in applied work, such as discrete choice models, limited dependent variable models, and duration models, among others.

This article describes several estimation methods that are available for nonlinear panel data models. An approach that is available for estimating certain linear and nonlinear parametric models with individual effects is the *conditional maximum likelihood* approach. This is described in section "The Conditional Maximum Likelihood (CML) Approach". Section "The Fixed Effects Approach" describes estimation techniques that have been recently developed for several semiparametric nonlinear panel data models. A common feature in the methods discussed in that section is that we do not make any assumptions about the nature of these individual effects, that is, whether they are fixed constants or random variables. Thus, we do not make any assumptions about whether they are related to the conditioning variables and, if so, in what manner. This approach is typically referred to as the *fixed effects* approach. Section "The Random Effects Approach" describes the so-called *random effects* approach in estimating nonlinear panel data models. In contrast to the fixed effects approach, the random effects approach does make assumptions about the individual effects.

The discussion distinguishes between two types of models, *static* and *dynamic*. In static models, the conditioning set includes past, present and future values of the variables. In this case the conditioning variables are said to be *strictly exogenous*. In dynamic models, the conditioning set may also include lags of the dependent variable and other endogenous variables, that is, variables that are only *weakly exogenous* or *predetermined*.

Our discussion is limited in several aspects. First, we focus only on the case when the time series dimension of the panel (T) is short so that it makes sense to consider the asymptotic properties of the estimators when the cross-sectional dimension (N) is large while T remains fixed. Second, we do not consider estimation of *random coefficient models*, that is, models where all the parameters are varying at the individual level. Finally, we do not discuss the Bayesian approach to estimating panel data models.

The Conditional Maximum Likelihood (CML) Approach

Suppose that a random variable y_{it} has density $f(\cdot,\theta,\alpha_i)$ where θ is the parameter of interest which is common across all units i, whereas α_i is a nuisance parameter which is allowed to differ across i. A *sufficient statistic* S_i for a_i is a function of the data such that the conditional distribution of the data given S_i does not depend on α_i. However, the conditional distribution may depend on θ. In this case, one can estimate θ by maximizing the *conditional likelihood function*, which conditions on the sufficient statistic(s). Such sufficient statistics are readily available for the exponential family that includes the normal, Poisson, gamma, logistic, and binomial distributions. The CML approach, when it exists, yields consistent and asymptotically normal estimators for parametric panel data models with individual effects (Andersen 1970). We will next demonstrate how the CML approach works in the case of a static and a dynamic logit model with individual effects.

The Static Panel Data Logit Model

Consider the binary choice logit model with individual effects

$$y_{it} = 1\{x_{it}\beta_0 + \alpha_i + \varepsilon_{it} \geq 0\}\, i = 1, \ldots, N; \\ t = 1, \ldots, T$$

where $1\{\text{A}\} = 1$ if A occurs and is 0 otherwise. Let $x_i \equiv (x_{i1}, \ldots, x_{iT})$. Here the error term ε_{it} is distributed i.i.d. over t with a logistic distribution conditional on (x_i, α_i). Note that this assumption implies that ε_{it} is in fact independent of α_i and x_{it} for all t. We can easily calculate that

$$\Pr(y_{it} = 1 | x_i, \alpha_i) = \frac{\exp(x_{it}\beta_0 + \alpha_i)}{1 + \exp(x_{it}\beta_0 + \alpha_i)}.$$

In this model it turns out that $\sum_t y_{it}$ is a sufficient statistic for α_i. Indeed, let $T = 2$. Note that

$$\Pr(y_{it} = 1 | y_{i1} + y_{i2} = 0, x_i, \alpha_i) \\ = 0\ \Pr(y_{it} = 1 | y_{i1} + y_{i2} = 2, x_i, \alpha_i) = 1$$

that is, individuals who do not switch states (i.e. who are 0 or 1 in both periods) do not offer any information about β_0. But it can be easily shown that

$$\Pr(y_{i1} = 1 | y_{i1} + y_{i2} = 1, x_i, \alpha_i) \\ = \frac{1}{1 + \exp((x_{i2} - x_{i1})\beta_0)}$$

and

$$\Pr(y_{i1} = 0 | y_{i1} + y_{i2} = 1, x_i, \alpha_i) \\ = \frac{\exp((x_{i2} - x_{i1})\beta_0)}{1 + \exp((x_{i2} - x_{i1})\beta_0)}.$$

In other words, conditional on the individual switching states (from 0 to 1 or from 1 to 0), the probability that y_{it} is 1 or 0 depends on β_0 (that is, contains information about β_0) but is independent of α_i.

The conditional log-likelihood is

$$\mathcal{L}_C(\beta) = \sum_{i=1}^{N} 1\{y_{i1} + y_{i2} = 1\} \\ \times \ln\left(\frac{\exp((x_{i2} - x_{i1})\beta)^{(1-y_{i1})}}{1 + \exp((x_{i2} - x_{i1})\beta)}\right)$$

and may be maximized over β to produce a consistent and root-N asymptotically normal estimator of β_0. Note that the approach uses a subset of the data, since only individuals who switch states enter the likelihood. For the expression of the conditional log-likelihood in the general T case, see Chamberlain (1984).

The Dynamic Panel Data Logit Model

Chamberlain (1985) noticed that the conditional maximum likelihood approach also applies to the 'AR(1)' logit model with individual effects:

$$y_{it} = 1\{\gamma_0 y_{it-1} + \alpha_i + \varepsilon_{it} \geq 0\}\, i = 1, \ldots, N; \\ t = 1, \ldots, T$$

where the error term ε_{it} is distributed i.i.d. with a logistic distribution conditional on α_i and the initial observation of the sample y_{i0}. Note that we are not making any assumption about the distribution of the initial y_{i0}. As we will see, the approach requires at least four observations for each individual (including the initial observation). In fact, let that be the case and consider the events:

$$A = \{y_{i0} = d_0, y_{i1} = 0, y_{i2} = 1, y_{i3} = d_3\} \\ B = \{y_{i0} = d_0, y_{it} = 1, y_{i2} = 0, y_{i3} = d_3\}$$

where d_0 and d_3 are either 0 or 1. It is rather easy to derive the following probabilities which condition on the individual switching states in the two middle periods

$$
\begin{aligned}
&\Pr(A|A\cup B,\alpha_i)\\
&=\frac{1}{1+\exp(\gamma_0(d_0-d_3))}\Pr(B|A\cup B,\alpha_i)\\
&=\frac{\exp(\gamma_0(d_0-d_3))}{1+\exp(\gamma_0(d_0-d_3))}.
\end{aligned}
$$

Note that these depend on γ_0 but are independent of α_i. The conditional log-likelihood of the model for four periods is:

$$
\begin{aligned}
\mathscr{L}_C(\beta) = &\sum_i 1\{y_{i1}+y_{i2}=1\}\\
&\ln\left(\frac{\exp(\gamma(y_{i0}-y_{i3}))^{y_{i1}}}{1+\exp(\gamma(y_{i0}-y_{i3}))}\right)
\end{aligned}
$$

and maximizing it with respect to γ produces a consistent and root-N asymptotically normal estimator. The approach generalizes to logit models with more than one lags of y_{it} (see Magnac 2000).

It is important to note that the CML approach described above does *not* work in the logit model

$$
\begin{aligned}
y_{it} &= 1\{\gamma_0 y_{it-1}+x_{it}\beta_0+\alpha_i+\varepsilon_{it}\geq 0\}\, i\\
&= 1,\ldots,N;\, t=1,\ldots,T
\end{aligned}
$$

that is, when the conditioning set also includes exogenous variables. Honoré and Kyriazidou (2000a) show that β_0 and γ_0 in the model above are in fact identified both for the case when the errors ε_{it} are logistic and when they are only assumed to have the same distribution over time conditional on (x_i, y_{i0}) (see below). In the logistic case identification is based on the fact that the following probabilities

$$
\Pr(A|A\cup B,x_{i2}=x_{i3},x_i,\alpha_i)=\frac{1}{1+\exp((x_{i1}-x_{i2})\beta_0+\gamma_0(d_0-d_3))}
$$

$$
\Pr(B|A\cup B,x_{i2}=x_{i3},x_i,\alpha_i)=\frac{\exp\big((x_{i1}-x_{i2})\beta_0+\gamma_0(d_0-d_3)\big)}{1+\exp((x_{i1}-x_{i2})\beta_0+\gamma_0(d_0-d_3))}
$$

are independent of α_i. Note that the probabilities above condition not only on the individual switching states in the middle two periods so that $y_{i1}+y_{i2}=1$ but also on the event that $x_{i2}=x_{i3}$. Honoré and Kyriazidou (2000a) propose estimating β_0 and γ_0 by maximizing

$$
\begin{aligned}
&\sum_i 1\{x_{i2}-x_{i3}=0\}1\{y_{i1}+y_{i2}=1\}\\
&\times\ln\left(\frac{\exp((x_{i1}-x_{i2})\beta+\gamma(y_{i0}-y_{i3}))^{y_{i1}}}{1+\exp((x_{i1}-x_{i2})\beta+\gamma(y_{i0}-y_{i3}))}\right)
\end{aligned}
$$

when $\Pr(x_{i2}=x_{i3})>0$. When $x_{i2}\, x_{i3}$ is continuously distributed with support around 0, β_0 and γ_0 can be obtained by maximizing

$$
\begin{aligned}
&\sum_i K\left(\frac{x_{i2}-x_{i3}}{h_N}\right)1\{y_{i1}+y_{i2}=1\}\\
&\times\ln\left(\frac{\exp((x_{i1}-x_{i2})\beta+\gamma(y_{i0}-y_{i3}))^{y_{i1}}}{1+\exp((x_{i1}-x_{i2})\beta+\gamma(y_{i0}-y_{i3}))}\right)
\end{aligned}
$$

where K () is a *kernel density function* and h_N is a *bandwidth sequence*, chosen so as to satisfy certain assumptions that guarantee consistency and asymptotic normality of the proposed estimators.

The Fixed Effects Approach

The conditional maximum likelihood approach is not always available. For example, there are no sufficient statistics for the binary choice model with individual effects when the errors are normally distributed. Furthermore, like all ML approaches, the approach suffers from the fact that the distribution of the unobserved idiosyncratic errors needs to be parametrically specified. There do exist, however, methods for some *semi-parametric* nonlinear panel data models with individual effects where the distribution of the underlying idiosyncratic errors is left unspecified. These include the binary choice model, the censored and truncated regression models, and the sample selection model.

The Semiparametric Panel Data Binary Choice Model

Manski (1987) considers the model

$$y_{it} = 1\{x_{it}\beta_0 + \alpha_i - \varepsilon_{it} \geq 0\}\, i = 1, \ldots, N;\\ t = 1, \ldots, T$$

where ε_{it} is identically distributed over time conditional on (x_i,α_i), with distribution function F that is a continuous and strictly increasing function on $\mathcal{R}$. Note that, in contrast to the models considered above, F here is not assumed to have a specific functional form, hence the characterization of the model as *semiparametric.*

He observes that for $T = 2$ the time invariance of F implies that

$$\Pr(y_{i1} = 1|x_i) \lesseqgtr \Pr(y_{i2} = 1|x_i) \,\text{if and only if}\, x_{i1}\beta_0 \lesseqgtr x_{i2}\beta_0$$

or equivalently that

$$sgn(\Pr(y_{i2} = 1|\, x_i, \alpha_i) - \Pr(y_{i1} = 1|\, x_i, \alpha_i))\\ = sgn((x_{i2} - x_{i1})\beta_0).$$

In fact it can be shown that, under appropriate regularity conditions on the joint distribution of $\Delta x_i \equiv (x_{i2} - x_{i1})$, β_0 uniquely (up to scale) maximizes the so-called population 'score function'

$$E[\Delta y_i \cdot sgn(\Delta x_i \beta_0)]$$

where $sgn(x)$ equals 1 if $x > 0$, equals $-$ 1 if $x < 0$ and is equal to 0 if $x = 0$. This suggests estimating β_0 by the so-called *conditional maximum score estimator* which maximizes the sample analog of the population score function

$$\hat{\beta} = \arg\max_{\beta} \sum_i \Delta y_i \cdot sgn(\Delta x_i \beta).$$

Note that only observations for which $y_{i1} \neq y_{i2}$ are used here, similarly to conditional logit. The estimator is consistent under some additional assumptions but is not asymptotically normal and its rate of convergence is not root-N.

Honoré and Kyriazidou (2000a) show that it is possible to extend the conditional maximum score approach to the dynamic binary choice model:

$$\begin{aligned}&\Pr(y_{i0} = 1|\, x_i, \alpha_i) = p_0(x_i, \alpha_i)\\ &\Pr(y_{it} = 1|\, x_i, \alpha_i, y_{i0}, \ldots, y_{it-1})\\ &\quad = F(x_{it}\beta_0 + \gamma_0 y_{it-1} + \alpha_i)\, t\\ &\quad = 1, \ldots, T\end{aligned}$$

where y_{i0} is assumed to be observed and F is strictly increasing.

We will next demonstrate their identification scheme. Assume $T = 3$ and define the events A and B as above. Then

$$\begin{aligned}&\Pr(A|\, x_i, \alpha_i, x_{i2} = x_{i3})\\ &\quad = p_0(x_i, \alpha_i)^{d_0}(1 - p_0(x_i, \alpha_i))^{1-d_0}\\ &\quad \times (1 - F(x_{i1}\beta_0 + \gamma_0 d_0 + \alpha_i))\\ &\quad \times F(x_{i2}\beta_0 + \alpha_i)\\ &\quad \times (1 - F(x_{i2}\beta_0 + \gamma_0 + \alpha_i))^{(1-d_3)}\\ &\quad \times F(x_{i2}\beta_0 + \gamma_0 + \alpha_i))^{d_3}\\ &\Pr(B|\, x_i, \alpha_i, x_{i2} = x_{i3})\\ &\quad = p_0(x_i, \alpha_i)^{d_0}(1 - p_0(x_i, \alpha_i))^{1-d_0}\\ &\quad \times F(x_{i1}\beta_0 + \gamma_0 d_0 + \alpha_i)\\ &\quad \times (1 - F(x_{i2}\beta_0 + \gamma_0 + \alpha_i))\\ &\quad \times (1 - F(x_{i2}\beta_0 + \alpha_i))^{(1-d_3)}\\ &\quad \times F(x_{i2}\beta_0 + \alpha_i))^{d_3}.\end{aligned}$$

If $d_3 = 0$, then,

$$\begin{aligned}&\frac{\Pr(A|\, x_i, \alpha_i, x_{i2} = x_{i3})}{\Pr(B|\, x_i, \alpha_i, x_{i2} = x_{i3})}\\ &= \frac{(1 - F(x_{i1}\beta_0 + \gamma_0 d_0 + \alpha_i))}{(1 - F(x_{i2}\beta_0 + \alpha_i))}\\ &\times \frac{F(x_{i2}\beta_0 + \alpha_i)}{F(x_{i1}\beta_0 + \gamma_0 d_0 + \alpha_i)}\\ &= \frac{(1 - F(x_{i1}\beta_0 + \gamma_0 d_0 + \alpha_i))}{(1 - F(x_{i2}\beta_0 + \gamma_0 d_3 + \alpha_i))}\\ &\times \frac{F(x_{i2}\beta_0 + \gamma_0 + \alpha_i)}{F(x_{i1}\beta_0 + \alpha_i)}\end{aligned}$$

while if $d_3 = 1$, then,

$$\begin{aligned}\frac{\Pr(A|\,x_i,\alpha_i,x_{i2}=x_{i3})}{\Pr(B|\,x_i,\alpha_i,x_{i2}=x_{i3})} &= \frac{(1-F(x_{i1}\beta_0+\gamma_0 d_0+\alpha_i))}{(1-F(x_{i2}\beta_0+\gamma_0+\alpha_i))} \\ &\times \frac{F(x_{i2}\beta_0+\gamma_0+\alpha_i)}{F(x_{i1}\beta_0+\gamma_0 d_0+\alpha_i)} \\ &= \frac{(1-F(x_{i1}\beta_0+\gamma_0 d_0+\alpha_i))}{(1-F(x_{i2}\beta_0+\gamma_0 d_3+\alpha_i))} \\ &\times \frac{F(x_{i2}\beta_0+\gamma_0 d_3+\alpha_i)}{F(x_{i1}\beta_0+\gamma_0 d_0+\alpha_i)}.\end{aligned}$$

Monotonicity of F implies that

$$\begin{aligned}&sgn(\Pr(A|\,x_i,\alpha_i,x_{i2}=x_{i3}) \\ &\quad -\Pr(B|\,x_i,\alpha_i,x_{i2}=x_{i3}) \\ &\quad = sgn((x_{i2}-x_{i1})\beta_0+\gamma_0(d_3-d_0)).\end{aligned}$$

This last equation suggests that β_0 and γ_0 can be estimated by conditional maximum score using only the observations satisfying $y_{i1}+y_{i2}=1$ and $x_{i2}=x_{i3}$, that is, by maximizing

$$\begin{aligned}&\sum_i 1\{x_{i2}-x_{i3}=0\}\,(y_{i2}-y_{i1}) \\ &\quad sgn\,((x_{i2}-x_{i1})\beta+\gamma(y_{i3}-y_{i0})).\end{aligned}$$

Similar to the logit case, when $x_{i2}-x_{i3}$ is continuously distributed with support around 0, estimation of β_0 and γ_0 can be obtained by maximizing

$$\begin{aligned}&\sum_i K\left(\frac{x_{i2}-x_{i3}}{h_N}\right)(y_{i2}-y_{i1})sgn\,((x_{i2}-x_{i1})\beta \\ &\quad +\gamma(y_{i3}-y_{i0}).\end{aligned}$$

The Semiparametric Panel Data Censored Regression Model

The standard censored panel data (or Type 1 Tobit) model with individual effects is given by

$$\begin{aligned}y_{it} &= \max\{x_{it}\beta_0+\alpha_i+\varepsilon_{it},0\}\,i=1,\ldots,N;t \\ &= 1,\ldots,T.\end{aligned}$$

Estimation of this model was first considered by Honoré (1992) and later by Honoré and Kyriazidou (2000b), who extend the results of the former paper. We will present here Honoré (1992), who assumes that (ε_{it}, ε_{is}) are *pairwise exchangeable* conditional on (x_i, a_i). This implies that ε_{it} and ε_{is} are identically distributed conditional on (x_i, a_i) although it does not require (conditional) independence over time. (Fristedt and Gray 1997, give the following definition of exchangeability: Let I e a countable set. A sequence $(X_i : i \in \mathscr{I})$, finite or infinite, of random variables on a probability space (Ω; F; P) is *exchangeable* if, for every permutation ρ of I, the distribution of $\left(X_{p(i)} : i \in \mathscr{I}\right)$ and $(X_i : i \in \mathscr{I})$) are identical. Note that a finite or infinite i.i.d. sequence is exchangeable and that exchangeability allows for certain types of serial correlation. Furthermore, exchangeability implies strict stationarity although the converse is not true.)

Consider the 'pseudo-error':

$$e_{ist}(\beta) = \max\{y_{is},(x_{is}-x_{it})\beta\}-x_{is}\beta.$$

With this definition, at the true β_0

$$\begin{aligned}e_{ist}(\beta_0) \;&= \max\{y_{is},(x_{is}-x_{it})\beta_0\}-x_{is}\beta_0 \\ &= \max\{\max\{x_{is}\beta_0+\alpha_i+\varepsilon_{is},0\},(x_{is}-x_{it})\beta_0\} \\ &-x_{is}\beta_0 = \max\{\max\{\alpha_i+\varepsilon_{is},-x_{is}\beta_0\},-x_{it}\beta_0\} \\ &= \max\{\alpha_i+\varepsilon_{is},-x_{is}\beta_0,-x_{is}\beta_0\}\end{aligned}$$

The conditional exchangeability assumption implies that ($e_{ist}(\beta_0)$, $e_{its}(\beta_0)$) is distributed like ($e_{its}(\beta_0)$, $e_{ist}(\beta_0)$) conditional on (x_{it}, x_{is}, a_i) and hence the difference $e_{its}(\beta_0)$ − $e_{ist}(\beta_0)$ is distributed symmetrically around 0 conditional on (x_{it}, x_{is}, a_i). Since this is true for any α_i this symmetry holds conditional only on (x_{it}, x_{is}). Therefore for any odd function ξ (that is, a function ξ that satisfies $\xi(-d)=-\xi(d)$) we have

$$E[\xi(e_{ist}(\beta_0)-e_{ist}(\beta_0))|\,x_{it},x_{is}]=0$$

which also implies the following moment restriction:

$$E\left[\xi(e_{ist}(\beta_0)-e_{ist}(\beta_0))(x_{is}-x_{it})'|\,x_{it},x_{is}\right]=0.$$

The left-hand side of the moment condition above may be thought of as the first order condition for the following population minimization problem

$$\min_{\beta} E[q(y_{is}, y_{it}, (x_{is} - x_{it})\beta) | x_{it}, x_{is}]$$

Where

$$q\left(y_i, y_j, \delta\right) = \begin{cases} \Xi(y_i) - \left(y_j + \delta\right)\xi(y_i) & \text{if} \quad \delta \leq -y_j \\ \Xi\left(y_i - y_j - \delta\right) & \text{if} \quad -y_j < \delta < y_i \\ \Xi\left(-y_j\right) - \left(\delta - y_j\right)\xi(-y_i) & \text{if} \quad y_i \leq \delta \end{cases}$$

and $\Xi(d)$:R →R)+ is an even function (that is, $\Xi(-d) = \Xi(d)$) which is convex, strictly increasing for $d > 0$ and has Ξ (0) = 0, and $\Xi'(d) = \xi(d)$ where $\xi(0) = 0$.Note that for Ξ to be convex, ξ has to be monotone. Obvious choices for Ξ are Ξ $(d) = d^2$ (which corresponds to $\xi(d) = 2d$) and Ξ $(d) = |d|$ (which corresponds to $\xi(d) = sgn(d)$).

The fact that the true β_0 solves the population minimization problem above suggests the following estimator for β_0:

$$\hat{\beta} = \arg\min_{\beta} \sum_i \sum_{s<t} q(y_{is}, y_{it}(x_{is} - x_{it})\beta).$$

Honoré (1992) shows that the estimators corresponding to Ξ $(d) = d^2$ and Ξ $(d) = |d|$ are root-N consistent and asymptotically normal.

Honoré (1993) considers a dynamic version of the model where the lag of the *observed* (censored) dependent variable appears in the model instead of the latent one. Hu (2002) considers the case where one lag of the *latent* (unobserved) dependent variable is included along with the set of exogenous variables x_{it}.

The Semiparametric Panel Data Sample Selection Model

The standard panel data sample selection (or Type 2 Tobit) model is defined as:

$$y^*_{it} = x^*_{it}\beta_0 + \alpha^*_i + \varepsilon^*_{it} y_{it} = d_{it} \cdot y^*_{it} d_{it} \\ = 1\{z_{it}\gamma_0 + \eta_i - u_{it} \geq 0\}$$

where $i = 1,2,\ldots,N$; $t = 1,\ldots T$. Kyriazidou (1997) considers estimation without any parametric assumptions on the form of the joint distribution of $\left(\varepsilon^*_{it}, u_{it}\right)$ or on the individual effects (α_i, η_i).

Consider the case where $T = 2$ and only those individuals for whom $d_{i1} = d_{i2} = 1$. Let $\xi_i = \left(z_{i1}, z_{i2}, x^*_{i1}, x^*_{i2}, \alpha_i, \eta_i\right)$ denote all the information about individual *i*. Note that

$$E(y_{i1} - y_{i2} | d_{i1} = d_{i2} = 1, \xi_i) \\ = \left(x^*_{i1} - x^*_{i2}\right)\beta_0 \\ + E\left(\varepsilon^*_{i1} - \varepsilon^*_{i2} | d_{i1} = d_{i2} = 1, \xi_i\right)$$

and hence OLS estimation of the first differenced model will not yield consistent estimation of β_0 since in general the so-called 'sample selection bias term'

$$\lambda_{it} \equiv E\left(\varepsilon^*_{it} | d_{i1} = d_{i2} = 1, \xi_i\right) \\ = E\left(\varepsilon^*_{it} | u_{i1} \leq z_{i1}\gamma_0 + \eta_i, u_{i2} \leq z_{i2}\gamma_0 + \eta_i, \xi_i\right)$$

is not zero. Nor do we have in general that $\lambda_{i1} = \lambda_{i2}$, so that first differencing removes the sample selection bias along with the individual effects. Kyriazidou (1997) makes a *conditional exchangeability assumption* that $\left(\varepsilon^*_{i1}, \varepsilon^*_{i2}, u_{i1}, u_{i2}\right)$ and $\left(\varepsilon^*_{i2}, \varepsilon^*_{i1}, u_{i2}, u_{i1}\right)$ are identically distributed conditional on ξ_i. Under this assumption, it is easy to see that if $z_{i1}\gamma_0 = z_{i2}\gamma_0$ then

$$\lambda_{i1} \\ = E\left(\varepsilon^*_{i1} | u_{i1} \leq z_{i1}\gamma_0 + \eta_i, u_{i2} \leq z_{i2}\gamma_0 + \eta_i, \xi_i\right) \\ = E\left(\varepsilon^*_{i2} | u_{i1} \leq z_{i1}\gamma_0 + \eta_i, u_{i2} \leq z_{i2}\gamma_0 + \eta_i, \xi_i\right) \\ = \lambda_{i2}$$

so that first differencing will eliminate both α_i and λ_{it} simultaneously. So β_0 can be estimated by first difference OLS for the subsample of individuals that are observed in both periods (that is, that have $d_{i1} = d_{i2} = 1$) and also have the selection index, $z_{it}\gamma_0$, constant (that is, $z_{i1}\gamma_0 = z_{i2}\gamma_0$). Of course, this estimation scheme cannot be directly implemented since γ_0 is unknown. And it is quite possible that no observation has $z_{i1}\gamma_0 = z_{i2}\gamma_0$ if $z_i\gamma_0$ is continuously distributed. If, however, λ_{it} is a sufficiently smooth function and $\hat{\gamma}$ is a consistent estimator of γ_0, $z_{i1}\gamma_0 \approx z_{i2}\gamma_0$ implies $\lambda_{i1} \approx \lambda_{i2}$, and the preceding augment holds approximately. Kyriazidou proposes a two-step estimation procedure, in the spirit of Powell (2001), and Ahn and Powell (1993) who consider estimation of cross-section versions of the sample selection model. In the first step, γ_0 is consistently estimated based on the selection equation. In the second step, the estimate $\hat{\gamma}$ is used to estimate β_0 based on those pairs of observations for which $z_{i1}\hat{\gamma}$ and $z_{i2}\hat{\gamma}$ are 'close'. To this end define

$$\hat{\psi}_i = \frac{1}{h_N} K\left(\frac{\Delta z_i \hat{\gamma}}{h_N}\right)$$

where K () is a kernel density function and h_N is a bandwidth sequence. The proposed estimator takes the form:

$$\hat{\beta} = \left[\sum_{i=1}^{N} \hat{\psi}_i \Delta x_i' \Delta x_i d_{i1} d_{i2}\right]^{-1} \sum_{i=1}^{N} \hat{\psi}_i \Delta x_i' \Delta y_i d_{i1} d_{i2}.$$

Under some assumptions and by appropriately choosing h_N, the estimator can be shown to be asymptotically normal although the rate of convergence is slower that the parametric $\sqrt{N}$ rate. Apart from the conditional exchangeability assumption, another important assumption that underlies the approach is that there is at least one variable in z_{it} not contained in x_{it}, which is an exclusion restriction common in semiparametric sample selection models.

A dynamic version of the panel data sample selection model, with the own lagged dependent variable appearing in each equation, is considered by Kyriazidou (2001).

The Random Effects Approach

Fixed effects methods and conditional maximum likelihood methods (when they exist) estimate the coefficients of time-varying regressors consistently without making any assumptions on how the individual effects are related to the observed covariates or to the time-varying errors or to the initial observations of the sample. However, these methods do not deliver estimates of coefficients of time-invariant regressors and of the individual effects, and hence cannot be used for prediction, or for computation of marginal effects and elasticities which are often the quantities of interest. Furthermore, none of these approaches allows for non-stationary errors and hence for time-series heteroskedasticity.

These problems do not arise in the random effects approach. The approach essentially consists of treating $(\alpha_I + \varepsilon_{it})$ as a two-component error term and making assumptions about its relationship with the observed covariates and, in the case of dynamic models, with the initial conditions as well. A downside of the approach is that misspecification of any part of the model typically yields inconsistent estimates.

Static Case

In the static panel data linear regression model, the traditional random effects approach (sometimes also called the *uncorrelated random effects approach*) assumes that the individual effects α_i along with the time-varying errors ε_{it} are uncorrelated with the observed covariates x_{it}. Then the coefficients of both time-varying and time-invariant regressors may be estimated consistently (albeit not efficiently) by pooled OLS. In static nonlinear models, the traditional random effects approach apart from parameterizing the conditional distribution of ε_{it} given x_{it}, also assumes that α_i is independent of x_{it} and ε_{it} for all t, and has a distribution, say H, that depends on a finite set of unknown parameters, say δ_0. For example, in the binary choice model,

$$y_{it} = 1\{x_{it}\beta_0 + \alpha_i + \varepsilon_{it} \geq 0\} i = 1, \ldots, N; t = 1, \ldots, T \tag{1}$$

assuming that ε_{it} are i.i.d. over time and independent of x_i and α_i with known distribution F (say, standard normal or logistic), we may estimate the unknown parameters (β_0, δ_0) via ML. The log-likelihood is

$$\ln L(\beta, \delta) = \sum_i \ln \int \prod_{T=1}^{T} F(x_{it}\beta + \alpha)^{y_{it}} (1 - F(x_{it}\beta + \alpha))^{1-y_{it}} dH(\alpha, \delta)$$

and involves a one-dimensional integral which may be calculated numerically, for example, by *quadrature procedures* (see Butler and Moffitt 1982).

However, things become quite complicated if we want to allow for arbitrary serial correlation in the ε_{it}'s. Consider the binary choice model

$$y_{it} = 1\{x_{it}\beta_0 - u_{it} \geq 0\}$$

where $u_{it} = \alpha_I + \varepsilon_{it}$ is the composite error term. For $T = 3$ there are 2^3 possible sequences of 0's and 1's. The likelihood for an individual for whom the sequence of observed y_{it}'s is (0,1,0) takes the form

$$\int_{x_{i1}\beta} \int^{x_{i2}\beta} \int_{x_{i3}\beta} f(u_1, u_2, u_3) du_1 du_2 du_3$$

where f is the trivariate density of (u_1,u_2,u_3) conditional on x_i. The log-likelihood is

$$\begin{aligned}\ln L(\beta,\delta) = & \sum_{i:(0,0,0)} \ln \int_{x_{i1}\beta} \int_{x_{i2}\beta} \int_{x_{i3}\beta} f(u_1,u_2,u_3) du_1 du_2 du_3 \\ & + \sum_{i:(0,0,1)} \ln \int_{x_{i1}\beta} \int_{x_{i2}\beta} \int^{x_{i3}\beta} f(u_1,u_2,u_3) \\ & \times du_1 du_2 du_3 + \ldots\end{aligned}$$

which requires the computation of multiple trivariate integrals. Multivariate integration is basically infeasible for large *T*. This is where simulation methods come in very handy.

The assumption that α_i is independent of x_i is often found unsatisfactory. A possible solution is to assume a specific functional form for the relationship of α_i with x_i. This approach (recently also called the *correlated random effects approach*) was first proposed by Chamberlain (1984). Suppose that

$$\alpha_i = \sum_{t=1}^{T} x_{it}\gamma_{0,t} + v_i$$

where v_i is independent of x_i, similarly to the time varying error component ε_{it}, and that the composite new error term $v_i + \varepsilon_{it}$ follows a specific distribution, say normal. In the case of the binary choice model, for example, assuming that $\varepsilon_{it} + v_i |$ x_i; α_i is $N\left(0, \sigma^2_{0,t}\right)$ implies that

$$\begin{aligned}\Pr(y_{it} = 1 | x_i) &= \Phi\left(\frac{x_{it}\beta_0 + \sum_{t=1}^{T} x_{it}\gamma_{0,t}}{\sigma_{0,t}}\right) \\ &= \Phi\left(x_{it}\theta_{0,t}\right).\end{aligned}$$

For computational simplicity, Chamberlain proposes to estimate the unknown parameters $\theta_{0,\mathrm{t}}$ via period-by-period probit. The 'structural parameters' β_0, $\left\{\sigma^2_{0,t}\right\}_{t=1}^{T}$, and $\left\{\gamma_{0,t}\right\}_{t=1}^{T}$ can then be recovered by *minimum distance* estimation. Note that the approach allows for time series heteroskedasticity and requires only one normalization e.g. that$\sigma^2_{0,t} = 1$.

Newey (1994) generalizes Chamberlain's approach by postulating that

$$\alpha_i = \rho(x_{i1}, \ldots, x_{it}) + v_i$$

where ρ () is an unknown function of x_i. Assuming again that v_i and ε_{it} are independent of x_i and that the composite new error term v_i + ε_{it} follows a specific distribution, say F_t, we obtain

$$\pi_t = \Pr(y_{it} = 1 | x_i) = F_t(\rho(x_i) + x_{it}\beta_0)$$

which for a strictly monotonic F_t implies that

$$F_t^{-1}(\pi_t) = \rho(x_i) + x_{it}\beta_0.$$

For example in the normal case

$$\Phi^{-1}(\pi_t) = \frac{\rho(x_i) + x_{it}\beta_0}{\rho_{0,t}}.$$

Thus for two periods *t* and *s* we obtain

$$\Phi^{-1}(\pi_t) = \frac{\sigma_{0,s}}{\sigma_{0,t}}\Phi^{-1}(\pi_s) + \frac{\sigma_{0,s}}{\sigma_{0,t}}(x_{it} - x_{is})\beta_0.$$

Normalizing $\sigma_{0,\mathrm{t}} = 1$ and estimating π_t and π_s nonparametrically, we can recover $\sigma_{0,\mathrm{s}}$ and β_0 from the regression of $\Phi^{-1}(\hat{\pi}_t)$ on $\Phi^{-1}(\hat{\pi}_s)$ and $(x_{it} - x_{is})$.

A criticism of all these correlated random effects approaches is that, although in the linear model writing $\alpha_i = \sum_{t=1}^{T} x_{it}\gamma_{0,t} + u_i$ where $E(u_i x_{it}) = 0$ for all *t* does not impose x_{it} $-x_{is}$ any restrictions on the joint distribution of α_i and x_i (apart from the requirement that it has second moments) since this is just the best linear projection of α_i on x_i, in the nonlinear model assuming $\alpha_i = \rho(x_{i1},\ldots,x_{it}) + u_i$, even without specifying the functional form of ρ, imposes implausible restrictions in the sense that, if this relationship holds for the T observations, a similar one will *not* in general hold for $T + 1$.

Dynamic Case

In the case where there are genuine dynamics in the model in the form of lags of the dependent variable or other endogenous regressors, random effects methods become even more complicated and require additional assumptions about the relationship of the individual effects with the initial observations. We next describe a general approach for estimating dynamic random effects models suggested by Wooldridge (2000). For simplicity we will drop the subscripts i.

We are interested in the conditional distribution of y_t given a vector of strictly exogenous variables $z^T \equiv (z_1, \ldots, z_T)$, own lags and lags of other endogenous variables $x^{t-1} \equiv (y_{t-1}, w_{t-1}, y_{t-2}, w_{t-2}, \ldots, y_0, w_0)$, and an unobserved scalar or vector random effect α. Here z_t is strictly exogenous in the sense that

$$F\left(w_t | z^T, x^{t-1}, \alpha\right) = F\left(w_t | z_t, x^{t-1}, \alpha\right).$$

The conditional density of $x_t \equiv (y_t, w_t)$ is

$$\begin{aligned} f_t\left(x_t | z^T, x^{t-1}, \alpha\right) &= f_t\left(x_t | z_t, x^{t-1}, \alpha\right) \\ &= f_t\left(y_t | w_t, z_t, x^{t-1}, \alpha\right) \\ &\quad \cdot f_t\left(w_t | z_t, x^{t-1}, \alpha\right) \end{aligned}$$

and the joint density for all T periods is

$$f\left(x_1, x_2, \ldots, x_T | z^T, x_0, \alpha\right) = \prod_{t=1}^{T} f_1\left(x_t | z_t, x^{t-1}, \alpha\right).$$

But a is unobserved. We need to integrate it out. One solution is to parameterize the distribution of α conditional on z^T and x_0, say $h(\alpha|z^T,x_0)$. Then

$$\begin{aligned} & f\left(x_1, x_2, \ldots, x_T | z^T, x_0\right) \\ &= \int \prod_{t=1}^{T} f_1\left(x_t | z_t, x^{t-1}, \alpha\right) h\left(\alpha | z^T, x_0\right) d\alpha. \end{aligned}$$

Notice that in the traditional random effects approach (in the line of Anderson and Hsiao 1981) we would have to make assumptions about the conditional distribution of x_0 conditional on a and z^T.

See Also

- ► Fixed Effects and Random Effects
- ► Maximum Likelihood

Bibliography

Ahn, H., and J.L. Powell. 1993. Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics* 58: 3–29.

Andersen, E. 1970. Asymptotic properties of conditional maximum likelihood estimators. *Journal of the Royal Statistical Society, Series B* 32: 283–301.

Anderson, T., and C. Hsiao. 1981. Estimation of dynamic models with error components. *Journal of the American Statistical Association* 76(375): 598–606.

Butler, J.S., and R. Moffitt. 1982. A computationally efficient quadrature procedure for the one-factor multinomial probit model. *Econometrica* 50: 761–764.

Chamberlain, G. 1984. Panel data. In *Handbook of econometrics*, ed. Z. Griliches and M. Intrilligator, Vol. 2. Amsterdam: North-Holland.

Chamberlain, G. 1985. Heterogeneity, omitted variable bias, and duration dependence. In *Longitudinal analysis of labor market data*, ed. J.J. Heckman and B. Singer. Cambridge: Cambridge University Press.

Fristedt, B., and L. Gray. 1997. *A modern approach to probability theory*. Boston: Birkhauser.

Honoré, B.E. 1992. Trimmed LAD and least squares estimation of truncated and censored regression models with fixed effects. *Econometrica* 60: 533–565.

Honoré, B.E. 1993. Orthogonality conditions for Tobit models with fixed effects and lagged dependent variables. *Journal of Econometrics* 59: 35–61.

Honoré, B.E., and E. Kyriazidou. 2000a. Panel data discrete choice models with lagged dependent variables. *Econometrica* 68: 839–874.

Honoré, B.E., and E. Kyriazidou. 2000b. Estimation of Tobit-type models with individual specific effects. *Econometric Reviews* 19: 341–366.

Hu, L. 2002. Estimation of a censored dynamic panel data model. *Econometrica* 70: 2499–2517.

Kyriazidou, E. 1997. Estimation of a panel data sample selection model. *Econometrica* 65: 1335–1364.

Kyriazidou, E. 2001. Estimation of dynamic panel data sample selection models. *Review of Economic Studies* 68: 543–572.

Magnac, T. 2000. Subsidised training and youth employment: Distinguishing unobserved heterogeneity from state dependence in labour market histories. *Economic Journal* 110: 805–837.

Manski, C. 1987. Semiparametric analysis of random effects linear models from binary panel data. *Econometrica* 55: 357–362.

Newey, W. 1994. The asymptotic variance of semiparametric estimators. *Econometrica* 62: 1349–1382.

Neyman, J., and E.L. Scott. 1948. Consistent estimation from partially consistent observations. *Econometrica* 16: 1–32.

Powell, J.L. 2001. Semiparametric estimation of bivariate latent variable models. In *Nonlinear statistical modeling: Proceedings of the thirteenth International Symposium in Economic Theory and Econometrics: Essays in honor of Takeshi Amemiya*, ed. C. Hsiao, K. Morimune, and J.L. Powell. Cambridge: Cambridge University Press.

Wooldridge, J.M. 2000. A framework for estimating dynamic, unobserved effects panel data models with possible feedback to future explanatory variables. *Economics Letters* 68: 245–250.

Non-linear Programming

Michael D. Intriligator

Keywords

Activity analysis; Concave programming; Demand functions; Dual problem; Firm, theory of the; Hyperplanes; Kuhn–Tucker conditions; Lagrange multipliers; Linear programming; Mathematical programming; Nonlinear programming; Quadratic programming; Saddlepoints; Slater constraint qualification

JEL Classifications

C6

The problem of *nonlinear programming* is that of maximizing (or minimizing) a given function subject to a set of inequality constraints. Such problems arise in many areas of economics, such as the microeconomic theory of the household and the firm. It has also had wide applicability in game theory and operations research. Historically, the subject developed from the work of mathematicians, primarily John in studying extremum problems with inequalities as side constraints and Kuhn and Tucker who made the fundamental contribution of characterizing the nature of the solution to such problems (John 1948; Kuhn and Tucker 1951).

The nonlinear programming problem is a special case of the general *mathematical programming problem* of maximizing a function subject to constraints. The *linear programming problem* can be considered a special case of the nonlinear programming problem, namely one of maximizing a given linear form subject to a set of linear inequality constraints.

Mathematical Programming: Resource Allocation in Economics

The more general *problem of mathematical programming* is that of maximizing a function subject to constraints. Using standard notation the problem can be written

$$\max_{\mathbf{x}} \; F(\mathbf{x}) \quad \textit{subject to} \;\; \mathbf{x}\in X. \qquad (1)$$

Here **x** is a (column) vector of n choice variables $F(x_1, x_2, \ldots, x_n)'$ (the prime denoting the transpose of the row vector); $F(\mathbf{x})$ is a given real-valued function of these variables $F(x_1, x_2, \ldots, x_n)$; and X is a given non-empty subset of Euclidean n-space (the space of all n-tuples of real numbers) (Hadley 1964; Intriligator 1971, 1981; Aoki 1971; Luenberger 1973; Hestenes 1975).

In economics the vector **x** is frequently called the vector of *instruments*, the function $F(\mathbf{x})$ is frequently called the *objective function* (or *criterion function*), and the set X of feasible instrument vectors (**x** satisfying $\mathbf{x} \in X$) is frequently called the *opportunity set*. The basic economic problem of allocating scarce resources among competing ends can thus be interpreted as one of mathematical programming, where a particular resource allocation is represented by the choice of a particular vector of instruments, the scarcity of the resources is represented by the opportunity set, reflecting constraints on the instruments; and the competing ends are represented by the objective function, which gives the value attached to each of the alternative allocations. Problem (1) can therefore be interpreted in the language of economics as that of choosing instruments within the opportunity set so as to maximize the objective function (Lancaster 1968; Intriligator 1971, 1981; Bazaraa and Shetty 1976; Takayama 1985).

Nonlinear Programming

The *problem of nonlinear programming*, a special case of (1), is that of choosing non-negative values of n variables so as to maximize a function of these variables subject to inequality constraints. Using the same type of notation the problem is

$$\max_{\mathbf{x}} F(\mathbf{x}) \text{ subject to } \mathbf{g}(\mathbf{x}) \leq \mathbf{b}, \quad \mathbf{x} \geq 0. \qquad (2)$$

Here the vector of instruments $\mathbf{x}$ and the objective function $F(\mathbf{x})$are as in (1), where $F(\mathbf{x})$is assumed to be a real-valued continuously differentiable function of n variables. The vector-valued function $\mathbf{g(x)}$ is a representation of m constraint functions, $[g_1, (x_1, x_2, \ldots, x_n), g_2, (x_1, x_2, \ldots, x_n), \ldots, g_m, (x_1, x_2, \ldots, x_n)]'$ and b is a column vector of m constraint constants $(b_1, b_2, \ldots, b_m)$, so the m inequality constraints in (2) can be written

$$g_i(x_1, x_2, \ldots, x_n) \leq b_i, \quad i = 1, 2, \ldots, m. \qquad (3)$$

The n non-negativity constraints in (2) state that all n instruments are non-negative. Thus the problem of nonlinear programming can be written

$$\max F(x_1, x_2, \ldots, x_n) \quad \text{by choice of} \quad x_1, x_2, \ldots, x_n \qquad (4)$$

$$\text{subject to} \begin{cases} g_i(x_1, x_2, \quad \ldots, x_n) \leq b_i, & i = 1, 2, \quad \ldots, m \\ x_j \geq 0, & j = 1, 2, \quad \ldots, n \end{cases}.$$

This problem is a special case of (1) in which the opportunity set can be written

$$X = \{\mathbf{x} \in E^n \,|\, \mathbf{g}(\mathbf{x}) \leq \mathbf{b}, \quad \mathbf{x} \geq 0\} \qquad (5)$$

where E^n is Euclidean n-space. Thus the problem is one of maximizing a given function subject to $m + n$ constraints – m inequality constraints and n non-negativity constraints (Kuhn and Tucker 1951; Hadley 1964; Mangasarian 1969; Zangwill 1969; Intriligator 1971, 1981; Luenberger 1973; Hestenes 1975; Martos 1975; Avriel 1976; Bazaraa and Shetty 1979; McCormick 1983).

Linear Programming

In spite of the contradictory terminology, the *problem of linear programming* is in fact an important special case of nonlinear programming. Here the problem is that of choosing non-negative values of n variables so as to maximize a linear from in these variables subject to m linear inequality constraints

$$\max_{\mathbf{x}} \quad \mathbf{cx} \quad \text{subject to} \quad \mathbf{Ax} \leq \mathbf{b}, \quad \mathbf{x} \geq 0, \qquad (6)$$

where $\mathbf{A}$ is a given $m \times n$ matrix, $\mathbf{b}$ is a given $m \times 1$ column vector, and $\mathbf{c}$ is a given $1 \times n$ row vector. Thus the linear programming problem represents the special case of nonlinear programming (2) which is doubly linear, being linear in the objective function $\mathbf{cx} = \sum_{j=1}^{n} c_j x_j$ and in each of the constraint functions $g_i(x_1, x_2, \ldots, x_n) = \sum_{j=1}^{n} a_{ij} x_j$. Thus the problem of linear programming can be written

$$\max \sum_{j=1}^{n} c_j x_j \text{ by choice of } x_1, x_2, \ldots, x_n$$

$$\text{subject to} \begin{cases} \sum\limits_{j=1}^{n} a_{ij} x_j \leq b_i, \quad i = 1, 2, \ldots, \mathrm{m} \\ x_j \geq 0, \qquad j = 1, 2, \ldots, n. \end{cases} \qquad (7)$$

(Dorfman et al. 1958; Gale 1960; Hadley 1963; Dantzig 1963; Intriligator 1971; Luenberger 1973; Gass 1975.)

Kuhn–Tucker Conditions

The Kuhn–Tucker conditions provide a characterization of the solution to the problem of nonlinear programming (2). These conditions are defined in terms of the *Lagrangian function*

$$\begin{aligned} L(\mathbf{x},\mathbf{y}) &= F(\mathbf{x}) + \mathbf{y}[\mathbf{b} - \mathbf{g}(\mathbf{x})] \\ &= F(x_1, x_2, \ldots, x_n) \\ &\quad + \sum_{i=1}^{m} y_i \left[b_i - g_i(x_1, x_2, \ldots, x_n)^i \right]. \end{aligned} \quad (8)$$

Here **y** is a (row) vector of *m Lagrange multipliers* (sometimes written as λ's), one for each of the inequality constraints defined by the $g_i(x_1, x_2, \ldots, x_n)$ constraint functions and *bi* constraint constants in (3). The *Kuhn–Tucker conditions* are then defined at the point $\mathbf{x}^*$, $\mathbf{y}^*$ as the $2n + 2\ m$ inequalities and 2 equalities

$$\begin{aligned} &\frac{\partial L}{\partial \mathbf{x}}(\mathbf{x}^*, \mathbf{y}^*) \le 0, \frac{\partial L}{\partial \mathbf{y}}(\mathbf{x}^*, \mathbf{y}^*) \ge 0 (n\ + m \text{ conditions}) \\ &\quad \times \frac{\partial L}{\partial \mathbf{x}}(\mathbf{x}^*, \mathbf{y}^*)\mathbf{x}^* 0, \mathbf{y}^* \frac{\partial L}{\partial \mathbf{y}}(\mathbf{x}^*, \mathbf{y}^*) \\ &\quad = 0 (2 \text{ conditions}) \mathbf{x}^* \ge 0,\ \mathbf{y}^* \\ &\quad \ge 0 (n\ + m \text{ conditions}). \end{aligned} \quad (9)$$

Half of the inequalities represent the constraints of the original problem

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{y}}(\mathbf{x}^*, \mathbf{y}^*) &= \mathbf{b} - \mathbf{g}(\mathbf{x}^*) \\ &\ge 0\ (m \text{ conditions}) \end{aligned} \quad (10)$$

$$\mathbf{x}^* \ge 0\ (n \text{ conditions}), \quad (11)$$

while the added $n + m$ inequalities require that

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{x}}(\mathbf{x}^*, \mathbf{y}^*) &= \frac{\partial F}{\partial \mathbf{x}}(\mathbf{x}^*) - \mathbf{y}^* \frac{\partial g}{\partial \mathbf{x}}(\mathbf{x}^*) \\ &\le 0\ (n \text{ conditions}) \end{aligned} \quad (12)$$

$$\mathbf{y}^* \ge 0\ (m \text{ conditions}). \quad (13)$$

The first *n* conditions, in (12), state that

$$\frac{\partial F}{\partial x_j} - \sum_{i=1}^{m} y_i^* \frac{\partial g_i}{\partial x_j} \le 0,\ j = 1, 2, \ldots, n, \quad (14)$$

and they are written as inequalities because of the non-negativity restrictions on **x**, which allow for the possibility of boundary solutions. The last *m* conditions, in (13), state that the Lagrange multipliers are non-negative, and they stem from the fact that the constraints are written as inequalities rather than as equalities. (If any constraint is imposed as an equality, then the corresponding Lagrange multiplier y_i^* is unrestricted.)

The two equality Kuhn–Tucker conditions

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{x}}(\mathbf{x}^*, \mathbf{y}^*)\mathbf{x}^* &= \sum_{j=1}^{n} \left[\frac{\partial F}{\partial x_j}(\mathbf{x}^*) - \mathbf{y}^* \frac{\partial \mathbf{g}}{\partial x_j}(\mathbf{x}^*) \right] x_j^* \\ &= 0 \end{aligned} \quad (15)$$

$$\mathbf{y}^* \frac{\partial L}{\partial \mathbf{y}}(\mathbf{x}^*, \mathbf{y}^*) = \sum_{i=1}^{m} y_i^* [b_i - g_i(\mathbf{x}^*)] = 0 \quad (16)$$

together with the other conditions in (9) require that every term in both of these sums vanishes. These *complementary slackness conditions of nonlinear programming* require that when one of the inequality constraints is satisfied at the solution as a strict inequality then the corresponding (dual) variable equals zero at the solution

$$\begin{aligned} &\frac{\partial F}{\partial x_j}(\mathbf{x}^*) - \mathbf{y}^* \frac{\partial \mathbf{g}}{\partial x_j}(\mathbf{x}^*) < 0 \text{ implies } x_j^* \\ &\quad = 0, j = 1, 2, \ldots, n \end{aligned} \quad (17)$$

$$\begin{aligned} &b_i - g_i(\mathbf{x}^*) > 0 \quad \text{i.e.} \quad g_i(\mathbf{x}^*) < b_i, \\ &\qquad \text{implies}\, y_i^* = 0 \quad i = 1, 2, \ldots, m. \end{aligned} \quad (18)$$

At the solution the value of the Lagrangian is the maximized value of the objective function

$$L(\mathbf{x}^*, \mathbf{y}^*) = F(\mathbf{x}^*) = F^*, \quad (19)$$

and the solutions for the Lagrange multipliers are to be interpreted as the sensitivities of the maximized value of the objective function to changes in the constraint constants

$$\mathbf{y}^* = \frac{\partial F^*}{\partial \mathbf{b}}, \quad \text{i.e.} \quad {y^*}_i = \frac{\partial F^*}{\partial b_i}, \quad i = 1, 2, \ldots, m. \tag{20}$$

In particular, from the complementary slackness conditions (18), if a constraint is met as a strict inequality at the solution then the corresponding Lagrange multiplier is zero, so increasing the constraint constant by a 'small' amount will not change the maximized value of the objective function.

If a suitably strong *constraint qualification condition* is satisfied the Kuhn–Tucker conditions are necessary conditions for the nonlinear programming problem in that if $\mathbf{x}^*$ solves (2) then there exists a vector of Lagrange multipliers $\mathbf{y}^*$ satisfying (9). There are, in fact, many alternative forms of the constraint qualification condition. One is the *Slater constraint qualification* requiring that there exist a point $\mathbf{x}^\mathbf{0} \geq \mathbf{0}$ such that $\mathbf{g}(\mathbf{x}^0) < \mathbf{b}$, that is, there exists a non-negative point at which all inequality constraints are satisfied as strict inequalities, thus excluding outward pointing cusps (Arrow et al. 1958, 1961; Mangasarian 1969; Zangwill 1969; Bazaraa et al. 1972; Bazaraa and Shetty 1976, 1979). For problems not satisfying the constraint qualification condition it is necessary to add another Lagrange multiplier y_0, for the objective function (John 1948).

As to sufficiency, a sufficient condition for $\mathbf{x}^*$ to solve the nonlinear programming problem (2) is that there exists a $\mathbf{y}^*$ such that $\mathbf{x}^*$, $\mathbf{y}^*$ solves the *saddle point problem*

$$\max_{\mathrm{x}} \min_{\mathrm{y}} L(\mathbf{x}, \mathbf{y}) \quad \text{subject to} \quad \mathbf{x} \geq \mathbf{0}, \mathbf{y} \geq \mathbf{0}, \tag{21}$$

where $\mathbf{x}^*$, $\mathbf{y}^*$ solves this problem if and only if, for all $\mathbf{x} \geq 0$, $\mathbf{y} \geq 0$

$$L(\mathbf{x}, \mathbf{y}^*) \leq L(\mathbf{x}^*, \mathbf{y}^*) \leq L(\mathbf{x}^*, \mathbf{y}). \tag{22}$$

Thus, if a pair of vectors $\mathbf{x}^*$, $\mathbf{y}^*$ satisfies (22) then $\mathbf{x}^*$ solves the nonlinear programming problem. Conversely, assuming both that a suitable constraint qualification condition is met and that the problem is one of *concave programming* in which $F(\mathbf{x})$ is a concave function and each constraint function $g_i(\mathbf{x})$ is a convex function, then if $\mathbf{x}^*$solves the nonlinear programming problem (2) there exists a nonzero vector $\mathbf{y}^*$ such that $\mathbf{x}^*$, $\mathbf{y}^*$ solves the saddle point problem (21) and the two problems are equivalent. In fact, if the problem is one of concave programming and a suitable constraint qualification condition is met, then the nonlinear programming problem (2), the problem of finding a solution to the Kuhn–Tucker conditions (9), and the saddle point problem (21) are all equivalent in that if $\mathbf{x}^*$ solves (2) then and only then there exists a $\mathbf{y}^*$ such that $\mathbf{x}^*$, $\mathbf{y}^*$ solves both (9) and (21).

Various computational approaches have been developed to solve nonlinear programming problems, and such approaches, in the form of computer codes, are widely available and routinely used to solve particular problems (Mangasarian 1969; Zangwill 1969; Polak 1971; Avriel 1976; Bazaraa and Shetty 1979; Schittkowski 1980; Dennis 1984).

N

The Kuhn–Tucker conditions imply that for an interior solution $\mathbf{x}^* > 0$ (or for a problem in which the non-negativity of the x's is not part of the problem)

$$\frac{\partial F}{\partial \mathbf{x}}(\mathbf{x}^*) = \mathbf{y}^* \frac{\partial \mathbf{g}}{\partial \mathbf{x}}(\mathbf{x}^*). \tag{23}$$

Thus at the solution the gradient vector of the objective function (the vector of first-order derivatives of the objective function, $\partial F/\partial \mathbf{x}(\mathbf{x}^*)$) must be a non-negative weighted combination of the gradient vectors of the constraint functions, the weights being the Lagrange multipliers. Geometrically this condition means that the gradient vector of the objective function must, at the solution, lie within the cone spanned by the outward pointing normals to the opportunity set, where the gradient vectors for the constraint functions define the outward pointing normals to the opportunity set.

For the special case of linear programming (6) the saddlepoint problem is

$$\max_{\mathbf{x}} \quad \max_{\mathbf{y}} \ L(\mathbf{x}, \mathbf{y}) = \mathbf{cx} + \mathbf{y}(\mathbf{b} - \mathbf{Ax}) \tag{24}$$

and the Kuhn–Tucker conditions (9) are

$$\begin{gathered}\frac{\partial L}{\partial \mathbf{x}} = \mathbf{c} - \mathbf{y}^{*}\mathbf{A} \leq 0, \quad \frac{\partial L}{\partial \mathbf{y}} = \mathbf{b} - \mathbf{Ax}^{*} \geq 0 \\ \frac{\partial L}{\partial \mathbf{x}}\mathbf{x}^{*} = (\mathbf{c} - \mathbf{y}^{*}\mathbf{A})\mathbf{x}^{*} = 0, \quad \mathbf{y}^{*}\frac{\partial L}{\partial \mathbf{y}} = \mathbf{y}^{*}(\mathbf{b} - \mathbf{Ax}^{*}) = 0 \\ \mathbf{x}^{*} \geq \mathbf{0}, \qquad \mathbf{y}^{*} \geq \mathbf{0},\end{gathered} \tag{25}$$

and they characterize the solution. The same conditions form the Kuhn–Tucker conditions for the *dual problem*

$$\min_{\mathbf{y}} \ \mathbf{yb} \ \text{ subject to } \ \mathbf{yA} \geq \mathbf{c}, \mathbf{y} \geq 0, \tag{26}$$

where the variables of the dual problem, **y**, are the Lagrange multipliers of the original (primal) problem. The dual problem uses the same matrix **A** and vectors **b** and **c** as the primal problem, but it is one of minimization, rather than maximization, and the constraint constants of the primal problem become the coefficients of the objective function of the dual problem while the coefficients of the objective function of the primal problem become the constraint constants of the dual problem. Vectors $\mathbf{x}^{*}$, $\mathbf{y}^{*}$ satisfying (24) or (25) solve both the primal problem (6) and dual problem (26). Geometrically, in the case of linear programming the opportunity set is a polyhedral closed convex set since it is the intersection of $m + n$ half spaces defined by the m inequality and n non-negativity constraints. The contours of the linear objective function are hyperplanes, and the problem is solved on the highest hyperplane within the polyhedral set. A solution must occur at a vertex, in which case it is unique, or along a bounding face, in which case it is non-unique. As in the more general case of nonlinear programming, however, the solution always occurs at a point where the gradient vector of the objective function (here **c**) lies in the cone spanned by the outward pointing normals to the opportunity set (here the relevant columns of **A** or the relevant outward pointing unit vectors corresponding to the non-negativity constraints).

Another special case of nonlinear programming, one which subsumes linear programming, is that of quadratic programming. The *problem of quadratic programming* is that of

$$\max_{\mathrm{x}} F(\mathbf{x}) = \mathbf{cx} + \frac{1}{2}\mathbf{x}'\mathrm{Q}\mathbf{x} \quad \text{subject to} \quad \mathbf{Ax} \leq \mathbf{b}, \quad \mathbf{x} \geq \mathbf{0} \tag{27}$$

where **A**, **b**, **c** are as in the linear programming problem and **Q** is a given $n \times n$ negative semidefinite symmetric matrix. The problem is one of concave programming because **Q** is negative semidefinite and the linear transformation **Ax** is convex. Furthermore the constraint qualification is met. The Kuhn–Tucker conditions are therefore both necessary and sufficient, that is, the vector $\mathbf{x}^{*}$ solves (27) if and only if there is a vector of Lagrange multipliers $\mathbf{y}^{*}$such that the pair $\mathbf{x}^{*}$, $\mathbf{y}^{*}$ satisfies the Kuhn–Tucker conditions

$$\begin{gathered}\frac{\partial L}{\partial \mathbf{x}} = \left(\mathbf{c} + \mathbf{x}^{*\prime}\mathrm{Q} - \mathbf{y}^{*}\mathbf{A}\right) \leq 0, \quad \frac{\partial L}{\partial \mathbf{y}} = \mathbf{b} - \mathbf{Ax}^{*} \geq 0 \\ \frac{\partial L}{\partial \mathbf{x}}\mathbf{x}^{*} = \left(\mathbf{c} + \mathbf{x}^{*\prime}\mathrm{Q} - \mathbf{y}^{*}\mathbf{A}\right)\mathbf{x}^{*} = 0, \quad \mathbf{y}^{*}\frac{\partial L}{\partial \mathbf{y}} = \mathbf{y}^{*}(\mathbf{b} - \mathbf{Ax}^{*}) = 0 \\ \mathbf{x}^{*} \geq \mathbf{0}, \qquad \mathbf{y}^{*} \geq \mathbf{0},\end{gathered} \tag{28}$$

where **L(x, y)** is the Lagrangian function

$$L(\mathbf{x}, \mathbf{y}) = \mathbf{cx} + \frac{1}{2}\mathbf{x}'\mathbf{Qx} + \mathbf{y}(\mathbf{b} - \mathbf{Ax}). \tag{29}$$

This case reduces to that of linear programming if **Q**, the matrix defining the quadratic form $\frac{1}{2}\mathbf{x}'\mathbf{Qx}$ in (27) vanishes.

Neoclassical Theory of the Household and the Firm

The problem of nonlinear programming can be applied in economics to the neoclassical theory of both the household and the firm, the two most important units of microeconomics. For the household, assume there are n goods (and services) available where $\mathbf{x} = (x_1, x_2, \ldots, x_n)'$ is the column vector of goods purchased and consumed by the household. Assume further that the household seeks to maximize a utility function, a real-valued function defined on these goods $U(\mathbf{x}) = U(x_1, x_2, \ldots, x_n)$. Assume finally that the household purchases non-negative quantities of each good so as to maximize the utility function subject to a budget constraint that states that expenditure on all n goods cannot exceed available income.

The *neoclassical problem of the household* is then

$$\max_{\mathbf{x}} \; U(\mathbf{x}) \text{ subject to } \mathbf{px} \leq I, \;\; \mathbf{x} \geq 0. \qquad (30)$$

Here $\mathbf{p}$ is a given row vector of (positive) prices of each of the n goods, and I is the given (positive) income available to the household. Thus the household chooses non-negative amounts of goods $\mathbf{x}$ so as to maximize the utility function $U(\mathbf{x})$ subject to the budget constraint $\mathbf{px} \leq I$, which states that expenditure on all n goods cannot exceed income.

This problem is one of nonlinear programming, so, introducing the (single) Lagrange multiplier y the Lagrangian is

$$L(\mathbf{x}, y) = U(\mathbf{x}) + y(I - \mathbf{px}). \qquad (31)$$

The Kuhn–Tucker conditions characterize an optimum point. Under the further regularity conditions that $\mathbf{x}^* > 0$ and $U(\mathbf{x})$ is a twice continuously differentiable function in a neighbourhood of $\mathbf{x}^*$ with a non-singular Hessian matrix of second-order derivatives there then exist solutions for the purchases of goods $\mathbf{x}^*$ as functions of the $n + 1$ parameters $\mathbf{p}$, I, which are the *demand functions* characterizing the optimum point (Hicks 1946; Samuelson 1947; Wold and Jureen 1953; Intriligator 1971; Barten and Böhm 1982; Phlips 1983).

For the firm, assume that the firm uses n inputs to produce a single output, where $\mathbf{x} = (x_1, x_2, \ldots, x_n)'$ is a column vector of inputs, q is output, and $f(\mathbf{x})$ is the production function of the firm. Assume further that the firm seeks to maximize a profit function, given as the difference between revenue and cost. Assume finally that the firm purchases non-negative inputs and produces non-negative output subject to the technology of the given production function so as to maximize profit. The *neoclassical theory of the firm* is then

$$\max_{\mathrm{x}} \pi(\mathbf{x}) = pf(\mathbf{x}) - \mathbf{wx} \quad \text{subject to} \quad \mathbf{x} \geq \mathbf{0}. \qquad (32)$$

Here $pf(\mathbf{x})$ is total revenue, p being the given price of output, and $\mathbf{wx}$ is the cost of production, the total expenditure on all inputs, $\mathbf{w}$ being the vector of given prices (wages) of inputs.

This problem is also one of nonlinear programming, and the Kuhn–Tucker conditions characterize an optimum point. Under the further regularity conditions that $\mathbf{x}^* > 0$ and $f(\mathbf{x})$ is twice continuously differentiable in a neighbourhood of $\mathbf{x}^*$ with a non-singular Hessian matrix of second-order derivatives there exist solutions for the purchase of inputs $\mathbf{x}^*$ and production of output q as functions of the $n + 1$ parameters $\mathbf{w}$, p which are the *input demand functions* and *output supply function* characterizing the optimum point (Hicks 1946; Samuelson 1947; Intriligator 1971; Nadiri 1982).

The problem of linear programming can be applied to a firm that produces output using an *activity analysis* technology. In such a case the firm produces n outputs $x_1, x_2, \ldots, x_n$ using m inputs $b1, b2, \ldots, bm$. To produce one unit of output x_j requires a_{ij} units of input i. In the short run all inputs are fixed so the only choice for the firm is that of deciding what mix of outputs to produce given these inputs. The problem is then

$$\max_{\mathbf{x}} \quad \mathbf{cx} \text{ subject to } \quad \mathbf{Ax} \leq \mathbf{b}, \mathbf{x} \geq 0, \qquad (33)$$

as in (6). Here the objective function to be maximized is total revenue, where c_j is the given price of output j, so the problem is one of choosing non-negative outputs so as to maximize profit, given the technology (the a_{ij}) and the inputs (the b_i). The dual problem is

$$\min_{y} \quad \mathbf{yb} \text{ subject to } \quad \mathbf{yA} \geq \mathbf{c}, \mathbf{y} \geq 0, \qquad (34)$$

as in (26), which can be interpreted as choosing non-negative values (shadow prices) for the inputs $y_1, y_2, \ldots, y_m$ so as to minimize the cost of the inputs $\mathbf{yb} = \sum y_i b_i$ where y_i is the chosen value and b_i is the given level of input i. The n constraints state that the unit cost of good j, obtained by summing the cost of producing one unit over all inputs, is no less than the price of this good. The dual to a problem of allocation, the primal problem (33), is one of valuation, the dual problem (34). According to the complementary slackness conditions in (25) if for any output j unit costs exceeds price (that is, the output is produced at a loss) then this output is not produced $\left(x_j^* = 0\right)$ and if for any input i not all of the input is used then it is valued at zero $\left(y_i^* = 0\right)$.

In conclusion, the problem of nonlinear programming is one with important applications to the microeconomic theory of the household and the firm, leading to conditions characterizing an equilibrium at an optimum point. The problem also has many other applications throughout economics.

See Also

- ▶ Calculus of Variations
- ▶ Convex Programming
- ▶ Demand Theory
- ▶ Duality
- ▶ Functional Analysis
- ▶ Lagrange Multipliers
- ▶ Linear Programming

Bibliography

Aoki, M. 1971. *Introduction to optimization techniques*. New York: Macmillan.

Arrow, K.J., L. Hurwicz, and H. Uzawa. 1958. Constraint qualifications in maximization problems. In *Studies in linear and nonlinear programming*, ed. K.J. Arrow, L. Hurwicz, and H. Uzawa. Stanford: Stanford University Press.

Arrow, K.J., L. Hurwicz, and H. Uzawa. 1961. Constraint qualifications in maximization problems. *Naval Research Logistics Quarterly* 8: 175–191.

Avriel, M. 1976. *Nonlinear programming: Analysis and methods*. Englewood Cliffs: Prentice-Hall.

Barten, A.P., and V. Böhm. 1982. Consumer theory. Chapter 9. In *Handbook of Mathematical Economics*, ed. K.J. Arrow and M.D. Intriligator, vol. 2. Amsterdam: North-Holland.

Bazaraa, M.S., and C.M. Shetty. 1976. *Foundations of optimization*. Berlin: Springer-Verlag.

Bazaraa, M.S., and C.M. Shetty. 1979. *Nonlinear programming: Theory and algorithms*. New York: Wiley.

Bazaraa, M.S., J.J. Goode, and C.M. Shetty. 1972. Constraint qualifications revisited. *Management Science* 18: 567–573.

Dantzig, G. 1963. *Linear programming and extensions*. Princeton: Princeton University Press.

Dennis, J.E. 1984. A user's guide to nonlinear optimization algorithms. *Proceedings IEEE* 72: 1765–1776.

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

Gale, D. 1960. *The theory of linear economic models*. New York: McGraw-Hill.

Gass, S.I. 1975. *Linear programming: Methods and applications*. 4th ed. New York: McGraw-Hill.

Hadley, G. 1963. *Linear programming*. Reading: Addison-Wesley.

Hadley, G. 1964. *Nonlinear and dynamic programming*. Reading: Addison- Wesley.

Hestenes, M.R. 1975. *Optimization theory: The finite dimensional case*. New York: Wiley.

Hicks, J.R. 1946. *Value and capital*. 2nd ed. New York: Oxford University Press.

Intriligator, M.D. 1971. *Mathematical optimization and economic theory*. Englewood Cliffs: Prentice-Hall.

Intriligator, M.D. 1981. Mathematical programming with applications to economics. Chapter 2 in *Handbook of mathematical economics*, vol. 1, ed. K.J. Arrow and M.D. Intriligator. Amsterdam: North-Holland.

John, F. 1948. Extremum problems with inequalities as side conditions. In *Studies and essays: courant anniversary volume*, ed. K.O. Friedrichs, O.W. Neugebauer, and J.J. Stoker. New York: Interscience Publishers.

Kuhn, H.W., and A.W. Tucker. 1951. Nonlinear programming. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, ed. J. Neyman. Berkeley: University of California Press.

Lancaster, K. 1968. *Mathematical economics*. New York: Macmillan.

Luenberger, D.G. 1973. *Introduction to linear and nonlinear programming*. Reading: Addison-Wesley.

Mangasarian, O.L. 1969. *Nonlinear programming*. New York: McGraw-Hill.

Martos, B. 1975. *Nonlinear programming: Theory and methods*. Amsterdam: North-Holland.

McCormick, G.P. 1983. *Nonlinear programming: Theory, algorithms, and applications*. New York: Wiley.

Nadiri, M.I. 1982. Producers theory. Chapter 10. In *Handbook of mathematical economics*, ed. K.J. Arrow and M.D. Intriligator, vol. 2. Amsterdam: North-Holland.

Phlips, L. 1983. *Applied consumption analysis*. Revised ed. Amsterdam: North-Holland.

Polak, E. 1971. *Computational methods in optimization: A unified approach*. New York: Academic Press.

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

Schittkowski, K. 1980. *Nonlinear programming codes*. Berlin: Springer-Verlag.
Takayama, A. 1985. *Mathematical economics*. 2nd ed. New York: Cambridge University Press.
Wold, H., and L. Jureen. 1953. *Demand analysis*. New York: Wiley.
Zangwill, W.I. 1969. *Nonlinear programming: A unified approach*. Englewood Cliffs: Prentice-Hall.

Non-linear Time Series Analysis

Bruce Mizrach

Keywords
Chaos; Cointegration; Ergodicity; Forecasting; GARCH models; Inference; Kernel estimators; Linear models; Lyapunov exponents; Markov switching models; Nonlinear time series analysis; Nonparametric estimation; Piecewise linear models; Regime dependence; Regime switching; Representation theory; Stochastic volatility models; Testing; Threshold autoregressions; Volterra expansion; Wavelets

JEL Classifications
C1

Since the early 1980s, there has been a growing interest in stochastic nonlinear dynamical systems of the form

$$x_{t+1} = f\left(x_t, x_{t-1}, \ldots, x_{t-p}\right) + \sigma(x_t)\varepsilon_t, \qquad (1)$$

where $\{x_t\}_{t=0}^{\infty}$ is a zero mean, covariance stationary process, $f\colon R^{p+1} \to R$, σ is the conditional volatility, and $\{\varepsilon_t\}_{t=0}^{\infty}$ is an independent and identically distributed noise process. The major recent developments in nonlinear time series are described here using this canonical model. The first section develops representation theory for a third order approximation. Nonparametric approaches follow; these rely on series expansions of the general model. Ergodic properties including path dependence and dimension are considered next. I then consider two widely utilized parametric models, piecewise linear models of f and autoregressive models for volatility. I conclude with a discussion of hypothesis testing and forecasting.

Volterra Expansion

There is no general causal representation for nonlinear time series as in the linear case. Series approximations rely on the *Volterra expansion*,

$$\begin{aligned} x_{t+1} \simeq f(0) +_{i=1}^{p} f_{i_1} x_{t-i_1} \\ +_{i_1=1}^{p}\ i_2 = i_1{}^{p} f_{i_1 i_2} x_{t-i_1} x_{t-i_2} \\ +_{i_1=1}^{p}\ i_2 = i_1{}^{p} \\ +_{i_3=i_2}^{p} f_{i_1 i_2 i_3} x_{t-i_1} x_{t-i_2} x_{t-i_3} + \cdots \end{aligned} \qquad (2)$$

Brockett (1976) shows any continuous map over [0; T] can be approximated by a finite Volterra series. Mittnik and Mizrach (1992) examine forecasts using generalized polynomial expansions like (2). Potter (2000) shows that in the cubic case, a one-sided Wold-type representation in terms of white noise v_t can be obtained,

$$\begin{aligned} x_{t+1} \simeq_{i=1}^{\infty}\ g_{i_1} v_{t-i_1} \\ +_{i_1=1}^{\infty}\ i_2 = i_1{}^{\infty} g_{i_1 i_2} v_{t-i_1} v_{t-i_2} \\ +_{i_1=1}^{\infty}\ {}_{i_2=i_1}^{\infty} i_3 = i_2{}^{\infty} g_{i_1 i_2 i_3} x_{t-i_1} x_{t-i_2} x_{t-i_3}. \end{aligned} \qquad (3)$$

Koop et al. (1996) note that the impulse response functions, $E[x_{t+n}|x_t, v_t] - E[x_{t+n}|x_t]$ will depend upon the size and sign of v_t as well as the current state x_t.

I now turn to nonparametric approaches which build on approximations like 2.

Nonparametric Estimation

Consider the local polynomial approximation to $f(.)$ around x_0,

$$\hat{f}(x) =_{j=0}^{m} \beta_j (x - x_0)^j. \qquad (4)$$

In the case $j = 0$, this corresponds to the *kernel regression* estimator of Nadaraya and Watson,

$$\hat{f}(x) = \frac{{}_{t=1}^{T} x_{t+1} K_h(x_t - x_0)}{{}_{t=1}^{T} K_h(x_t - x_0)}. \tag{5}$$

The K_h are *kernels*, usually functions with a support on a compact set, assigning greater weight to observations closer to x_0. h is the *bandwidth* parameter, determining the size of the histogram bin. *Nearest neighbours* estimation is the case where h is adjusted to find a fixed number of nearby observations k.

More generally, the *local linear approximation* solves,

$$\alpha_0, \beta_0 \min (x_{t+1} - \alpha_0 - \beta_0 (x_t - x_0))^2 K_h(x_t - x_0). \tag{6}$$

The estimator (5) corresponds to the case where the only regressor in (6) is the constant term.

The application of these methods in the time series case is a fairly recent development. Conditions for consistency and asymptotic normality rely on *mixing conditions* where the dependence between x_{t+j} and x_t becomes negligible as j grows large.

A closely related approach involves the use of a *recurrent neural network*,

$$\begin{array}{ll} \Psi_i(x_t, h_{t-1}) & = \Psi\left(\gamma_{i0} + \gamma_{i1} x_t +_{k=1}^{r} \delta_{ik} h_{k,t-k}\right), \\ x_{t+1} & = \Phi\left(\beta_0 +_{i=1}^{p} \beta_i \Psi_i(x_t, h_{t-1})\right). \end{array} \tag{7}$$

Kuan et al. (1994) provide convergence results for bounded Ψ (most commonly the logistic) as p grows large.

A popular approach in the frequency domain is wavelets. The *discrete wavelet transform* is

$$x_{t+1} =_{j=-\infty}^{\infty} k = -\infty^{\infty} \gamma(j,k) \Psi_{j,k}(t), \tag{8}$$

where the *mother wavelet* $\Psi(t)$,

$$\Psi_{j,k}(t) = \frac{1}{\sqrt{s_0^j}} \Psi\left(\frac{t - k\tau_0 s_0^j}{s_0^j}\right), \tag{9}$$

is parameterized by scale s_0 and translation τ, and the wavelet coefficients are given by

$$\gamma(j,k) = \left\langle \Psi_{j,k}(t), x(t) \right\rangle. \tag{10}$$

Daubechies (1992) orthonormal basis functions,

$$E\left[\Psi_{j,k}(t) \Psi_{m,n}(t)\right] = 0, \ \forall j \neq m, k \neq n, \tag{11}$$

have received the widest application.

Even when very little is known about f or σ, nonlinear time series analysis can shed light on the long run average or *ergodic* properties of the dynamical system.

Ergodic Properties

Mathematicians have known since Poincaré that even simple maps like (1) can produce very complex dynamics. The nonlinear time series literature has developed tools for estimation of ergodic properties of these systems. Denote by $Df(\bar{x})$ the Jacobian matrix of partial derivatives of (1),

$$\begin{bmatrix} \partial f_1 / \partial x_1 & \cdots & \partial f_1 / \partial x_p \\ \vdots & \ddots & \vdots \\ \partial f_p / \partial x_1 & \cdots & \partial f_p / \partial x_p \end{bmatrix} \tag{12}$$

evaluated at $(\bar{x})$: Replacing 12 with a sample analog,

$$J_t = \begin{bmatrix} \Delta f_1 / \Delta x_{1,t} & \cdots & \Delta f_1 / \Delta x_{p,t} \\ \vdots & \ddots & \vdots \\ \Delta f_p / \Delta x_{1,t} & \cdots & \Delta f_p / \Delta x_{p,t} \end{bmatrix} \tag{13}$$

we compute eigenvalues V_i,

$$V_i\left(Q_T' Q_T\right) \tag{14}$$

rank ordered from 1,. . ., p, where

$$Q_T = J_{T-p} \cdot J_{T-p-1} \cdots J_1 \tag{15}$$

The *Lyapunov exponents* are defined for the positive eigenvalues V_i^+ as

$$T \to \infty \lim \lambda_i = \frac{1}{2(T-p)} \ln V_i^+, \tag{16}$$

and a single exponent greater than 1 characterizes a system with sensitive dependence.

Popularly known as ‘chaos’, this property implies that dynamic trajectories become unpredictable even when the state of the system is known with certainty. Gençay and Dechert (1992) and Shintani and Linton (2004) provide methods for estimating these. Shintani and Linton (2003, 2004) reject the presence of positive Lyapunov exponents in both real output and stock returns.

The sum of the Lyapunov exponents also provides a measure of the Kolmogorov–Sinai *entropy* of the system. This tells the researcher how quickly trajectories separate. Mayfield and Mizrach (1991) estimate this time at about 15 minutes for the S&P 500 index.

A final quantity of interest is the *dimension p* of the dynamical system. Nonlinear econometricians try to estimate the dimension from a scalar m-history. A powerful result due to Takens (1981) says this can be done as long as $m \geq 2p + 1$. Diks (2004) has shown that the scaling of correlation exponents seems to be consistent with the stochastic volatility model.

A great deal of progress has been made with parametric models of (1) as well. I begin with the widely utilized piecewise linear models.

Piecewise Linear Models

The most widely applied parametric nonlinear time series specification has been the Markov switching model introduced by James Hamilton (1989). The function f is a piecewise linear function,

$$f(x_t) = \left\{ \begin{array}{l} \mu^{(1)} + \Sigma_{j=0}^{p} \varphi_j^{(1)} \left(x_{t-j} - \mu^{(1)}, S_t = s_t^{(1)} \right) \\ \vdots \\ \mu^{(m)} + \Sigma_{j=0}^{p} \varphi_j^{(m)} \left(x_t - \mu^{(m)}, S_t = s_t^{(m)} \right) \end{array} \right\}, \tag{17}$$

where the changes among states are governed by an unobservable regime switching process, $S_t = s_t^{(i)}, i = 1, \ldots m$, an $m \times m$ transition matrix Π, and $E\left[x_t \middle| S_t = s_t^{(i)} \right] = \mu^{(i)}$. When S_t is unobserved, $Pr(S_t|x_{t\,-\,1})$ is nonlinear in x_{t-1}. Hamilton has shown that a two-dimensional switching model describes well the business cycle dynamics in the United States. This model has been extended to include regime dependence in volatility (Kim 1994) and time varying transition probabilities (Filardo 1994).

The latent state vector requires forming prior and posterior estimates of which regime you are in. The EM algorithm (Hamilton 1990) and Bayesian Gibbs sampling methods (Albert and Chib, 1993) have proven fruitful in handling this problem. Hypothesis testing is also non-standard because under the alternative of m 1 regimes, the conditional mean parameters are nuisance parameters. Hansen (1996) has explored carefully these issues.

A closely related framework is the *threshold autoregressive* (TAR) model,

$$f(x_t) = \left\{ \begin{array}{c} \left[\mu^{(1)} + \Sigma_{j=0}^{p} \varphi_j^{(1)} \left(x_{t-j} - \mu^{(1)} \right) \right] I\left(q\left(x_{t-d}, Z_t \right) \leq \gamma_1 \right) \\ \left[\mu^{(2)} + \Sigma_{j=0}^{p} \varphi_j^{(2)} \left(x_{t-j} - \mu^{(2)} \right) \right] I\left(\gamma_1 < q\left(x_{t-d}, Z_t \right) \leq \gamma_2 \right) \\ \vdots \\ \left[\mu^{(m)} + \Sigma_{j=0}^{p} \varphi_j^{(m)} \left(x_t - \mu^{(m)} \right) \right] I\left(q\left(x_{t-d}, Z_t \right) > \gamma_{m-1} \right) \end{array} \right\}. \tag{18}$$

$I(.)$ is the indicator function, and $q(x_{t-d}, Z_t)$, the regime switching variable, is assumed to be an observable function of exogenous variables Z_t and lagged x’s. The integer d is known as the *delay parameter*. When q depends only upon x, the model is called *self-exciting*.

Teräsvirta (1994) has developed a two-regime version of the TAR model in which regime changes are governed by a smooth transition function

$$G(x_{t\,-\,d}, Z_t) : R^k \rightarrow [0, 1],$$

$$f(x_t) = G(x_{t-d}, Z_t)_{j=0}^{p}\varphi_j^{(1)}\left(x_{t-j} - \mu^{(1)}\right)$$
$$+ (1 - G(x_{t-d}, Z_t))_{j=0}^{p}\varphi_j^{(2)}\left(x_{t-j} - \mu^{(2)}\right). \tag{19}$$

Luukkonen, Saikkonen and Teräsvirta (1988) have shown that inference and hypothesis testing in this model is often much simpler than in the piecewise linear models. Van Dijk and Franses (1999) have extended this model to multiple regimes. Applications of this framework have been widespread from macroeconomics (Teräsvirta and Anderson 1992) to empirical finance (Franses and van Dijk 2000).

Krolzig (1997) considers the multivariate case where $x_t = (x_{1,t}, x_{2,t}, \ldots, x_{k,t})'$ is $k \times 1$. Balke and Fomby (1997) introduced threshold cointegration by incorporating error correction terms into the thresholds. Koop, Pesaran and Potter (1996) develop a bivariate model of US GDP and unemployment where the threshold depends upon the depth of the recession.

I now turn to models that introduce nonlinearity through the error term.

Models of Volatility

Engle and Bollerslev have introduced the generalized autoregressive conditional heteroskedasticity (GARCH) model,

$$h_t = \alpha_0 +_{i=1}^{q} \alpha_i \sigma^2(x_{t-i})\varepsilon_{t-i}^2 +_{i=1}^{p} \beta_i h_{t-i}, \tag{20}$$

where $h_t = E[(x_t - E[x_t| \Omega_{t-1}])^2| \Omega_{t-1}]$ is the *conditional variance*. This is just a Box–Jenkins model in the squared residuals of 1 of order (max [p, q], p). The model is nonlinear because the disturbances are uncorrelated, but their squares are not.

The GARCH model describes the volatility clustering and heavy-tailed returns in financial market data, and has found wide application in asset pricing and risk management applications.

Volatility modelling has been motivated by the literature on options pricing. Popular alternatives to the GARCH model include the stochastic volatility (SV) model (Ghysels et al. 1996), and the realized volatility approach of Andersen et al. (2003) and Barndorff-Nielsen and Shephard (2002). The discretetime SV model takes the form,

$$x_t = \sigma_\varepsilon \exp(h_t/2)\varepsilon_t, \tag{21}$$
$$h_t = \beta h_{t-1} + \sigma_h \eta_t,$$

where x_t is the demeaned log asset return, and ε_t and η_t are noise terms. Realized volatility sums high-frequency squared returns as an approximation of lower frequency volatility. Both GARCH and SV have been successful in explaining the departures from the Black–Scholes observed empirically.

The final two sections address the marginal contribution of nonlinear modelling to goodness of fit and forecasting.

Testing for Linearity and Gaussianity

There is a large literature on testing the importance of the nonlinear components of a model. The most widely used test is due to Brock, Dechert, Scheinkman and LeBaron (BDSL 1996). Their nonparametric procedure is built upon U-statistics. Serfling (1980) is a good introduction.

The first step is to form m-histories of the data,

$$x_t^m = (x_t, x_{t+1}, \ldots, x_{t+m-1}), \tag{22}$$

with joint distribution $F\left(x_t^m\right)$. Introduce the kernel $h : R^m \times R^m \to R$,

$$h\left(x_t^m, x_s^m\right) = I\left(x_t^m, x_s^m, \varepsilon\right)$$
$$\equiv I\left[\left\|x_t^m - x_s^m\right\| < \varepsilon\right], \tag{23}$$

where I(.) is the indicator function. The *correlation integral* of Grassberger and Procaccia (1983),

$$C(m, \varepsilon) \equiv \int_X \int_X I\left(x_t^m, x_s^m, \varepsilon\right) dF\left(x_t^m\right) dF\left(x_s^m\right), \tag{24}$$

is the expected number of m-vectors in an ε neighbourhood. A U-statistic,

$$C(m,N,\varepsilon) \equiv \frac{2}{N(N-1)} {}_{t=1}^{N-1} {}_{s=t+1}^{N} I\left(X_t^m, X_s^m, \varepsilon\right), \tag{25}$$

is a consistent estimator of 24. BDSL demonstrate the asymptotic normality of the statistic

$$\sqrt{N}\frac{S(m,N,\varepsilon)}{\sqrt{Var[S(m,N,\varepsilon)]}}\sqrt{N}d\times \rightarrow N(0,1), \tag{26}$$

where

$$S(m,N,\varepsilon) = C(m,N,\varepsilon) - C(m,N,\varepsilon)^m. \tag{27}$$

There is a multi-dimensional extension due to Baek and Brock (1992). De Lima (1997) explores the use of the BDSL under moment condition failure.

There is a direct relationship between nonlinear and non-Gaussian time series. In the model (1), even if the disturbance term ε_t is normal, nonlinear transformations of Gaussian noise will make x_t non-Gaussian. Testing for Gaussianity is then an instrumental part of the nonlinear time series toolkit.

Hinich (1982) has developed testing in the time domain using the *bicorrelation*,

$$\gamma(r,s) = {}_{t=1}^{s} x_{t+r}x_{t+s}/(N-s), \quad 0 \leq r \leq s, \tag{28}$$

and in the frequency domain using the *bispectrum*,

$$B(\omega_1,\omega_2) = {}_{r=-\infty}^{\infty} {}_{s=-\infty}^{\infty} \gamma(r,s) \times \exp[-i(\omega_1 r + \omega_2 s)]. \tag{29}$$

For a Gaussian time series, the bicorrelation should be close to zero, and the bispectrum should be flat across all frequencies. Both tests have good power against skewed alternatives.

Ramsey and Rothman (1996) have proposed a related time domain procedure that looks for *time reversibility*,

$$\begin{aligned} &F(X_t, X_{t+1}, \ldots, X_{t+r}) \\ &= F(X_{s-t}, X_{s-t-1}, \ldots, X_{s-t-r}) \end{aligned} \tag{30}$$

for any r, s and t, where $F(.)$ is the joint distribution. This condition is stronger than stationarity because of the triple index. The authors find evidence of business cycle asymmetry using this diagnostic.

Forecasting

For many, the bottom line on nonlinear modelling is the ability to generate superior forecasts. In this respect, the results from the nonlinear literature are decidedly mixed. Harding and Pagan (2002) are prominent sceptics. Teräsvirta et al. (2005) provide a very wide set of evidence in favour of nonlinear models.

Aside from the comparison of point forecasts from model i, $u_{i,t+1} = x_{t+1} - f_i(x_t)$, with a particular loss function $g(.)$,

$$H_0 : \int \left[p_i(x_{t+1}|f_i(x_i)) - p_j\left(x_{t+1}|f_j(x_t)\right)\right]dx = 0. \tag{31}$$

there has been growing interest in comparing forecast densities $p_i(x_{t+1}|f_i(x_t))$,

$$H_0 : \int \left[p_i(x_{t+1}|f_i(x_i)) - p_j\left(x_{t+1}|f_j(x_t)\right)\right]dx = 0. \tag{32}$$

Corradi and Swanson (2005) provide a comprehensive overview of available tools.

See Also

- ▶ Forecasting
- ▶ Linear Models
- ▶ Stochastic Volatility Models

Acknowledgment I would like to thank Cees Diks, James Hamilton, Sebastiano Manzan, Simon Potter, Phil Rothman, Dick van Dijk and Steven Durlauf for helpful comments.

Bibliography

Albert, J., and S. Chib. 1993. Bayesian analysis via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. *Journal of Business and Economic Statistics* 11: 1–15.

Andersen, T.G., T. Bollerslev, F.X. Diebold, and P. Labys. 2003. Modeling and forecasting realized volatility. *Econometrica* 71: 579–625.

Barndorff-Nielsen, O.E., and N. Shephard. 2002. Econometric analysis of realised volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society, Series B* 64: 253–280.

Baek, B., and W.A. Brock. 1992. A nonparametric test for independence of a multivariate time series. *Statistica Sinica* 2: 137–156.

Balke, N., and T. Fomby. 1997. Threshold cointegration. *International Economic Review* 38: 627–645.

Brock, W.A., W.D. Dechert, J.A. Scheinkman, and B. LeBaron. 1996. A test for independence based on the correlation dimension. *Econometric Reviews* 15: 197–235.

Brockett, R.W. 1976. Volterra series and geometric control theory. *Automatica* 12: 167–176.

Corradi, V., and N.R. Swanson. 2005. Predictive density evaluation. In *Handbook of economic forecasting*, ed. C.W.J. Granger, A. Timmermann, and G. Elliott. Amsterdam: North-Holland.

Daubechies, I. 1992. *Ten lectures on wavelets*. 2nd ed. Philadelphia: SIAM.

De Lima, P. 1997. On the robustness of nonlinearity tests due to moment condition failure. *Journal of Econometrics* 76: 251–280.

Diks, C. 2004. The correlation dimension of returns with stochastic volatility. *Quantitative Finance* 4: 45–54.

Filardo, A.J. 1994. Business cycle phases and their transitional dynamics. *Journal of Business and Economic Statistics* 12: 299–308.

Franses, P.H., and D. van Dijk. 2000. *Nonlinear time series models in empirical finance*. New York: Cambridge University Press.

Gallant, A.R., and G. Tauchen. 1987. Seminonparametric maximum likelihood estimation. *Econometrica* 55: 363–390.

Gallant, A.R., and G. Tauchen. 1996. Which moments to match? *Econometric Theory* 12: 657–681.

Gençay, R., and W.D. Dechert. 1992. An algorithm for the n Lyapunov exponents of an n-dimensional unknown dynamical system. *Physica D* 59: 142–157.

Ghysels, E., A. Harvey, and E. Renault. 1996. Stochastic volatility. In *Handbook of statistics 14: Statistical methods in finance*, ed. G.S. Maddala and C.R. Rao. Amsterdam: North-Holland.

Granger, C.W.J., and J. Hallman. 1991. Nonlinear transformations of integrated time series. *Journal of Time Series Analysis* 12: 207–224.

Grassberger, P., and J. Procaccia. 1983. Measuring the strangeness of strange attractors. *Physica D* 9: 189–208.

Hamilton, J.D. 1989. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* 57: 357–384.

Hamilton, J.D. 1990. Analysis of time series subject to changes in regime. *Journal of Econometrics* 45: 39–70.

Hansen, B.E. 1996. Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica* 64: 413–430.

Harding, D., and A. Pagan. 2002. Dissecting the cycle: a methodological investigation. *Journal of Monetary Economics* 49: 365–381.

Hinich, M.J. 1982. Testing for Gaussianity and linearity of a stationary time series. *Journal of Time Series Analysis* 3: 169–176.

Kim, C.J. 1994. Dynamic linear models with Markov-switching. *Journal of Econometrics* 60: 1–22.

Koop, G., H. Pesaran, and S. Potter. 1996. Impulse response analysis in nonlinear multivariate models. *Journal of Econometrics* 74: 119–148.

Krolzig, H.-M. 1997. *Markov switching vector autoregressions: Modelling, statistical inference and application to business cycle analysis*. Berlin: Springer.

Kuan, C.-M., K. Hornik, and H. White. 1994. A convergence result for learning in recurrent neural networks. *Neural Computation* 6: 620–640.

Luukkonen, R., P. Saikkonen, and T. Teräsvirta. 1988. Testing linearity against smooth transition autoregressive models. *Biometrika* 75: 491–499.

Mayfield, S., and B. Mizrach. 1991. Nonparametric estimation of the correlation exponent. *Physical Review A* 88: 5298–5301.

Mittnik, S., and B. Mizrach. 1992. Parametric and seminonparametric analysis of nonlinear time series. In *Advances in GLIM and statistical modeling*, ed. L. Fahrmeir, B. Francis, R. Gilchrist, and G. Tutz. New York: Springer.

Potter, S. 2000. Nonlinear impulse response functions. *Journal of Economic Dynamics and Control* 24: 1425–1446.

Ramsey, J.B., and P. Rothman. 1996. Time irreversibility and business cycle asymmetry. *Journal of Money, Credit, and Banking* 28: 1–21.

Serfling, R.J. 1980. *Approximation theorems of mathematical statistics*. New York: John Wiley.

Shintani, M., and O. Linton. 2003. Is there chaos in the world economy? A nonparametric test using consistent standard errors. *International Economic Review* 44: 331–358.

Shintani, M., and O. Linton. 2004. Nonparametric neural network estimation of Lyapunov exponents and a direct test for chaos. *Journal of Econometrics* 120: 1–33.

Takens, F. 1981. Detecting strange attractors in turbulence. In *Springer lecture notes in mathematics*, ed. D. Rand and L.-S. Young, vol. 898. Berlin: Springer.

Teräsvirta, T. 1994. Specification, estimation, and evaluation of smooth transition autoregressive models. *Journal of the American Statistical Association* 89: 208–218.

Teräsvirta, T., and H.M. Anderson. 1992. Characterizing nonlinearities in business cycles using smooth transition autoregressive models. *Journal of Applied Econometrics* 7: S119–S136.

Teräsvirta, T., D. van Dijk, and M.C. Medeiros. 2005. Linear models, smooth transition autoregressions, and neural networks for forecasting macroeconomic time series: a re-examination. *International Journal of Forecasting* 21: 755–774.

Van Dijk, D., and P.H. Franses. 1999. Modeling multiple regimes in the business cycle. *Macroeconomic Dynamics* 3: 311–340.

Non-nested Hypotheses

M. Hashem Pesaran and M. Rodrigo Dupleich Ulloa

Abstract

This article provides an overview of the literature on hypotheses testing when the hypotheses or models under consideration are non-nested. Two models are said to be non-nested if neither can be obtained from the other by some limiting process, including the imposition of equality and/or inequality constrains on one of the model's parameters. Relevant concepts such as closeness measures and pseudo-true values are discussed and alternative approaches to testing non-nested hypotheses, including the Cox procedure, artificial nesting and the encompassing approach, are reviewed. The Vuong approach to model selection is also covered.

Keywords

Artificial nesting; Cox's test; Encompassing test; Hypothesis testing; Kullback-Leibler information criteria; Linear regression models; Maximum likelihood estimation; Model selection; Non-nested hypotheses; Pseudo-true values

JEL Classifications

C1; C20; C52

In economics, as in many other disciplines, there are competing explanations of the same phenomena, often characterized by alternative statistical models. Different models may represent, for example, different theoretical paradigms, or could be the result of alternative formulations from the same paradigm. Within the classical framework, the problem of model adequacy is approached through 'general specification tests', the 'diagnostic tests', and the 'non-nested tests'. All three approaches can be used to test the same explanation or hypothesis of interest (the null or the maintained hypothesis), but they differ in their consideration of the alternative(s). General specification tests intentionally consider a broad class of alternatives, while the alternatives considered under diagnostic and non-nested testing procedures are much more specific. In the case of non-nested tests the null hypothesis is contrasted with a specific alternative. Non-nested tests are appropriate when rival hypotheses are advanced for the explanation of the same economic phenomenon, and the aim is to devise a powerful test against a specific alternative.

When the null hypothesis is nested within the alternative, standard classical procedures such as those based on the likelihood ratio, Wald and Lagrange multiplier (or score) principles can be utilized. But if the null and the alternative hypotheses belong to 'separate' families of distributions, classical testing procedures cannot be applied directly and need to be suitably modified.

This article provides an overview of the concepts and some of the most widely used non-nested hypotheses tests and applies these procedures to the classical regression models. Our discussion of non-nested hypothesis testing will necessarily omit many topics. Survey articles on this subject include McAleer and Pesaran (1986), Gourieroux and Monfort (1994), and Pesaran and Weeks (2001).

Non-nested Models

Suppose the object of interest is the process generating the random variable Y, observed over a sample of size n, $\mathbf{y} = (y_1, y_2, \ldots, y_n)'$. Assume that the true process generating **y** is characterized by a joint probability density function, $f_0(\mathbf{y})$, which is unknown, and two models (hypotheses) are advanced as possible explanations of *Y*, represented by the joint probability density functions:

$$H_g = \{g(\mathbf{y};\ \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}, H_h \\ = \{h(\mathbf{y}; \gamma), \gamma \in \Gamma\}. \qquad (1)$$

These functions are known but depend on a finite number of unknown parameters denoted by $\theta \in \Theta$ and $\gamma \in \Gamma$, respectively. The sets Θ and Γ represent the 'admissible' parameter space for which the respective densities g(**y**; $\boldsymbol{\theta}$) and h(**y**; γ) are well defined. The aim is to ascertain which of the two alternatives, H_g and H_h, if any, can be viewed as belonging to $f0(\mathbf{y})$. In this set-up there is no natural null hypothesis; either of the two hypotheses under consideration can be taken as the null. In practice, the analysis of non-nested hypotheses is carried out with both alternatives taken in turn as the null hypothesis. Four outcomes are possible: (i) H_g rejected against H_h and not vice versa, (ii) H_h rejected against H_g and not vice versa, (iii) neither hypothesis is rejected against the other, and finally (iv) both hypotheses are rejected against one another. The first two outcomes are familiar from the classical test results and are straightforward to interpret. The third outcome can arise when the two models are very close to $f_0(\mathbf{y})$, and hence equivalent observationally. The fourth outcome suggests the existence of a third possible model which shares important features from both models under consideration.

Pseudo-True Values and Closeness Measures

Given the observations **y**, the maximum likelihood (ML) estimators of $\boldsymbol{\theta}$ and γ are given by

$$\widehat{\boldsymbol{\theta}}_n = \underset{\theta \in \Theta}{\arg\max}\, L_g(\boldsymbol{\theta}), \widehat{\gamma}_n = \underset{\gamma \in \Gamma}{\arg\max}\, L_h(\gamma),$$

where the corresponding log-likelihood functions are defined by $L_g(\boldsymbol{\theta}) =$ log (g(**y**; $\boldsymbol{\theta}$)) and $L_h(\gamma) =$ log (h(**y**; γ)). Throughout we shall assume that probability densities satisfy the usual regularity conditions as established, for example in White (1982), such that $\widehat{\boldsymbol{\theta}}_n$ and $\widehat{\gamma}_n$ have asymptotically normal limiting distributions under the 'true' model, $f_0(\mathbf{y})$. In the general case where neither of the models under consideration coincide with f_0(y), $\widehat{\boldsymbol{\theta}}_n$ and $\widehat{\gamma}_n$ are known as quasi-ML estimators and their probability limits under f_0(y) are referred to as (asymptotic) pseudo-true values, such that

$$\boldsymbol{\theta}_{*f} = \underset{\theta \in \Theta}{\arg\max}\, \mathbb{E}_f\{L_g(\boldsymbol{\theta})\}\gamma_{*f} \\ = \underset{\gamma \in \Gamma}{\arg\max}\, \mathbb{E}_f\{L_h(\gamma)\} \qquad (2)$$

where $\mathbb{E}_f(\cdot)$ denotes expectations under the true density $f_0(\mathbf{y})$. In what follows, we assume that the above asymptotic pseudo-true values exist; and $\boldsymbol{\theta}_{*f}$ and γ_{*f} are the *unique* maxima to the respective optimization problem given in (2), such that *global identifiability* is ensured. For the case in which $f_0(\mathbf{y})$ belongs to H_g, we have that $\boldsymbol{\theta}_{*g} = \boldsymbol{\theta}_0$ and $\gamma_{*g} = \gamma^*(\boldsymbol{\theta}_0)$, where the 'true' value of $\boldsymbol{\theta}$ under H_g, is denoted by $\boldsymbol{\theta}_0$. Given the symmetry of our setting, under H_h we have $\boldsymbol{\theta}_{*h} = \boldsymbol{\theta}^*(\gamma_0)$ and $\gamma_{*h} = \gamma_0$, where γ_0 is the 'true' value of γ under H_h. The relationship between the parameters of the two models under consideration is given by the functions $\gamma_{*g} = \gamma^*(\boldsymbol{\theta}_0)$ and $\boldsymbol{\theta}_{*h} = \boldsymbol{\theta}_0(\gamma_0)$, known as the *binding functions.*

Using closeness measures and pseudo-true values, Pesaran (1987) provides a formalization of the concepts of nested and non-nested hypotheses. The closeness of H_g with respect to H_h is given by

$$C_{gh}(\boldsymbol{\theta}_0) = \boldsymbol{I}_{gh}(\boldsymbol{\theta}_0, \gamma^*(\boldsymbol{\theta}_0)) = \min_{\gamma} \boldsymbol{I}_{gh}(\boldsymbol{\theta}_0, \gamma) \quad (3)$$

$$= \mathbb{E}_g\{L_g(\boldsymbol{\theta}_0) - L_h(\gamma^*(\boldsymbol{\theta}_0))\} \qquad (4)$$

where $\boldsymbol{I}_{gh}(\boldsymbol{\theta}_0,\ \gamma_*)$ is known as the *Kullback-Leibler information criterion* (KLIC) measure, introduced by Kullback (1959). Similarly, the closeness of H_h to H_g, is defined by $C_{hg}(\gamma_0) = I_{hg}(\gamma_0,\ \boldsymbol{\theta}^*(\gamma_0))$.

- H_g is *nested* within H_h if and only if $C_{gh}(\boldsymbol{\theta}_0) = 0$, for all values of $\boldsymbol{\theta}_0 \in \Theta$, and $C_{hg}(\gamma_0) \neq 0$ for some $\gamma_0 \in \Gamma$.
- H_g and H_h are *globally non-nested* if and only if $C_{gh}(\boldsymbol{\theta}_0)$ and $C_{hg}(\gamma_0)$ are both non-zero for all values of $\boldsymbol{\theta}_0 \in \Theta$ and $\gamma_0 \in \Gamma$.
- H_g and H_h are *partially non-nested* if $C_{gh}(\boldsymbol{\theta}_0)$ and $C_{hg}(\gamma_0)$ are both non-zero for some values of $\boldsymbol{\theta}_0 \in \Theta$ and $\gamma_0 \in \Gamma$.
- H_g and H_h are *observationally equivalent* if and only if $C_{gh}(\boldsymbol{\theta}_0) = 0$ and $C_{hg}(\boldsymbol{\gamma}_0) = 0$ for all values of $\boldsymbol{\theta}_0 \in \Theta$ and $\gamma_0 \in \Gamma$.

Tests of Non-nested Hypotheses

There are three general approaches to testing non-nested hypotheses. The first, due to the pioneering contributions of Cox (1961, 1962), involves centring the log-likelihood ratio statistic under the null hypothesis and then deriving its asymptotic null distribution. This is known as the *Cox test.* A second approach, also suggested by Cox (1962) and explored extensively by Atkinson (1970), is based on an artificially constructed general model. The basic idea is to introduce a third hypothesis in which both H_g and H_h are nested as special cases. A third approach, originally considered by Deaton (1982) and Dastoor (1983), and further developed by Mizon and Richard (1986) known as the *encompassing procedure*, focuses on the ability of one model in explaining particular features of an alternative model. In a related contribution, Gourieroux et al. (1983) extend the Wald and score-type tests to non-nested models. Their statistics are based on the difference between two estimators of the pseudo-true values.

The Cox test statistic is derived by modifying the log-likelihood ratio statistic, $L_g\left(\widehat{\boldsymbol{\theta}}_n\right) - L_h(\widehat{\gamma}_n)$, so that it is appropriately centred. Specifically, for testing H_g against H_h, the numerator of the Cox statistic is given by

$$S_n^{gh} = \left\{L_g\left(\widehat{\boldsymbol{\theta}}_n\right) - L_h(\widehat{\gamma}_n)\right\} - \widehat{\mathbb{E}}_g\left\{L_g\left(\widehat{\boldsymbol{\theta}}_n\right) - L_h(\widehat{\gamma}_n)\right\}, \tag{5}$$

where $\widehat{\mathbb{E}}_g\left\{L_g\left(\widehat{\boldsymbol{\theta}}_n\right) - L_h(\widehat{\gamma}_n)\right\}$, is a consistent estimator of $C_{gh}(\boldsymbol{\theta}_0)$. In the case where H_g is nested within H_h we have $C_{gh}(\boldsymbol{\theta}_0) = 0$ for all $\boldsymbol{\theta}_0$, and S_n^{gh} reduces to the standard log-likelihood ratio statistic. An application to linear regression models has been proposed by Pesaran (1974) and subsequently extended to simultaneous non-linear equations systems by Pesaran and Deaton (1978). As pointed out previously, since there is no natural null hypothesis in this testing framework, one also needs to consider the modified log-likelihood statistic for testing H_h against H_g which is denoted by S_n^{hg}. Under a suitable normalization (that is $\sqrt{n}$), both statistics are asymptotically normally distributed under their respective nulls with a zero mean and a finite asymptotic variance. When the null hypothesis of H_g is considered against H_h, we have

$$N_n^{gh} = \frac{\sqrt{n} S_n^{gh}}{\sqrt{V_{gh}}} \underset{\sim}{a} N(0,1)$$

where V_{gh} is the asymptotic variance of $\sqrt{n} S_n^{gh}$ and $\underset{\sim}{a}$ denotes asymptotical equivalence in distribution (for details see Pesaran and Deaton 1978). Based on the results of the two statistics, N_n^{gh} and N_n^{hg}, four outcomes are possible:

- reject H_g but not H_h if $\left|N_n^{gh}\right| < c_\alpha$ and $\left|N_n^{hg}\right| \geq c_\alpha$,
- reject H_h but not H_g if $\left|N_n^{gh}\right| \geq c_\alpha$ and $\left|N_n^{hg}\right| < c_\alpha$,
- reject both H_g and H_h if $\left|N_n^{gh}\right| \geq c_\alpha$ and $\left|N_n^{hg}\right| \geq c_\alpha$,
- reject neither H_g and H_h if $\left|N_n^{gh}\right| < c_\alpha$ and $\left|N_n^{hg}\right| < c_\alpha$,

where the $(1 - \alpha)$ per cent critical value of the standard normal distribution is denoted by c_α. In the case of non-nested hypotheses, there is no way of ranking the models by the level of their generality. As a consequence, the test results may provide a consistent outcome such as the rejection of H_g (or H_h) by both tests. But it is also not unusual, given the data, for both non-nested models to be simultaneously rejected or fail to be rejected. For the case of a simultaneous rejection of H_g and H_h, we need to find some other model that fits the data better. If neither model is rejected, this may indicate lack of power.

The second approach, named Atkinson's comprehensive method (Atkinson 1970), is based on an artificial nesting of the two models with a general model such as,

$$f_c(\mathbf{y};\, \boldsymbol{\theta}, \gamma, \lambda) = \left\{ \frac{(g(\mathbf{y};\, \boldsymbol{\theta}))^{\lambda} (h(\mathbf{y}; \gamma))^{1-\lambda}}{\int (g(\mathbf{y};\, \boldsymbol{\theta}))^{\lambda} (h(\mathbf{y}; \gamma))^{1-\lambda} d\mathbf{y}}, \lambda \in [0,1], \boldsymbol{\theta} \in \Theta, \gamma \in \Gamma \right\} \tag{6}$$

Atkinson's comprehensive approach considers families that are obtained by mixing the probability distributions of H_g and H_h. It requires the existence of the integral appearing in the denominator in eq. (6). This component ensures that the combined function $f_c(\mathbf{y};\, \boldsymbol{\theta}, \gamma, \lambda)$, is in fact a proper density function. In equation (6), the comprehensive model is based on an exponential combination (that is, a geometric mean); alternatively the compound model can also be derived from an arithmetic mean (see for instance Quandt 1974). In this set-up, the hypothesis H_g is obtained by imposing $\lambda = 1$, while the hypothesis H_h is obtained by imposing $\lambda = 0$. Thus, in principle, by testing $\lambda = 1$ or $\lambda = 0$, we can test H_g or H_h, respectively. The 'mixing' parameter, λ, varies over the range [0, 1] and measures the relative weights attached to H_g and H_h. As a consequence, tests for the restriction of $\lambda = 1$ ($\lambda = 0$) against the alternative that $\lambda \neq 1$ ($\lambda \neq 0$) can be performed based on standard techniques from the literature of nested hypothesis testing (see Atkinson 1970; Pesaran 1982b).

Atkinson's approach is, however, subject to a number of drawbacks. The first one arises from the fact that under $\lambda = 1$ (or $\lambda = 0$), the unknown parameter vector $\boldsymbol{\gamma}$ (or $\boldsymbol{\theta}$) disappears from the combined model written in (6). This is known as Davies' problem (Davies 1977) which can be circumvented in various ways, as discussed, for example, by Pesaran (1982c). The second limitation is due to the fact that testing $\lambda = 1$ against $\lambda \neq 1$ is not equivalent to performing the test of H_g against H_h, which is the primary object of the non-nested testing exercise. Finally, there is some degree of arbitrariness in the choice of the comprehensive model (see Pesaran 1981).

The *encompassing approach* generalizes Cox's original idea and examines the extent to which H_g explains one or more features of the rival model, H_h. When all the features of the model H_h can be explained by model H_g, then H_g is said to *encompass* H_h. This condition is denoted by

$$H_g \varepsilon H_h \;:\; \gamma_{*f} = \gamma^*\left(\theta_{*f}\right). \tag{7}$$

Likewise, $H_h \varepsilon H_g$ implies that all features of model H_g can be explained by the model H_h, that is H_h *encompasses* H_g, such that

$$H_g \varepsilon H_h \;:\; \theta_{*f} = \theta^*\left(\gamma_{*f}\right). \tag{8}$$

Recall that λ_{*f} and $\boldsymbol{\theta}_{*f}$ are the pseudo-true values of γ and $\boldsymbol{\theta}$, with respect to the true model H_f. Moreover, $\gamma_*(\cdot)$ and $\boldsymbol{\theta}_*(\cdot)$ are the binding functions linking the parameters of the models under H_g and H_h. The encompassing hypothesis $H_g \varepsilon H_h$ (resp. $H_g \varepsilon H_h$) can be tested using the statistic $\sqrt{n}$ $\left(\widehat{\gamma}_n - \gamma^*\left(\widehat{\boldsymbol{\theta}}_n\right)\right)$, respectively $\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}^*\left(\widehat{\gamma}_n\right)\right)$. Gourieroux and Monfort (1995) show that under the encompassing hypothesis, $H_g \varepsilon H_h$ and a set of regularity conditions, $\sqrt{n}\left(\widehat{\gamma}_n - \gamma^*\left(\widehat{\boldsymbol{\theta}}_n\right)\right)$ is asymptotically normal with zero mean and a finite covariance matrix. Based on this result, two testing procedures are proposed by Gourieroux and Monfort (1995): the Wald encompassing test (WET) and the score encompassing test (SET). In practice, the implementation of these tests tends to be difficult. First, the binding functions $\gamma_*(\cdot)$ and $\boldsymbol{\theta}_*(\cdot)$ are not easy to derive and, second, the variance–covariance matrices appearing in the test statistics tend to be difficult to compute in practice. Chen and Kuan (2002) suggest the use of 'pseudo-true score' as a way of avoiding the need to estimate pseudo-true values.

Vuong's Model Selection Test

Vuong's (1989) criterion is motivated by testing that H_g and H_h are observationally equivalent, using the *Kullback-Leibler information criterion* (KLIC) as a closeness measure. The focus of this approach is to test the hypothesis that the models under consideration are 'equally' close to the true unknown model, $H_f : f_0(\mathbf{y})$. It provides a natural link between model selection and hypothesis testing approaches. Under model selection, a model is selected even if the 'best' model happens to be very close to the second best model. Vuong's approach allows the statistical significance of the differences between models to be tested using classical testing procedures. It is based on the closeness measures of H_g and H_h with respect to the true model, H_f, namely (for closeness of H_g to H_f)

$$C_{fg}\left(\boldsymbol{\theta}_{*f}\right) = E_f\left\{L_f(\cdot) - L_g\left(\boldsymbol{\theta}_{*f}\right)\right\}$$

and (for closeness of H_h to H_f)

$$C_{fg}\left(\gamma_{*f}\right) = E_f\left\{L_f(\cdot) - L_h\left(\gamma_{*f}\right)\right\}.$$

The null hypothesis of interest, 'H_g and H_h are equivalent', is then formally defined by

$$H_V : C_{fg}\left(\boldsymbol{\theta}_{*f}\right) = C_{fh}\left(\gamma_{*f}\right) \tag{9}$$

which is equivalent to the unknown quantity $H_V : E_f\left\{L_g\left(\boldsymbol{\theta}_{*f}\right) - L_h\left(\gamma_{*f}\right)\right\} = 0$, that depends on $f_0(\mathbf{y})$, the unknown true distribution. However, the latter difference can be consistently estimated by $T^{-1}\left\{L_g\left(\widehat{\boldsymbol{\theta}}_n\right) - L_h(\widehat{\gamma}_n)\right\}$, an average of the log-likelihood ratio statistic. Vuong derives an asymptotic standard normal distribution for the related test statistic under H_V.

Rivers and Vuong (2002) provide a number of generalizations and show that the test can be applied to nonlinear dynamic models and other closeness measures.

Application to Linear Regression Models

An important application of the non-nested tests in econometrics has been to linear regression models. Consider the following classical normal regression models:

$$\begin{aligned} H_g &: \mathbf{y} = \mathbf{X}\beta + \mathbf{u}_g, \mathbf{u}_g \sim N(0, \sigma^2 \boldsymbol{I}_n), 0 < \sigma^2 < \infty, \\ H_h &: \mathbf{y} = \mathbf{Z}\alpha + \mathbf{u}_h, \mathbf{u}_h \sim N(0, \omega^2 \boldsymbol{I}_n), 0 < \omega^2 < \infty, \end{aligned} \tag{10}$$

where $\mathbf{X}$ and $\mathbf{Z}$ are $n \times k_g$ and $n \times k_h$ matrices of observations on the explanatory variables of models H_g and H_h, respectively. These variables are assumed to be distributed independently of the $n \times 1$ disturbance vectors $\mathbf{u}_g$ and $\mathbf{u}_h$. The parameters $\theta = (\beta', \sigma^2)'$ and $\gamma = (\boldsymbol{\alpha}', \omega^2)'$ are the $(k_g + 1) \times 1$ and $(k_h + 1) \times 1$ vectors of unknown regression coefficients, and $\mathbf{I}_n$ is the identity matrix of order n. It is also assumed that the probability limits of $\hat{\Sigma}_{xx} = n^{-1}(\mathbf{X}'\mathbf{X}), \hat{\Sigma}_{zz} = n^{-1}(\mathbf{Z}'\mathbf{Z})$ and $\hat{\Sigma}_{xz} = n^{-1}(\mathbf{X}'\mathbf{Z})$ exist with population values denoted by the non-singular matrices Σ_{xx}, Σ_{zz} and Σ_{xz}. At the same time, define $\hat{\Sigma}_g = \hat{\Sigma}_{xx} - \hat{\Sigma}_{xz}\hat{\Sigma}_{zz}^{-1}\hat{\Sigma}_{zx} > 0$ and $\hat{\Sigma}_h = \hat{\Sigma}_{zz} - \hat{\Sigma}_{zx}\hat{\Sigma}_{xx}^{-1}\hat{\Sigma}_{xz} > 0$. The link between these strict inequality restrictions and the nesting properties of the models in (10) will be made clear below.

Suppose that neither H_g nor H_h belong to the true DGP, and the data is generated by

$$H_f : \mathbf{y} = \mathbf{W}\delta + \mathbf{u}_f, \mathbf{u}_f \sim N\left(0, v^2\mathbf{I}_n\right), 0 < v^2 < \infty. \tag{11}$$

As before, assume that $\hat{\Sigma}_{ww} = n^{-1}(\mathbf{W}'\mathbf{W}), \hat{\Sigma}_{wx} = n^{-1}(\mathbf{W}'\mathbf{X})$ and $\hat{\Sigma}_{wz} = n^{-1}(\mathbf{W}'\mathbf{Z})$ can be replaced by their population values given by Σ_{ww}, Σ_{wx} and Σ_{wz}. The pseudotrue values given in (2) can be obtained for this case by maximizing $\mathrm{E}_f\{n^{-1}L_g(\theta)\}$ with respect to $\boldsymbol{\theta}$ which yields

$$\theta_{*f} = \begin{pmatrix} \beta_{*f} \\ \sigma^2_{*f} \end{pmatrix} = \begin{pmatrix} \Sigma_{xx}^{-1}\Sigma_{xw}\delta \\ v^2 + \delta'\left(\Sigma_{ww} - \Sigma_{wx}\Sigma_{xx}^{-1}\Sigma_{xw}\right)\delta \end{pmatrix}. \tag{12}$$

Similarly, for model *Hh* we have

$$\gamma_{*f} = \begin{pmatrix} \alpha_{*f} \\ \omega^2_{*f} \end{pmatrix} = \begin{pmatrix} \Sigma_{zz}^{-1}\Sigma_{zw}\delta \\ v^2 + \delta'\left(\Sigma_{ww} - \Sigma_{wz}\Sigma_{zz}^{-1}\Sigma_{zw}\right)\delta \end{pmatrix}. \tag{13}$$

Note that, for the case in which *Hf* belongs to the family of models given by *Hg*, the latter result can be rewritten as

$$\gamma_{*g} = \begin{pmatrix}\alpha_{*g}\\ \omega^2_{*g}\end{pmatrix} = \begin{pmatrix}\Sigma_{zz}^{-1}\Sigma_{zw}\beta \\ \sigma^2+\beta'\Sigma_g\beta\end{pmatrix}. \tag{14}$$

In terms of our previous discussion, these regression models are non-nested if it is not possible to write **X** as an exact linear function of **Z** and *vice versa*, or more formally if $\mathbf{X}\not\subseteq\mathbf{Z}$ and $\mathbf{Z}\not\subseteq\mathbf{X}$. The model H_g is said to be nested in H_h if $\mathbf{X}\subset\mathbf{Z}$ and $\mathbf{Z}\subset\mathbf{X}$. The two models are observationally equivalent if **X C Z** and **Z C X**. These conditions can be written in terms of the KLIC measure given by (3) as in McAleer and Pesaran (1986), who derive the closeness measure of H_h with respect to H_g as

$$C_{gh}(\theta) = \frac{1}{2}\log\left[1+\frac{\beta'\sum_g\beta}{\sigma^2}\right].$$

Similarly, the KLIC measure of closeness of H_g with respect to H_h is

$$C_{hg}(\gamma) = \frac{1}{2}\log\left[1+\frac{\alpha'\sum_h\alpha}{\omega^2}\right].$$

It can be easily seen from this example that a necessary and sufficient condition for H_g to be nested within H_h is $\beta\ '\ \sum_g\beta/\sigma^2=0$ for all admissible values of β with $\alpha'\sum_h\alpha/\omega^2\neq 0$ for some α. Note that $\beta'\sum_g\beta/\sigma^2=0$ is implied by either $\mathbf{\Sigma}_g\beta=0$ or$\mathbf{\Sigma}_g=0$.

Given the linear set-up and using results in Pesaran (1974), the adjusted log-likelihood ratio statistic for testing H_g against H_h is given by

$$S_n^{gh} = \frac{n}{2}\log\left(\frac{\hat\omega^2_{*ng}}{\hat\omega^2_n}\right) \tag{15}$$

where $\hat\omega^2_n$ is the estimate of ω^2 under the alternative H_h, and $\hat\omega^2_{*ng}$ is the estimated pseudo-true value of the residual variance of H_h under $\mathrm{H_g}$, such that

$$\begin{aligned}\hat\omega^2_n &= n^{-1}(\mathbf{y}-\mathbf{X}\hat\alpha_n)'(\mathbf{y}-\mathbf{X}\hat\alpha_n)\hat\omega^2_{*ng}\\ &= \hat\sigma^2_n+\hat\beta'_n\hat\Sigma_g\hat\beta'_n\end{aligned}$$

in which the estimates under the true model $\mathrm{H_g}$ are given by$\hat\sigma^2_n=n^{-1}\left(\mathbf{y}-\mathbf{X}\hat\beta_n\right)'\left(\mathbf{y}-\mathbf{X}\hat\beta'_n\right)$, $\hat\beta'_n=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ and $\hat\alpha'_n=(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$. As pointed out earlier, since we do not have a natural null hypothesis in this framework, one also needs to evaluate H_g against H_h, for which the modified log-likelihood statistic is given by

$$S_n^{hg} = \frac{n}{2}\log\left(\frac{\hat\sigma^2_{*nh}}{\hat\sigma^2_n}\right). \tag{16}$$

For the statistic given by (15), the asymptotic variance of $\sqrt{n}S_n^{gh}$, denoted by V_{gh}, can be computed as follows:

$$V_{gh} = \frac{\hat\sigma^2_n\left(\hat\beta'_n\mathbf{X}'\mathbf{M}z\mathbf{M}x\mathbf{M}z\mathbf{X}\hat\beta'_n\right)}{n\left(\hat\sigma^2_n+\hat\beta'_n\hat\Sigma_g\hat\beta'_n\right)} \tag{17}$$

where $\mathbf{M}_x=\mathbf{I}_n-\mathbf{X}'(\mathbf{X}'\mathbf{X})^{-1}\mathrm{X}'$ and $\mathbf{M}_z=\mathbf{I}_n-\mathbf{Z}'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ are orthogonal projection matrices. Combining (15) and (17), the associated standardized Cox statistic, $N_n^{gh}=\sqrt{n}S_n^{gh}/\sqrt{V_{gh}}$, can now be calculated as described in Pesaran (1974). Similar derivations lead to the analogue statistic for the test of H_h against H_g, N_n^{hg}.

The application of the comprehensive approach to the above linear regression models yields the following exponential combination as presented in (6):

$$H_\lambda:\mathbf{y}=(1-\xi)\mathbf{X}\beta+\xi\mathbf{Z}\alpha+\mathbf{u},\mathbf{u}\sim N\left(0,\sigma^2_\lambda\mathbf{I}_n\right) \tag{18}$$

where $\xi=\lambda\sigma^2_\lambda/\sigma^2$ and $\sigma_\lambda^{-2}=(1-\xi)\sigma^{-2}+\xi\omega^{-2}$. Given that the error variances σ^2 and ω^2 are strictly positive, performing a test of $\xi=0$ is equivalent to testing $\lambda=0$ when we consider the null hypothesis of H_g against H_h. As pointed out earlier, the Davies problem arises when under the null hypothesis of H_g ($\lambda=0$), the unknown parameter vector $\boldsymbol{\alpha}$ disappear from the mixture model. The presence of this nuisance parameter results in a Student-type of test statistic associated with λ that depends on the value of $\boldsymbol{\alpha}$, such that

$$t_\lambda(a) = \frac{a'\mathbf{Z}'\mathbf{M}_x\mathbf{y}}{\hat{\sigma}_\lambda^2(a'\mathbf{Z}'\mathbf{M}_x\mathbf{Z}a)^{1/2}} \tag{19}$$

where $\hat{\sigma}_\lambda^2$ denotes the usual estimator of the variance of the errors. One possible way to solve this identification problem would be to construct a test statistic based on $\mathrm{F}_\lambda = \max_\alpha t_\lambda(\alpha)$.

A different approach to deal with the identification problem was proposed by Davidson and MacKinnon (1981), who propose a J-test by replacing the nuisance parameter $\boldsymbol{\alpha}$ by its estimator, $\hat{\alpha}_n$, under H_h. An exact version of this test, proposed by Fisher and McAleer (1981) and known as the JA-test (indicating the Atkinson variation of this test), substitutes a by the estimate of its pseudo-true value under H_h given in (14), that is $\hat{\boldsymbol{\alpha}}^*\left(\hat{\beta}_n\right) = \hat{\Sigma}_{zz}^{-1}\hat{\Sigma}_{zx}\hat{\beta}_n$. By symmetry of our testing problem, the J and JA versions of the t-test can also be calculated for H_h against H_g. Davidson and MacKinnon (1981) show that the t-ratio statistic, $t_\lambda(\hat{\alpha}_n)$, has asymptotically a standard normal distribution under the null.

Based on the application of Roy's union-intersection principle, McAleer and Pesaran (1986) show that the test for $\xi = 0$ in (18) is equivalent to the standard F-statistic for the test of $\boldsymbol{\delta}_2 = 0$ in the combined model $\mathbf{y} = \mathbf{X}\delta_1 + \mathbf{Z}\delta_2 + \mathbf{u}$.

In order to frame the linear regression models into the encompassing type tests, we can focus on the discrepancy between the OLS estimator of the regression coefficients, denoted by $\hat{\alpha}_n$, and the estimator of the pseudo-true value in finite samples, such that $\sqrt{n}\left(\hat{\alpha}_n - \hat{\alpha}^*\left(\hat{\beta}_n\right)\right) = \sqrt{n}(\mathbf{Z}'\mathbf{Z})^{-1}$ $\mathbf{Z}'\mathbf{M}_x\mathbf{y}$. Using this, we can build an encompassing statistic for testing $H_g\varepsilon H_h$, as follows:

$$\sqrt{n}\left(\hat{\alpha}_n - \hat{\alpha}^*\left(\hat{\beta}_n\right)\right) = \sqrt{n}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{M}_x\mathbf{W}\delta + \sqrt{n}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{M}_x\mathbf{u}_f,$$

if H_f is taken as the true model given in (11). As a consequence, under some regularity conditions, $\sqrt{n}\left(\hat{\alpha}_n - \hat{\alpha}^*\left(\hat{\beta}_n\right)\right)$ is asymptotically normally distributed with mean zero and the covariance matrix $v^2\Sigma_{xx}^{-1}\Sigma_g\Sigma_{xx}^{-1}$. Using these results the WET statistic for testing $H_g\varepsilon H_h$, is given by

$$\varepsilon_{gh} = \frac{\mathbf{y}'\mathbf{M}_x\mathbf{Z}_1\left(\mathbf{Z}_1'\mathbf{M}_x\mathbf{Z}_1\right)^{-1}\mathbf{Z}_1'\mathbf{M}_x\mathbf{y}}{\hat{V}^2} \tag{20}$$

where $\mathbf{Z}_1$ are the components in $\mathbf{Z}$ that are orthogonal to $\mathbf{X}$. Similarly, a *variance* encompassing test of $Hg\varepsilon H_h$ can be constructed for the discrepancy between a consistent estimate of ω^2 and its pseudo-true value ω_{*f}^2, which takes the form of $\hat{\omega}_n^2 - \hat{\omega}_*^2\left(\hat{\theta}_n\right)$. For the case in which H_g contains the true model, H_h, the variance encompassing test is asymptotically equivalent to the Cox and the J-tests.

Vuong's test criterion for the comparison of H_g and H_h is computed as

$$G_{gh} = \frac{\sum_{i=1}^{n} d_i}{\left[\sum_{i=1}^{n}\left(d_i - \overline{d}\right)^2\right]^{1/2}} \tag{21}$$

where $\overline{d} = n^{-1} \sum_{i=1}^{n} d_i$; $d_i = -1/2\left(\hat{\sigma}_n^2/\hat{\omega}_n^2\right)$ $-1/2\left(\left(\hat{u}_{ig}^2/\hat{\sigma}_n^2\right) - \left(\hat{u}_{ig}^2/\hat{\omega}_n^2\right)\right)$; and $\hat{u}_{ig}$ and $\hat{u}_{ih}$ are the estimated residuals of the underlying linear models given by (10). Under the null hypothesis H_V, H_g and H_h are equivalent and G_{gh} is approximately distributed as a standard normal variate.

Extensions and Empirical Applications

Non-nested tests have also been derived for a number of other models, including tests of non-nested linear regression models with serially correlated errors (McAleer et al. 1990), regression models estimated by instrumental variables (Ericsson 1983), models estimated by generalized method of moments (Smith 1992), generalized empirical likelihood (Ramalho and Smith 2002), conditional empirical likelihood (Otsu and Whang 2005), non-nested Euler equations (Ghysels and Hall 1990), autoregressive versus moving average models (Walker 1967), autoregressive conditional heteroskedastic models (Bera and Higgins 1997; McAleer and Ling 1998), logit and probit models (Pesaran and Pesaran 1993), non-nested threshold autoregressive models (Altissimo and Violante 2001; Pesaran and Potter 1997), and stochastic volatility models (Kim et al. 1998).

Further theoretical contributions include a robust version of Cox-type statistics that controls for the effect of contamination in the data (Victoria-Feser 1997), conditional tests on sufficient statistics (Pace and Salvan 1990), asymptotic improvements to Davidson and MacKinnon's approach (Royston and Thompson 1995), score-type statistics which are constructed from linear combinations of the likelihood functions (Santos Silva 2001) and the enhancement of finite-sample performance of non-nested tests by bootstrap methods (Godfrey 1998; Davidson and MacKinnon 2002).

Various economic applications of non-nested hypothesis testing have appeared in the literature. Among them, savings and consumption functions (Deaton 1982), Keynesian and new classical models of unemployment (Pesaran 1982a), wage-employment bargaining models (Vannetelbosch 1996), effects of dividend taxes on corporate investment decisions (Poterba and Summers 1983), money demand functions (McAleer et al. 1982; Elyasiani and Nasseh 1994), autoregressive and moving-average schemes for unanticipated inflation series (Pagan et al. 1983), exchange rates models (Backus 1984), alternative crop response models (Ackello-Ogutu et al. 1985; Frank et al. 1990), agricultural marketing margins (Lyon and Thompson 1993), economic growth models (Ram 1986; Dowrick and Gemmell 1991; Bleaney and Nishiyama 2002), and hedonic house prices (Dubin and Sung 1990; Goodman and Dubin 1991). In the literature of empirical industrial organization, non-nested hypothesis testing is applied to compare a Nash and collusive pricing in an industry with vertical product differentiation (Bresnahan 1987). Non-nested tests are also applied in game-theoretic contexts by Gasmi et al. (1992) and Sandler and Murdoch (1990), in sociological research by Halaby and Weakliem (1993), and in political science by Clarke (2001).

Non-nested tests for rival linear regression models can be computed using various econometric packages. See, for example, Pesaran and Pesaran (1997).

See Also

- ▶ Econometrics
- ▶ Encompassing
- ▶ Linear Models
- ▶ Maximum Likelihood
- ▶ Model Selection
- ▶ Specification Problems in Econometrics
- ▶ Statistical Inference

Bibliography

Ackello-Ogutu, C., Q. Paris, and W.A. Williams. 1985. Testing a von Liebig crop response function against polynomial specifications. *American Journal of Agricultural Economics* 67: 873–880.

Altissimo, F., and G.L. Violante. 2001. The nonlinear dynamics of output and unemployment in the US. *Journal of Applied Econometrics* 16: 461–486.

Atkinson, A. 1970. A method for discriminating between models. *Journal of the Royal Statistical Society B* 32: 323–353.

Backus, D. 1984. Empirical models of the exchange rate: Separating the wheat from the chaff. *Canadian Journal of Economics* 17: 824–846.

Bera, A.K., and M.L. Higgins. 1997. ARCH and bilinearity as competing models for nonlinear dependence. *Journal of Business and Economic Statistics* 15: 43–50.

Bleaney, M., and A. Nishiyama. 2002. Explaining growth: A contest between models. *Journal of Economic Growth* 7: 43–56.

Bresnahan, T.F. 1987. Competition and collusion in the American automobile market: The 1955 price war. *Journal of Industrial Economics* 35: 457–482.

Chen, Y.T., and C.M. Kuan. 2002. The pseudo-true score encompassing test for nonnested hypotheses. *Journal of Econometrics* 106: 271–295.

Clarke, K.A. 2001. Testing nonnested models of international relations: Reevaluating realism. *American Journal of Political Science* 45: 724–744.

Cox, D.R. 1961. Tests of separate families of hypothesis. *Proceedings of the Forth Berkeley Symposium on Mathematical Statistics and Probability* 1: 105–123.

Cox, D.R. 1962. Further results on tests of separate families of hypothesis. *Journal of the Royal Statistical Society B* 24: 406–424.

Dastoor, N.K. 1983. Some aspects of testing nonnested hypothesis. *Journal of Econometrics* 21: 213–228.

Davidson, R., and J.G. MacKinnon. 1981. Several tests for model specification in the presence of alternative hypothesis. *Econometrica* 49: 781–793.

Davidson, R., and J.G. MacKinnon. 2002. Bootstrap J tests of nonnested linear regression models. *Journal of Econometrics* 109: 167–193.

Davies, R.B. 1977. Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* 64: 247–254.

Deaton, A.S. 1982. Model selection procedures, or, does the consumption function exist? In *Evaluating the reliability of macro-economic models*, ed. G.C. Chow and P. Corsi. New York: Wiley.

Dowrick, S., and N. Gemmell. 1991. Industrialisation, catching up and economic growth: A comparative study across the world's capitalist economies. *Economic Journal* 101: 263–275.

Dubin, R.A., and C.-H. Sung. 1990. Specification of hedonic regressions: Non-nested tests on measures of neighborhood quality. *Journal of Urban Economics* 27: 97–110.

Elyasiani, E., and A. Nasseh. 1994. The appropriate scale variable in the U.S. money demand: An application of nonnested tests of consumption versus income measures. *Journal of Business and Economic Statistics* 12: 47–55.

Ericsson, N. 1983. Asymptotic properties of instrumental variables statistics for testing nonnested hypothesis. *Review of Economic Studies* 50: 287–304.

Fisher, G.R., and M. McAleer. 1981. Alternative procedures and associated tests of significance for nonnested hypothesis. *Journal of Econometrics* 16: 103–119.

Frank, M.D., B.R. Beattie, and M.E. Embleton. 1990. A comparison of alternative crop response models. *American Journal of Agricultural Economics* 72: 597–603.

Gasmi, F., J.J. Laffont, and Q. Vuong. 1992. Econometric analysis of collusive behavior in a soft-drink market. *Journal of Economics and Management Strategy* 1: 277–311.

Ghysels, E., and A. Hall. 1990. Testing non-nested Euler conditions with quadrature based method approximation. *Journal of Econometrics* 46: 273–308.

Godfrey, L.G. 1998. Tests of non-nested regression models: Some results on small sample behaviour and the bootstrap. *Journal of Econometrics* 84: 59–74.

Goodman, A.C., and R.A. Dubin. 1991. Sample stratification with non-nested alternatives: Theory and a hedonic example. *The Review of Economics and Statistics* 72: 168–173.

Gourieroux, C., and A. Monfort. 1994. Testing nonnested hypothesis. In *Handbook of econometrics*, ed. R.F. Engle and D.L. McFadden, vol. 4. Amsterdam: North-Holland.

Gourieroux, C., and A. Monfort. 1995. Testing, encompassing, and simulating dynamic econometric models. *Econometric Theory* 11: 195–228.

Gourieroux, C., A. Monfort, and A. Trognon. 1983. Testing nested and nonnested hypothesis. *Journal of Econometrics* 21: 83–115.

Halaby, C.N., and D.L. Weakliem. 1993. Ownership and authority in the earnings function: Nonnested tests of alternative specifications. *American Sociological Review* 58: 16–30.

Kim, S., N. Shephard, and S. Chib. 1998. Stochastic volatility: Likelihood inference and comparison with ARCH models. *Review of Economic Studies* 65: 361–393.

Kullback, S. 1959. *Statistics and information theory.* New York: Wiley.

Lyon, C.C., and G.D. Thompson. 1993. Temporal and spatial aggregation: Alternative marketing margin models. *American Journal of Agricultural Economics* 75: 523–536.

McAleer, M., G. Fisher, and P. Volker. 1982. Separate misspecified regressions and the U.S. long run demand for money function. *The Review of Economics and Statistics* 64: 572–583.

McAleer, M., and S. Ling. 1998. *A nonnested tests for GARCH and EGARCH models*. Working paper, Department of Economics, University of Western Australia.

McAleer, M., and M.H. Pesaran. 1986. Statistical inference in nonnested econometric models. *Applied Mathematics and Computation* 20: 271–311.

McAleer, M., M.H. Pesaran, and A.K. Bera. 1990. Alternative approaches to testing nonnested models with autocorrelated disturbances: An application to models of U.S. unemployment. *Communications in Statistics A* 19: 3619–3644.

Mizon, G., and J.F. Richard. 1986. The encompassing principle and its applications to testing nonnested hypothesis. *Econometrica* 3: 657–678.

Otsu, T., and Y.J. Whang. 2005. Testing for non-nested conditional moment restrictions via conditional empirical likelihood. Discussion Paper No. 1533, Cowles Foundation, Yale University.

Pace, L., and A. Salvan. 1990. Best conditional tests for separate families of hypotheses. *Journal of the Royal Statistical Society B* 52: 125–134.

Pagan, A.R., A.D. Hall, and P.K. Trivedi. 1983. Assessing the variability of inflation. *Review of Economic Studies* 50: 585–596.

Pesaran, M.H. 1974. On the general problem of model selection. *Review of Economic Studies* 41: 153–171.

Pesaran, M.H. 1981. Pitfalls of testing nonnested hypotheses by the Lagrange multiplier method. *Journal of Econometrics* 17: 323–331.

Pesaran, M.H. 1982a. A critique of the proposed tests of the natural rate-rational expectations hypothesis. *Economic Journal* 92: 529–554.

Pesaran, M.H. 1982b. Comparison of local power of alternative tests of nonnested regression models. *Econometrica* 50: 1287–1305.

Pesaran, M.H. 1982c. On the comprehensive method of testing nonnested regression models. *Journal of Econometrics* 18: 263–274.

Pesaran, M.H. 1987. Global and partial nonnested hypothesis and asymptotic local power. *Econometric Theory* 3: 69–97.

Pesaran, M.H., and A.S. Deaton. 1978. Testing nonnested nonlinear regression models. *Econometrica* 46: 677–694.

Pesaran, M.H., and B. Pesaran. 1993. A simulation approach to the problem of computing Cox's statistic for testing nonnested models. *Journal of Econometrics* 57: 377–392.

Pesaran, M.H., and B. Pesaran. 1997. *Working with Microfit 4.0*. Oxford: Oxford University Press.

Pesaran, M.H., and S. Potter. 1997. A floor and ceiling model of US output. *Journal of Economic Dynamics and Control* 21: 661–696.

Pesaran, M.H., and M. Weeks. 2001. Nonnested hypothesis testing: An overview. In *Companion to theoretical econometrics*, ed. B.H. Baltagi. Oxford: Basil Blackwell.

Poterba, J.M., and L.H. Summers. 1983. Dividend taxes, corporate investments, and 'Q'. *Journal of Public Economics* 22: 135–167.

Quandt, R.E. 1974. A comparison of methods for testing nonnested hypothesis. *The Review of Economics and Statistics* 56: 92–99.

Ram, R. 1986. Government size and economic growth: A new framework and some evidence from cross-section and time-series data. *American Economic Review* 76: 191–203.

Ramalho, J.J.S., and R.J. Smith. 2002. Generalized empirical likelihood nonnested tests. *Journal of Econometrics* 107: 99–125.

Rivers, D., and Q. Vuong. 2002. Model selection tests for nonlinear dynamic models. *The Econometrics Journal* 5: 1–39.

Royston, P., and S.G. Thompson. 1995. Comparing nonnested regression models. *Biometrics* 51: 114–127.

Sandler, T., and J.C. Murdoch. 1990. Nash–Cournot or Lindahl behavior?: An empirical test for the NATO allies. *Quarterly Journal of Economics* 105: 875–894.

Santos Silva, J.M.C. 2001. A score test for non-nested hypothesis with applications to discrete data models. *Journal of Applied Econometrics* 16: 577–597.

Smith, R.J. 1992. Nonnested for competing models estimated by generalized method of moments. *Econometrica* 4: 973–980.

Vannetelbosch, V.J. 1996. Testing between alternative wage-employment bargaining models using Belgian aggregate data. *Labour Economics* 3: 43–64.

Victoria-Feser, M.-P. 1997. A robust tests for non-nested hypothesis. *Journal of the Royal Statistical Society B* 59: 715–727.

Vuong, Q.H. 1989. Likelihood ratio tests for model selection and nonnested hypothesis. *Econometrica* 57: 307–333.

Walker, A.M. 1967. Some tests of separate families of hypothesis in time series analysis. *Biometrika* 54: 39–68.

White, H. 1982. Regularity conditions for Cox's test of nonnested hypothesis. *Journal of Econometrics* 19: 301–318.

Non-Parametric Statistical Methods

Joseph L. Gastwirth

Basic statistics and econometrics courses stress methods based on assuming that the data or error term in regression models follow the normal distribution. Indeed, the efficiency of least squares estimates relies on the assumption of normality. In order to lessen the dependence of statistical inference on that assumption statisticians developed methods based on rank tests whose sampling distribution, under the null hypothesis, do not depend on the form of the underlying density function.

The simplest and oldest (Arbuthnott 1710) non-parametric test is the sign test used to test whether the median of a population equals a specified value υ_0. Let $x_1,\ldots, x_n$ be a sample of size n and let $s(x, \upsilon_0) = 1$ if $x > \upsilon_0$; $= 0$ otherwise. Then the statistic

$$S = \sum_{i=1}^{n} s(x_i, \upsilon_0) \tag{1}$$

has a binomial distribution with mean $n/2$ and variance $n/4$ if υ_0 is the true median, as $P[s(x) = 1] = 1/2$ regardless of the form of the density function. This contrasts with the classical t-test whose exact sampling distribution depends on normality.

The sign test also yields an estimate and confidence interval for υ the median when that parameter is unknown. The idea (see Hettmansperger 1984 for details) is that we can vary υ in the definition of $s(x)$ until we find that value or values for which S equals its expected value ($n/2$), that is, the estimate $\hat{\upsilon}$ satisfies

$$\sum_{i=1}^{n} s(x_i,\hat{\upsilon}) = 0 \tag{2}$$

and is simply the sample median. In contrast with the test statistic, the estimator derived from the sign test is not distribution-free although distribution-free confidence intervals are available. These are based on the binomial distribution of S and it can be demonstrated that the interval

$$\left[x_{(k+1)}, x_{(n-k)}\right], \tag{3}$$

where $x_{(i)}$ are the ordered observations, is a $100(1 - \alpha)$ per cent confidence interval for υ when k satisfies $P[S \leq k] = \alpha/2$.

More interesting uses of non-parametric tests occur when samples from two populations are compared. Suppose we desire to see whether male and female college graduates earn similar wages after working for five years. We are testing whether the earning distribution of females, $G(x)$ equals that, $F(x)$, of males. One possible alternative is that $G(x) = F(x - \Delta)$, that is, the female distribution is shifted up by Δ. If the distributions are found to be significantly different we will then estimate Δ.

To resolve the issue, we take random samples $x_1, \ldots, x_m$ from $F(x)$ and $y_1, \ldots, y_n$ from $G(x)$. Consider the combined sample of $N = m + n$ observations. Under the null hypothesis that the two distributions are the same, that is, $F(x) = G(x)$, it can be shown that each of the original observations has probability $1/N$ of being the kth largest in the pooled sample. Thus the ranks in the combined sample that the n y's have can be considered as a random sample of n integers chosen from $1, \ldots, N$, irrespective of the form of the distribution function $F(x)$. Any test which is solely a function of the ranks that one group of observations has in the combined sample is called a rank test. If we let R_i be the rank y_i has in the ordered combined sample of N, then the Wilcoxon (1945) test is defined as $W = \sum R_i$, and its distribution is that of the sum of n randomly selected integers from $1, 2, \ldots, N$. Since the average of the first N integers is $(N + 1)/1$, under the null hypothesis the expected value of W is $n(N + 1)/2$. Furthermore, its variance is $nm(N + 1)/12$ and its standardized form

$$\frac{W - n(N+1)/2}{\sqrt{nm(N+1)/12}} \tag{4}$$

rapidly approaches the unit normal variate as n and m increase. If the observed value of W is much larger than expected, for example the standardized from exceeds $\pm z_{\alpha/2}$ where α is the preset significance level we reject the hypothesis that the x's and y's have the same distribution in favour of the alternative and conclude that the distribution of the y's is shifted to the right, i.e. $\Delta > 0$.

So far we have limited our attention to ensuring that the probability of rejecting the hypothesis that both populations are the same when it is indeed true, is small (5 per cent or less). The advantage of rank statistics is that this calculation is the same regardless of the form of the density function of the variable. On the other hand, we also desire to reject the null hypothesis when the two populations truly differ. If c is the critical point of a test of size α (often 0.05), that is, the probability, when the null hypothesis is true, of obtaining a value $> c$, is $\leqslant \alpha$ written $P_0[W > c] \leq \alpha$, then the power of the test is the probability of $[W > c]$ calculated under the alternative assumption (e.g. $\Delta = 1$). While the size (α) of the Wilcoxon (or any rank test) does not depend on the form of the underlying density function, its power does. The remarkable fact about the Wilcoxon test is that it is about 95 per cent as powerful as the usual t-test for normal data. Hence, one pays a rather small price in terms of loss of power for guaranteeing that the Type I error (size) is not affected by the form of the density.

The Wilcoxon test also has an equivalent form, due to Mann and Whitney (1947), based on comparing each of the x's with each of the y's. Let

$$I_{ij} = \begin{cases} 1, \text{if } y_i > x_j \\ 1/2, \text{if } y_i = x_j \\ 0, 0 \text{otherwise}, \end{cases} \tag{5}$$

The statistic $W = \Sigma\Sigma \mathrm{I}_{ij}$, counts the number of times a y observation exceeds an x observation. Notice that

$$\Sigma R_i = W + n(n+1)/2 \tag{6}$$

since, if the y's are the smallest n observations in the total of $m + n$, $W = 0$ and $\Sigma R_i = n(n + 1)/2$. As we move the y's up to obtain our sample ranks every time a y exceeds an x both ΣR_i and

W increase by 1. This form of the Wilcoxon test has two desirable features. First, *W*/*mn* estimates an interesting parameter, $P[X < Y]$, the probability that a randomly selected y (female earnings) exceeds a randomly selected x (male earnings) under the null hypothesis that $F = G$, $P[X < Y]$ should equal 1/2. This measure can also indicate whether 'progress' towards equality is made over time. Secondly, the amount of shift, Δ, that needs to be added to the y's so that

$$\Sigma R_i\left(x_i, y_j + \Delta\right) - n(N+1)/2 = 0, \qquad (7)$$

that is, the Wilcoxon test calculated on the old x's and the new $(y_i + \Delta)$'s equals its expected value under H_0, can be expressed as the *median* of the *mn* differences $(y_i - x_j) = D_{ij}$ (see Lehmann 1975, p. 82) and is an alternative estimate of the difference between the location parameters of the two distributions.

So far we have only used the sum of the ranks of the observation of our sample (y'·s) as a test statistic. More generally one can use a statistic of the form $\Sigma a(R_i)$ where R_i is the rank of y_i and $a(R_i)$ is specified by $a[R_i/(N+1)]$, where $a(u)$ is a function on (0, 1). The following basic result, due to Chernoff and Savage (1958), shows that there is a non-parametric test with the same large sample power as the best parametric test for the problem when the density function $f = F'$ is known:

Theorem 1 Let $x_1, \ldots, x_m; y_1, \ldots, y_n$ be two independent samples from the distributions $F(x)$ and $G(x) = F(x - \Delta)$, respectively, and assume that $f = F'$ has finite Fisher information, that is $I = \int (f'/f)^2 f\,dx < \infty$. The asymptotically most powerful rank test of H_0: $\Delta = 0$ against $\Delta \neq 0$ is based on the function

$$a(u) = I^{-1/2} f'\left[F^{-1}(u)\right] / f\left[F^{-1}(u)\right],\ 0 < u < 1, \qquad (8)$$

and is asymptotically as powerful as the best parametric (maximum likelihood) procedure. In particular, if $F(x) = \Phi(x)$, the standard normal distribution, $a(u) = \Phi^{-1}(u)$ and generates the normal scores test. The Wilcoxon test is the optimal test for data from the logistic law. For further examples see Hajek and Sidak (1967).

The next problem one faces is how to choose the rank test or score function $a(u)$ as almost any reasonable test is consistent. To guide this choice we use the Pitman efficiency $e(T_1, T_2)$ which compares the power of two tests of the $H_0 : \Delta = 0$ against a sequence of alternatives $\Delta = \sigma/\surd N$ which approach the null hypothesis as the sample size increases. The Pitman efficiency $e(T_1, T_2)$ can be interpreted as the limiting ratio of the sample sizes required by the tests T_1 and T_2 to achieve the same limiting power π against the same sequence of alternatives. For example, if the Pitman efficiency of test T_1 relative to T_2 $e(T_1, T_2) = 1/2$, then the test T_2 requires approximately half as many observations as the test T_1 to achieve the same large sample power for critical regions of the same size α. Moreover, the relative efficiencies of the corresponding estimates is also given by $e(T_1, T_2)$. The Pitman efficiency can be easily computed. The efficiency of T_1 (based on $a_1(u)$) relative to the best test, T_2, on data from the density f_2 is given by

$$\langle a_1, a_2 \rangle = \int a_i(u) a_2(u) \mathrm{d}u, \qquad (9)$$

where a_2 is obtained from (8). By the symmetry of the inner product T_2 has this same efficiency relative to the best test, T_1, for data from the density f_1. The fact that the functions $a(u)$ generating the most powerful rank tests for a wide family of densities are in $L_2(0, 1)$ yields insight into other problems as well. If one truly knew the form of the density, why use a rank test instead of the usual maximum likelihood test? Suppose one knew something about the density, for example that it was either f_1 or f_2 (normal or double exponential): Is there a reasonably powerful rank test for this problem? Considering the functions a_1, a_2 as vectors in $L^2(0, 1)$, it is clear that a test corresponding to the angle bisector will maximize the minimum efficiency when data come from either density. In fact, the robust test obtained in this manner is nearly 90 per cent as efficient as the best tests when they fit the model. On the other hand, if the normal scores test is used when the data are from a double-exponential, it has only 64 per cent efficiency. This general problem is discussed in Gastwirth (1966) and Birnbaum and Laska (1967).

The reason we reviewed the two-sample problem at length is that the relative efficiencies of the tests and derived parameter estimates (Hodges and Lehmann 1963; Bauer 1972) typically extend to their analogues in regression and linear models. Thus, once an appropriate nonparametric test is selected it can be used for the same family of possible error distributions in more complex general linear models.

Before discussing regression models we note that an alternative approach to account for the effect of covariates is to stratify the data into homogeneous subgroups, compare the two samples in each subgroup using the same rank test and combine the *results* into our summary statistic. In the male–female earnings example one might stratify the data by occupation. The Wilcoxon procedure was generalized by van Elteren (1960) and developed further by others (see Oosterhoff 1969) and yields a summary estimate of the parameter $P[X < Y]$. General rank tests were considered by Puri (1965).

The first analogues of rank tests for analysis of variance (Friedman 1937; Brown and Mood 1951; Kruskal and Wallis 1952) and regression models (Theil 1950) were based on extending the Wilcoxon and median tests and Kendall's measure, τ, of dependence of bivariate data (Sen 1968). The generalizability of tests based on score functions, $a(u)$ to these more general situations was made possible by the results of Jureckova (1969, 1971). The analogues of the normal equations of least squares are a non-linear system of equations which can be 'linearized' by her techniques. We next introduce these ideas in the simple linear regression model

$$Y_i = \alpha + x_i\beta + e_i, \qquad (10)$$

where e_i are i.i.d. with mean 0 and the x_i's are fixed known numbers. To test whether the slope $\beta = 0$, ordinary least squares theory uses

$$\widehat{\beta} = \frac{\sum (x_i - \overline{x})y_i}{\sum (x_i - \overline{x})^2} \qquad (11)$$

which has a variance $\sigma^2 \sum (x_i - \overline{x})^2$ and mean 0 if $\beta = 0$. The extension of the Wilcoxon procedure replaces the y_i in the numerator of (11) by their ranks, R_i; that is, considers the statistic

$$T = \sum (x_i - \overline{x})R_i. \qquad (12)$$

If $H_0 : \beta = 0$ holds, then the y_i's are i.i.d. variates with location parameter α and the ranks of the y_i's, just like their numerical values, should be uncorrelated with $(x_i - x)$, that is $E(T) = 0$. Moreover,

$$\mathrm{var}(T) = \frac{nN(N+1)}{12} \sum (x_i - \overline{x})^2.$$

To estimate β we consider T as a function of a possible value of β that is

$$T(\beta) = \sum (x_i - \overline{x})R(y_i - \alpha - \beta x_i), \qquad (13)$$

where $R(y_i - \alpha - \beta x_i)$ is the rank of the residual $y_i - \alpha - \beta x_i$. Because the ranks of the y's do not depend on α that is if the intercept of the line were increased (or decreased), the ranks of the $y_i - \alpha - \beta x_i$ would remain the same, one can take $\alpha = 0$. If β^* is the true value of β then

$$E[T(\beta*)] = 0 \qquad (14)$$

and we can estimate β by the value $\widehat{\beta}$ which satisfies

$$T\left(\widehat{\beta}\right) = \sum (x_i - \overline{x})R\left(y_i - \widehat{\beta}x_i\right) = 0. \qquad (15)$$

Unfortunately, $\widehat{\beta}$ typically must be obtained by numerical means although its distribution in large samples approaches a normal law. If an estimate of α is also desired, the median of the $\left\{y_i - \widehat{\beta}x_i\right\}$ can be used.

Rank tests and estimates of β based on other score functions, $a(u)$ replace R_i by $a(R_i/N + 1)$ etc. The relative efficiency of these procedures is the same as their values in the two sample problem so that knowledge of nature of the error distribution should be used in selecting a nonparametric test. A large literature (Hettmansperger 1984) has been devoted to obtaining nonparametric methodology for the multiple regression and linear models. The basic idea of estimation is to find the vector β minimizing

$$\sum a\left[\frac{R\left(Y_i - x_i'\beta\right)}{N+1}\right]\left(Y_i - x_i'\beta\right), \tag{16}$$

in the model

$$y = [1\,X]\begin{pmatrix}\alpha\\ \beta\end{pmatrix} + e,$$

where $Y = (Y_1,\ldots Y_n)'$, 1 is an $n \times 1$ column of 1's, X is the $N \times p$ matrix of regression constants and β the regression parameters. Various conditions may be imposed on the score function, a, in (16) for example $a(u) = a(1 - u)$ and $a(u)$ is an increasing function, in order that the measure (16) is a proper measure of dispersion. The mathematical methods of finding the estimate β in (16) involve solving a set of non-linear equations satisfying

$$\sum_{i=1}^{N}\left(x_{ij} - \overline{x}_j\right)a\left[R\left(y_i - x_i'\beta\right)\right]\check{Z}0, \quad j = 1, \ldots, p \tag{17}$$

which play the role of the normal equations of OLS.

Although methods based on rank tests are less sensitive to the distributional assumptions of classical procedures, they lose their distribution free character when the observations are dependent. For example, the distribution of the sampling distribution of the sign test or first order autoregressive processes with the same ρ depends on the form of the underlying distribution (Wolff et al. 1967). Thus, the usual diagnostic checks based on examining the residuals should be carried out even when the regression model is fitted by nonparametric or other robust methods (Huber 1972; Bickel 1973; Hogg 1974). Of course, the dependence also affects the distribution of least squares estimates and Gastwirth and Rubin (1971) show that the level of the test using the sample mean is more sensitive to dependence than the sign or Wilcoxon procedures.

Bibliography

Arbuthnott, J. 1710. An argument for divine providence taken from the constant regularity observed in the birth of both sexes. *Philosophical Transactions* 27: 186–190.

Bauer, D.F. 1972. Constructing confidence sets using rank statistics. *Journal of the American Statistical Association* 67: 687–690.

Bickel, P.J. 1973. On some analogs to linear combination of order statistics in the linear model. *Annals of Statistics* 1: 597–616.

Birnbaum, A., and E. Laska. 1967. Efficiency robust two-sample rank tests. *Journal of the American Statistical Association* 62: 1241–1251.

Brown, G.W. and A.M. Mood. 1951. On median tests for linear hypotheses. *Proceedings of the 2nd Berkeley Symposium,* 159–166.

Chernoff, H., and I.R. Savage. 1958. Asymptotic normality and efficiency of certain nonparametric test statistics. *Annals of Mathematical Statistics* 29: 972–994.

van Elteren, P. 1960. On the combination of independent two sample tests of Wilcoxon. *Bulletin de l'Institut Internationale de Statistique* 37(3): 351–360.

Friedman, M. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association* 32: 675–701.

Gastwirth, J.L. 1966. On robust procedures. *Journal of the American Statistical Association* 61: 929–948.

Gastwirth, J.L., and H. Rubin. 1971. The behavior of the level of rank tests on dependent data. *Journal of the American Statistical Association* 66: 816–820.

Hajek, J., and Z. Sidak. 1967. *Theory of rank tests*. New York: Academic Press.

Hettmansperger, T.P. 1984. *Statistical inference based on ranks*. New York: Wiley.

Hodges Jr., J.L., and E.L. Lehmann. 1963. Estimates of location based on rank tests. *Annals of Mathematical Statistics* 34: 598–611.

Hogg, R.V. 1974. Adaptive robust procedures: A partial review and some suggestions for future research. *Journal of the American Statistical Association* 69: 909–927.

Hollander, M., and D.A. Wolfe. 1973. *Nonparametric statistical methods*. New York: Wiley.

Huber, P.J. 1972. Robust statistics: A review. *Annals of Mathematical Statistics* 43: 1041–1067.

Jureckova, J. 1969. Asymptotic linearity of a rank statistic in regression parameter. *Annals of Mathematical Statistics* 40: 1889–1950.

Jureckova, J. 1971. Nonparametric estimate of regression coefficients. *Annals of Mathematical Statistics* 42: 1328–1338.

Kruskal, W.H., and W.A. Wallis. 1952. Use of ranks in one criterion variance analysis. *Journal of the American Statistical Association* 57: 583–621.

Lehmann, E.L. 1975. *Nonparametrics: Statistical methods based on ranks*. San Francisco: Holden-Day.

Mann, H.B., and D.R. Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics* 18: 50–60.

Oosterhoff, J. 1969. *Combination of one-sided statistical tests*. Amsterdam: Mathematical Centre.

Puri, M.L. 1965. On the combination of independent two sample tests of a general class. *Review of the ISI* 33: 229–241.

Sen, P.K. 1968. Estimates of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association* 63: 1379–1389.

Theil, H. 1950. A rank invariant method of linear and polynomial regression analysis. *Proceedings: Koninklijke Nederlandse Akademie van Wetenschappen, Series A* 53: 386–392.

Wilcoxon, F. 1945. Individual comparison by ranking methods. *Biometrics* 1: 80–83.

Wolff, S.S., Rubin, H., and J.L. Gastwirth. 1967. The effect of autoregression dependence on a nonparametric test. Professional Group on Information Theory IEEE, IT–13: 311–313.

Non-parametric Structural Models

Rosa L. Matzkin

Abstract

Nonparametric structural models facilitate the analysis of counterfactuals without making use of parametric assumptions. Such methods make use of the behavioural and equilibrium assumptions specified in economic models to define a mapping between the distribution of the observable variables and the primitive functions and distributions that are used in the model. Using these methods, one can infer elements of the model, such as utility and production functions, that are not directly observed. We review some of the latest works that have dealt with the identification and estimation of nonparametric structural models.

Keywords

Additivity; Average derivative methods; Control functions; Convergence; Curse of dimensionality; Endogeneity; Estimation; Identification; Instrumental variables; Maximum likelihood; Nonadditivity; Nonparametric structural models; Nonseparable models; Observable and unobservable explanatory variables; Partial integration methods; Quantile structural functions; Simultaneous equations

JEL Classifications

C14

The interplay between economic theory and econometrics comes to its full force when analysing structural models. These models are used in industrial organization, marketing, public finance, labour economics and many other fields in economics. Structural econometric methods make use of the behavioural and equilibrium assumptions specified in economic models to define a mapping between the distribution of the observable variables and the primitive functions and distributions that are used in the model. Using these methods, one can infer elements of the model, such as utility and production functions, that are not directly observed. This allows one to predict behaviour and equilibria outcomes under new environments and to evaluate the welfare of individuals and profits of firms under alternative policies, among other benefits.

To provide an example, suppose that one would like to predict the demand for a new product. Since the product has not previously been available, no direct data exists. However, one could use data on the demand for existent products together with a structural model, as shown and developed by McFadden (1974). Characterize the new product and the existent competing products by vectors of common attributes. Assume that consumers derive utility from the observable and unobservable attributes of the products, and that each chooses the product that maximizes his or her utility of those attributes among the existent products. Then, from the choice of consumers among existent products, one can infer their preferences for the attributes, and then predict what the choice of each of them would be in a situation when a new vector of attributes, corresponding to the new product, is available. Moreover, one could get a measure of the differences in the welfare of the consumers when the new product is available.

Economic theory seldom has implications regarding the parametric structures that functions and distributions may possess. The behavioural and equilibrium specifications made in economic models typically imply shape restrictions, such as monotonicity, concavity, homogeneity, weak separability, and additive separability, and exclusion restrictions, but typically not parametric specifications, such as linearity of conditional expectations, or normal distributions for unobserved variables. Nonparametric methods, which do not require specification of parametric structures for the functions and distributions in a model, are ideally fitted, therefore, to analyse structural models, using as few a priori restrictions as possible. Nonparametric techniques have been applied to many models, such as discrete choice models, tobit models, selection models, and duration models. We will concentrate here, however, on the basic models and on those, indicate some of the latest works that have dealt with identification and estimation.

Nonparametric Structural Econometric Models

As with parametric models, a nonparametric econometric model is characterized by a vector X of variables that are determined outside the model and are observable, a vector ε of variables that are determined outside the model and are unobservable, a vector Υ of outcome variables, which are determined within the model and are unobservable, and a vector $Y$of outcome variables that are determined within the model and are observable. These variables are related by functional relationships, which determine the causal structure by which Υ and $Y$are determined from X and ε. The functional relationships are characterized by some functions that are known and some that are unknown. Similarly, some distributions may be known, some are unknown, and the others should be derived from the functional relationships and the known and unknown functions and distributions. Let $\underline{h}^*$ denote the vector of all the unknown functions in the model, $\underline{F}^*$ denote the vector of all unknown distributions, and $\zeta^* = (\underline{h}^*, \underline{F}^*)$-In contrast to parametric models, in nonparametric models, none of the coordinates of ζ^* is assumed known up to a finite dimensional parameter.

Only restrictions such as continuity or values of the conditional expectations are assumed. The specification of the model should be such that from any vector $\zeta = (\underline{h}, \underline{F})$, satisfying those same restrictions that ζ^* is assumed to satisfy, one is able to derive a distribution for the observable variables, $F_{YX}(\cdot, \zeta)$.

Nonparametric Identification

When specifying an econometric model, we may be interested in testing it, or we may be interested in estimating $\zeta^* = (\underline{h}^*, \underline{F}^*)$ or some feature of ζ^*, such as only one of the elements of $\underline{h}^*$, or even the value of that element at one point. Suppose that interest lies on estimating a particular feature, $\psi(\zeta^*)$ of ζ^*. The first question one must answer is whether that feature is identified. Let Ω denote the set of all possible values that $\psi(\zeta)$ may attain, when ζ is restricted to satisfy the properties that ζ^* is assumed to satisfy. Given $\psi \in \Omega$, define $\Gamma_{Y, X}(\psi)$ to be the set of all probability distributions of (Y, X) that are consistent with ψ. This is the set of all distributions that can be generated by a ζ, satisfying the properties that ζ^* is assumed to satisfy, and with $\psi(\zeta) = \psi$. We say that two values $\psi, \psi' \in \Omega$ are *observationally equivalent* if

$$\left[\Gamma_{Y,X}(\psi) \cap \Gamma_{Y,X}(\psi')\right] \neq \varnothing,$$

that is, they are observationally equivalent if there exist a distribution of the observable variables that could have been generated by two vectors ζ and ζ' with $\psi(\zeta) = \psi$ and $\psi(\zeta') = \psi'$. The feature $\psi^* = \psi(\zeta^*)$ is said to be identified if there is no $\psi \in \Omega$ such that $\psi \neq \psi^*$ and ψ is observationally equivalent to ψ^*. That is, $\psi^* = \psi(\zeta^*)$ is identified if a change from ψ^* to $\psi \neq \psi^*$ cannot be compensated by a change in other unknown elements of ζ, so that a same distribution of observable variables could be generated by both, vectors ζ^* and ζ with $\psi^* = \psi\,(\zeta^*)$ and $\psi = \psi\,(\zeta)$.

When ψ^* can be expressed as a continuous functional of the distribution of observable variables (Y, X) one can typically estimate ψ^* nonparametrically by substituting the distribution by a nonparametric estimator for it.

Additive and Nonadditive Models with Exogenous Explanatory Variables

The current literature on nonparametric econometrics methods considers additive and nonadditive models. In *additive models,* the unobservable exogenous variables ε are specified as affecting the value of Y though an additive function. Hence, for some functions m and v and some unobservable η

$$Y = m(X) + v(X, \varepsilon) = m(X) + \eta.$$

In these models, the object of interest is typically the function m. Depending on the restrictions that one may impose on η, m may denote a conditional expectation, a conditional quantile, or some other function. Many methods exist to estimate conditional means and conditional quantiles nonparametrically. Prakasa Rao (1983), Härdle and Linton (1994), Matzkin (1994, 2007b), Pagan and Ullah (1999), Koenker (2005), and Chen (2007), among others, survey parts of this literature. Some nonparametric tests for these models include Wooldridge (1992), Yatchew (1992), Hong and White (1995), and Fan and Li (1996).

In *nonadditive models*, one is interested in analysing the interaction between the unobservable and observable explanatory variables. These models are specified, for some function m as

$$Y = m(X, \varepsilon).$$

Nonparametric identification and estimation in models of this type was studied in Roehrig (1988), Olley and Pakes (1996), Brown and Matzkin (1998), Matzkin (1999, 2003, 2004, 2005, 2006), Chesher (2003), Imbens and Newey (2003), and Athey and Imbens (2006), among others.

Dependence Between Observable and Unobservable Explanatory Variables

In econometric models, it is often the case that in an equation of interest, some of the explanatory variables are endogenous; they are not distributed independently of the unobservable explanatory variables in that same equation. This typically occurs when restrictions such as agent's optimization and equilibrium conditions generate interrelationships among observable variables and unobservable variables, ε, that affect a common observable outcome variable, Y. In such cases, the distribution of the observable outcome and observable explanatory variables does not provide enough information to recover the causal effect of those explanatory variables on the outcome variable, since changes in those explanatory variables do not leave the value of ε fixed. A typical example of this is when Y denotes quantity demanded for a product, X denotes the price of the product, and ε is an unobservable demand shifter. If the price that will make firms produce a certain quantity increases with quantity, this change in ε will generate an increment in the price X. Hence, the observable effect of a change in price in demanded quantity would not correspond to the effect of changing the value of price when the value ε stays constant. Another typical example arises when analysing the effect of years of education on wages. An unobservable variable, such as ability, affects wages and also affects the decision about years of education.

Estimation Techniques for Additive and Nonadditive Functions of Endogenous Variables

The estimation techniques that have been developed to estimate nonparametric models with endogenous explanatory variables typically make use of additional information, which provides some exogenous variation on either the value of the endogenous variable or on the value of the unobservable variable. The common procedures are based on conditional independence methods and on instrumental variable methods. In the first set of procedures, independence between the unobservable and observable explanatory variables in a model is typically achieved by either *conditioning on observable* variables, or *conditioning on unobservable* variables. A *control function* approach (Heckman and Robb 1985) models the unobservable as a function, so that conditioning on that function purges the dependence between the explanatory observable and unobservable variables in the model. Instrumental variable methods derive identification from an independence condition between the unobservable and an external variable (an instrument) or function, which is correlated with the endogenous variable and which might be estimable.

Conditioning on unobservable variables often requires the estimation of those unobservable variables. Two-step procedures, where they are first estimated, and then used as additional regressors in the model of interest have been developed for additive models by Ng and Pinkse (1995), Pinkse (2000), and Newey et al. (1999), among others. Two-step procedures for nonadditive models have been developed by Altonji and Matzkin (2001), Blundell and Powell (2003), Chesher (2003), and Imbens and Newey (2003), among others. Conditional moment estimation methods or quasi-maximum likelihood estimation methods can also be used (see, for example, Ai and Chen 2003). Altonji and Ichimura (2000), Altonji and Matzkin (2001, 2005), and Matzkin (2004), among others, considered conditioning on observables for estimation of nonadditive models with endogenous explanatory variables. Matzkin (2004) provides insight into the sources of exogeneity that are generated when conditioning on either observables or unobservables, and which allow identification and estimation in nonadditive models. In particular, if $Y = m(X, \varepsilon)$, with m strictly increasing in ε, and ε is independent of X conditional on W, she shows that there exists functions $s(W,\eta)$ and $r(W, \delta)$ such that δ is independent η conditional on W, $X = s(W, \eta)$ and $\varepsilon = r(W, \delta)$. Hence,

$$Y = m(X, \varepsilon) = m(s(W, \eta), r(W, \delta)).$$

Instrumental variable methods for additive models were considered by Newey and Powell (1989, 2003), Ai and Chen (2003), Darolles et al. (2003), and Hall and Horowitz (2003), among others. To develop estimators for m in the model

$$Y_1 = m(Y_2) + \varepsilon \;\; E[\varepsilon | Z] = 0.$$

they use the moment condition

$$E[Y_1 | Z = z] = \int m(y_2) f_{Y_2|Z=z}(y_2) dy_2,$$

which depends on the conditional expectation $E[Y_1|Z = z]$ and the conditional density $f_{Y_2|Z=z}(y_2)$, which can be estimated nonparametrically. For nonadditive models, of the form

$$Y_1 = m(Y_2, \varepsilon) \quad \varepsilon \quad \text{independent of} \;\; Z$$

where m is strictly increasing in ε, Chernozhukov and Hansen (2005) and Chernozhukov et al. (2007) developed estimation methods using the moment condition that for al τ

$$\tau = E[1(\varepsilon < \tau) = E1(\varepsilon < \tau) | Z]$$

from which m can be estimated using the conditional moment restriction

$$E[1(Y_1 < m(Y_2, \varepsilon)) - \varepsilon | Z] = 0.$$

Matzkin (2006) considered the model

$$Y_1 = m_1(Y_2, \varepsilon) \; Y_2 = m_2(Y_1, \;\; Z, \eta)$$

where Z is distributed independently of (ε, η). She established restrictions on the functions m_1 and m_2 and on the distribution of (ε, η, Z) under which

$$\left[\frac{\partial r_1(y_1, \;\; y_2)}{\partial y_2}\right]^{-1} \left[\frac{\partial r_1(y_1, \;\; y_2)}{\partial y_1}\right]$$

can be expressed as a function of the conditional density $f_{Y_1,Y_2|Z=z^*}(y_1, y_2)$, where r_1 is the inverse of m_1 with respect to ε, and the value z^* of the instrument Z is easily identified (see also Matzkin 2005, 2007a, b).

Estimation of Averages and Average Derivatives

Nonparametric estimators are notorious by their slow rate of convergence, which worsens as the dimension of the number of arguments of the nonparametric function increases. A remedy for this is to consider averages of the nonparametric function. The average derivative method in Powell et al. (1989) and the partial integration methods of Newey (1994) and Linton and Nielsen (1995), for example, show how rates of convergence can increase by averaging a nonparametric function or its derivatives. This approach has been extended to cases where the explanatory variables are endogenous, using additional variables to deal with the endogeneity, and averaging over them. Examples are Blundell and Powell's (2003) *average structural function,* Imbens and Newey's (2003) *average quantile function,* and Altonji and Matzkin's (2001, 2005) *local average response* function.

Suppose, for example, that the model of interest is

$$Y_1 = m(Y_2, \varepsilon)$$

and W is such that Y_2 and ε are independent conditional on W. Then, the Blundell and Powell (2003) average structural function is

$$G(y_2) = \int m(y_2, \varepsilon) f_\varepsilon(\varepsilon) d\varepsilon$$

which can be derived from a nonparametric estimator for the distribution of (Y_1, Y_2, W) as

$$G(y_2) = \int E(Y_1 | Y_2 = y_2, W = w) f_W(w) dw.$$

Imbens and Newey's (2003) quantile structural function is defined for the τ-th quantile of ε, $q_\varepsilon(\tau)$, as

$$r(y_2, y_1) = \Pr(m(Y_2, q_\varepsilon(\tau)) \leq y_1 | Y_2 = y_2)$$

which can be estimated by

$$r(y_2, y_1) = \int \Pr(Y_1 \leq y_1 | Y_2 = y_2, W = w) f_W(w) dv.$$

Altonji and Matzkin's (2001, 2005) local average response function is

$$\beta(y_2) = \int \frac{\partial m(y_2, \varepsilon)}{\partial y_2} f_{\varepsilon|Y_2=y_2}(\varepsilon) d\varepsilon$$

which can be derived from a nonparametric estimator for the distribution of (Y_1, Y_2, W) as

$$\beta(y_2) = \int \frac{\partial E(Y_1 | Y_2 = y_2, W = w)}{\partial y_2} f_{W|Y_2=y_2}(w) \times dw.$$

Conclusions

The literature on nonparametric structural models has been rapidly developing in recent years. The new methods allow one to analyse counterfactuals without making use of parametric assumptions. Estimation of some features of the model rather than the functions themselves may reduce the curse of dimensionality, therefore providing improved properties and reducing the need for large data-sets.

See Also

▶ Endogeneity and Exogeneity
▶ Identification
▶ Quantile Regression
▶ Simultaneous Equations Models

Bibliography

Ai, C., and X. Chen. 2003. Efficient estimation of models with conditional moments restrictions containing unknown functions. *Econometrica* 71: 1795–1843.

Altonji, J.G., and H. Ichimura. 2000. Estimating derivatives in nonseparable models with limited dependent variables. Mimeo, Northwestern University.

Altonji, J.G., and R.L. Matzkin. 2001. Panel data estimators for nonseparable models with endogenous regressors. NBER Working paper T0267.

Altonji, J.G., and R.L. Matzkin. 2005. Cross section and panel data estimators for nonseparable models with endogenous regressors. *Econometrica* 73: 1053–1102.

Athey, S., and G. Imbens. 2006. Identification and inference in nonlinear difference-in-differences models. *Econometrica* 74: 431–497.

Blundell, R., and J.L. Powell. 2003. Endogeneity in nonparametric and semiparametric regression models. In *Advances in economics and econometrics, theory and applications, eighth world congress*, ed. M. Dewatripont, L.P. Hansen, and S.J. Turnovsky, vol. 2. Cambridge: Cambridge University Press.

Brown, D.J, and R.L. Matzkin. 1998. Estimation of nonparametric functions in simultaneous equations models, with an application to consumer demand. Discussion paper no. 1175, Cowles Foundation, Yale University.

Chen, X. 2007. Large sample sieve estimation of semi-nonparametric models. In *Handbook of econometrics*, ed. E. Leamer and J.J. Heckman, vol. 6. Amsterdam: North-Holland.

Chernozhukov, V., and C. Hansen. 2005. An IV model of quantile treatment effects. *Econometrica* 73: 245–261.

Chernozhukov, V., G. Imbens, and W. Newey. 2007. Instrumental variable estimation of nonseparable models. *Journal of Econometrics* 139: 4–14.

Chesher, A. 2003. Identification in nonseparable models. *Econometrica* 71: 1404–1441.

Darolles, S., J.P. Florens, and E.Renault. 2003. Nonparametric instrumental regression. IDEI Working paper no. 228.

Fan, Y., and Q. Li. 1996. Consistent model specification tests: Omitted variables and semiparametric functional forms. *Econometrica* 64: 4.

Florens, J.P. 2003. Inverse problems and structural econometrics: The example of instrumental variables. In *Advances in economics and econometrics, theory and applications*, ed. M. Dewatripont, L.P. Hansen, and S. Turnovsky, vol. 2. Cambridge: Cambridge University Press.

Hall, P., and J.L. Horowitz. 2003. Nonparametric methods for inference in the presence of instrumental variables. Working paper no. 102/03, CMMD.

Härdle, W., and O. Linton. 1994. Applied nonparametric methods. In *Handbook of econometrics*, ed. R.F. Engel and D.F. McFadden, vol. 4. Amsterdam: North-Holland.

Heckman, J.J., and R. Robb. 1985. Alternative methods for evaluating the impact of interventions. In *Longitudinal analysis of labor market data*, ed. J.J. Heckman and B. Singer. Cambridge: Cambridge University Press.

Hong, Y., and H. White. 1995. Consistent specification testing via nonparametric series regression. *Econometrica* 63: 1133–1159.

Imbens, G.W., and W.K. Newey. 2003. Identification and estimation of triangular simultaneous equations models without additivity. Mimeo, Massachusetts Institute of Technology.

Koenker, R.W. 2005. *Quantile regression*. Cambridge: Cambridge University Press.

Linton, O.B., and J.B. Nielsen. 1995. A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika* 82: 93–100.

Matzkin, R.L. 1994. Restrictions of economic theory in nonparametric methods. In *Handbook of econometrics*, ed. R.F. Engel and D.L. McFadden, vol. 4. Amsterdam: North-Holland.

Matzkin, R.L. 1999. Nonparametric estimation of nonadditive random functions. Mimeo, Northwestern University.

Matzkin, R.L. 2003. Nonparametric estimation of nonadditive random functions. *Econometrica* 71: 1339–1375.

Matzkin, R.L. 2004. Unobservable instruments. Mimeo, Northwestern University.

Matzkin, R.L. 2005. Identification in nonparametric simultaneous equations. Mimeo, Northwestern University.

Matzkin, R.L. 2006. Estimation of nonparametric simultaneous equations. Mimeo, Northwestern University.

Matzkin, R.L. 2007a. Heterogenous choice. Invited lecture, ninth world congress of the econometric society. In *Advanced in economics and econometrics, theory and applications, ninth world congress*, ed. R. Blundell, W. Newey, and T. Persson, vol. 3. Cambridge: Cambridge University Press.

Matzkin, R.L. 2007b. Nonparametric identification. In *Handbook of econometrics*, ed. E. Leamer and J.J. Heckman, vol. 6. Amsterdam: North-Holland.

McFadden, D. 1974. Conditional logit analysis of qualitative choice behavior. In *Frontiers in econometrics*, ed. P. Zarembka. New York: Academic Press.

Newey, W.K. 1994. Kernel estimation of partial means and a general variance estimator. *Econometric Theory* 10: 233–253.

Newey, W., and J. Powell. 1989. Instrumental variables estimation of nonparametric models. Mimeo, Princeton University.

Newey, W., and J. Powell. 2003. Instrumental variables estimation of nonparametric models. *Econometrica* 71: 1565–1578.

Newey, W.K., J.L. Powell, and F. Vella. 1999. Nonparametric estimation of triangular simultaneous equations models. *Econometrica* 67: 565–603.

Ng, S., and J. Pinkse. 1995. Nonparametric two-step estimation of unknown regression functions when the regressors and the regression error are not independent. Mimeo, CIREQ.

Olley, G.S., and A. Pakes. 1996. The dynamics of productivity in the telecommunications equipment industry. *Econometrica* 64: 1263–1297.

Pagan, A., and A. Ullah. 1999. *Nonparametric econometrics*. Cambridge: Cambridge University Press.

Pinkse, J. 2000. Nonparametric two-step regression estimation when regressors and errors are dependent. *Canadian Journal of Statistics* 28: 289–300.

Powell, J.L., J.H. Stock, and T.M. Stoker. 1989. Semiparametric estimation of index coefficients. *Econometrica* 51: 1403–1430.

Prakasa Rao, B.L.S. 1983. *Nonparametric functional estimation*. New York: Academic Press.

Roehrig, C.S. 1988. Conditions for identification in nonparametric and parametric models. *Econometrica* 56: 433–447.

Wooldridge, J. 1992. Nonparametric regression tests based on an infinite dimensional least squares procedure. *Econometric Theory* 8: 435–451.

Yatchew, A.J. 1992. Nonparametric regression tests based on an infinite dimensional least squares procedure. *Econometric Theory* 8: 435–451.

Non-price Competition

K. J. Lancaster

In markets for any goods but those which are absolutely homogeneous in both reality and perception, there are many ways in which firms may compete, price being one, but only one of these. The others include advertising and other forms of increased selling effort, product differentiation, improvement in product quality, customer service, warranties and the like, and bundling of other goods or services without charge or at low prices.

In general, firms can be expected to make optimal use of all strategic variables together, choosing price in conjunction with non-price elements. However, there was a historical progression in the output of the typical firm from homogeneous commodities, in which price is the only criterion for the purchaser, to heterogeneous products amenable to other forms of competition. This was reflected in economic analysis and probably in the development of real business strategies, so that non-price competition was initially treated as an alternative to, rather than as cooperating with, price competition.

For this reason the term *non-price competition* has come to be used mainly to describe the specific situation in which, for some particular reason, variation in price is ruled out and competing with the non-price variables alone is either the only permissible mode of competition or the optimal strategy.

The Incentive for Non-price Competition

There are two main contexts in which firms will be constrained to compete by non-price methods alone. The first is when the price is fixed by regulation or by a binding cartel agreement, the second in a small group oligopoly when it is not in the strategic interest of any party to upset a fragile equilibrium that has been reached in price alone. Such competition requires perceived product heterogeneity in the group, even if it is only because the products are branded, labelled, or bundled with seller-specific services.

There is a considerable grey area of disguised price competition between overt competition in the money sale price and true non-price competition. Apart from such practices as secret discounts and under-the-counter rebates (obviously to be treated as price competition), firms might compete by offering more services in conjunction with the product or by improving its quality. While it might be argued that provision of these services is equivalent to a quality increase and thus to a reduction in quality-adjusted price, there may also be a very large element of product differentiation since the firms often compete by the kind of service as much as by its quantity.

Prior to the burst of de-regulation in the late 1970s and early 1980s, there were many industries in the United States in which regulation of price was accompanied by non-price competition and disguised price competition. Brokerage houses in the securities markets were constrained by fixed commission rates – they competed by offering a variety of informational and advisory services to clients, as well as by advertising and other selling activities. In the airline industry, also price regulated, it was found necessary to regulate such additional things as the exact size of steaks provided in airline meals in order to prevent obvious competition in product quality, leaving non-price competition to unmeasurable and thus uncontrollable service properties like cabin decor and the behaviour of employees, as well as advertising and selling effort.

Elimination of price competition does not necessarily result in non-price competition. If the industry is initially unregulated and in equilibrium, firms will have been able to choose optimal values for both price and non-price variables. Fixing prices at these levels will, of itself, create no incentive to increase non-price competition. Such an incentive will occur only if prices are regulated at a level different from the equilibrium or if the equilibrium changes due to a change in costs, market parameters, or the number of firms.

The typical picture of the use of non-price competition is that of the low cost firm in a group in which price is regulated or collusively set at a level to suit higher cost firms. The low cost firm's preferred strategy would be to compete by lowering its price (perhaps in conjunction with other moves): denied this possibility, it aggressively pursues non-price methods of competition. As the formal analysis below shows, this may not be the outcome in all cases. It might be noted in passing that aggressive cost cutting is not considered as competition in most of the economics literature because of the conventional assumption that all firms are always operating at the lowest cost for their chosen output – a point on which the business literature would differ.

Formal Analysis

Consider an initial situation in which all firms are in equilibrium with respect to both price and non-price variables. Some regulatory or collusive arrangement now imposes a price which represents a shift away from the equilibrium value for at least one firm. Consider such a firm, in particular one for which the regulated price is above its equilibrium value, and analyse the effect on the non-price variables of this increase in the price. Will the firm wish to increase the levels of its non-price variables?

The situation can be modelled relatively simply with a non-price variable z (this can be thought of as advertising or selling effort), which directly affects quantity sold at a given price, and which will be taken to be measured in dollars or the standard numeraire of the system. The behaviour of other firms is taken to be fixed. The firm's profit function then has the form

$$\pi(p,z) = pq(p,z) - C(q(p,z)) - z$$

The condition for the profit maximizing level of p is the standard one, that marginal revenue with z constant equals marginal production cost. The profit maximizing level of z is given by the following relationship:

$$q_z(p,z) = 1/(p - C'(q(p,z))$$

where q_z is the marginal increase is quantity sold per unit increase in z, and C' is the marginal production cost of the good.

Take an initial situation of full equilibrium with both p and z at profit maximizing levels p^*, z^*, and consider the effect on z of a move in p away from p^*, such as might occur under price regulation, where z is optimally adjusted to the change. The effect on z is found by varying p in the equilibrium condition above, which gives:

$$\frac{\mathrm{d}z}{\mathrm{d}p} = \frac{1 - C''q_p - (p - C')^2 q_{zp}}{C''q_z - (p - C')^3 q_{zz}} (p - C').$$

Some of the signs of the derivatives on the right hand are clear, such as $q_p < 0$ (downward sloping demand curve), $q_z > 0$ (the variable z increases sales) and we would normally assume $C' \leq 0$ (non-decreasing marginal cost) and $p > C'$ (some monopoly mark-up). If z was a physical input, $q_{zz} < 0$ (diminishing marginal effect of z on sales) should be a normal assumption, but z is measured in cost terms and economies of scale, common in advertising, could give $q_{zz} > 0$. It seems reasonable to suppose that $q_{zp} < 0$ and small (increasing the price does not affect the influence of z much, and certainly does not increase it), although that might be in dispute.

If there are no economies of scale in the non-price variable z, $\mathrm{d}z/\mathrm{d}p$ is positive so that a low cost firm in a regulated industry, for which price is constrained above the optimal level, will engage in non-price competition – that is will increase the variable *z*. However if there are economies of scale in *z* and marginal production cost is rising little or is constant or falling, the firm would actually cut back on its non-price effort if constrained to sell at above its optimum pricing level.

Differentiating the Product

Product differentiation is less a form of non-price competition than a means of reducing the impact of price changes by one firm on the sales of others and thus reducing the instability of price competition in an oligopoly situation. The more different are the products of the rival firms, the lower the cross effects between their markets with respect to *all* variables, non-price as well as price. However, product differentiation may also be used as a sophisticated form of hidden price competition, by making the product simultaneously different and of higher quality. A low cost producer may be able to provide higher quality at the same nominal price by carefully chosen product differentiation that circumvents agreements or regulations yet is recognized as higher quality by consumers.

A large multiproduct firm may use product differentiation as a pre-emptive strategy to forestall competition rather than engage in it. Consider the case of a monopolist in a market where the consumers are distributed uniformly and will buy from the nearest outlet. A legally protected monopolist might operate at a single location, particularly if there are scale economies. Suppose there are such economies, and a firm needs a market width of ten miles to break even at the current price. If the monopolist sets up outlets nineteen miles apart, no firm will find it worthwhile to enter between these outlets because it will have a market area of less than ten miles, thus pre-empting competition from other firms by differentiating its location. An analogous effect holds for differentiation in characteristics other than location.

Other forms of non-price competition may also be used pre-emptively. The possibility of using advertising this way, when it shows economies of scale, was recognized in the early work on barriers to entry.

Non-price Competition and Excess Profits

A recurring theme in the literature on non-price competition is whether positive profits accruing to the members of an oligopolistic group of firms, which can certainly be pushed down to zero by competitive price undercutting, can also be competed away by advertising or other non-price activities. This is a relevant question for a regulated industry without free entry (like airlines or stock brokers in the fully regulated era) or a cartel arrangement, especially if the objective of the price regulation (perhaps hidden) is to preserve profit levels.

There has been a traditional presumption that non-price competition was less potentially ferocious than price competition, in the sense that it would be less likely to dissipate potential completely excess profit. Formal models (like those of Stigler 1968, or Schmalensee 1986), show, however, that the outcome depends on the values of the various system parameters and can go either way. The perceived difference between price and non-price competition may be due to comparing classic cases of price competition in homogeneous product industries (gasoline before the era of octane ratings and additives) with events in differentiated product markets with inherently softer competition in all variables.

See Also

- ▶ Advertising
- ▶ Competition
- ▶ Hotelling, Harold (1895–1973)
- ▶ Product Differentiation
- ▶ Spatial Competition

Bibliography

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

Schmalensee, R. 1986. Advertising and market structure. In *New developments in the analysis of market structure*, ed. J.E. Stiglitz and G.F. Mathewson. Cambridge, MA: MIT Press.

Spence, A.M. 1977. Nonprice competition. *American Economic Review* 67: 255–259.

Stigler, G.J. 1968. Price and nonprice competition. *Journal of Political Economy* 76: 149–154.

Non-profit Organizations

Richard Steinberg and Burton A. Weisbrod

Abstract

Non-profit organizations are hybrids – private but with restricted ownership rights. This defining 'nondistribution constraint' reduces incentives to exploit underinformed customers and allows non-profits to depart from profit-maximizing behaviour, although costly enforcement of this constraint limits effectiveness. Non-profits' GDP share in the United States is about 30 per cent of the governmental non-defence share. Worldwide they employ about four per cent of the labour force. Non-profits receive public subsidies potentially justifiable by their provision of public goods. Sales of goods and services constitute the main source of non-profit revenues, but government grants and private donations are also important. Extensive research on the economic behaviour of non-profit, for-profit, and governmental organizations in mixed industries has disclosed systematic differences.

Keywords

Agency problems; Crowding out; Free-rider problem; Lindahl prices; Nondistribution constraint; Non-profit organizations; Price discrimination

JEL Classifications

H0

Variously termed voluntary, philanthropic and charitable, non-governmental, as well as non-profit, these organizations constitute a sizeable and growing share of economic activity. Non-profit organizations contribute some four per cent of GDP in the United States, up from three per cent in 1970 and two per cent during the Second World War, but their GDP share is about 40 per cent of that of government. In the social service sector where they predominate, non-profits account for some 20–25 per cent of outputs. There are at least 1.6 million non-profit organizations in the United States, a number that grew by 27 per cent during the decade 1994–004 alone. The United States is not alone in the prominence or growth of the non-profit sector. Figures gathered by the Johns Hopkins International Comparison Project reveal a 'global associational revolution', with paid and volunteer labour in non-profits involving an average of 4.4 per cent of the economically active population in the 35 countries studied, ranging from a high of 14 per cent in the Netherlands to a low of 0.4 per cent in Mexico.

Non-profits are a form of institutional hybrid, combining attributes of profit-maximizing firms with those of government. Their organization and control are exercised through private initiatives rather than through the political process, and they cannot levy taxes. But, like government, they are constrained from distributing any profit or surplus to managers – the 'non-distribution constraint' (NDC). Many non-profits are granted a variety of tax subsidies such as eligibility for tax-deductible donations, special postal rates and exemption from taxation of income, property and sales. The NDC implies that non-profits cannot sell ownership shares, and so they can pursue social objectives other than profit maximization without fearing a hostile takeover. The NDC also implies that non-profits must rely for capital on sources other than equity shares. Thus, non-profits have both advantages and disadvantages in relation to private firms, with which they compete in industries such as health care (hospitals, nursing homes, hospices), education, and the arts.

Non-profits have provided public-type services, similar to those of government, for centuries. Jews created communal soup kitchens for travellers and collective charity funds for the needy in the second century BCE. In 16th-century

England, private ‘philanthropies’ were engaged in such wide-ranging social services as schools, hospitals, non-toll roads, fire-fighting apparatus, public parks, bridges, dykes and causeways, drainage canals, docks, harbour cleaning, libraries, care of prisoners in jails, and charity to the poor. In short, non-profits supplied the gamut of non-military goods and services that we identify today as governmental responsibilities.

Recent economic theorizing about the role of non-profits has examined both the nature of demand and the source of supply. Research on demand has two strands, one emphasizing failures of private markets, the other emphasizing governmental failures. In markets where valued attributes of the product are hard for consumers to observe and not verifiable by third parties, profit-maximizing firms can exploit their informational advantage. Alone, this outcome is inefficient, but the inefficiency is reduced if consumers deal with non-profit organizations. Non-profits have less incentive to short-change consumers because there are no shareholders or managers who can lawfully profit from this act. Nursing homes, day care for children, blood banks, medical research, environmental protection, and organizations claiming to aid the needy illustrate industries in which consumer information problems are not left to the private market alone. It is difficult for a nursing home patient or family member to determine whether the supplier is providing ‘tender loving care’; it is difficult to determine whether a day care centre is providing the attention that parents expect; and it is sometimes difficult for a patient or even a physician to determine the quality of blood available for a transfusion. Non-profits are a major force in all these industries characterized by informational asymmetries.

The quality assurance provided by the non-profit label, however, is limited. Enforcement of the NDC is spotty, and it is difficult to prevent distribution of profits in non-financial forms. Even when the NDC is well enforced, non-profits may short-change some under-informed customers in order to cross-subsidize missions that are not popular enough to generate direct donations. On the other hand, the sorting of entrepreneurs and consumers across ownership sectors, and the religious affiliation of many non-profits, enhances their credibility. The occasional failure of all these mechanisms is revealed, for example, by the collapse in 1995 of the Foundation for New Era Philanthropy following the revelation that the founder was benefiting from an illegal ‘Ponzi’ scheme in which colleges and universities, religious charities and individual donors were assured that their donations to the Foundation would be matched by a secret donor, resulting in grants that would double their ‘investment’ in three months.

Non-profits are a response to government failures as well as private market failures. The quantity and quality of outputs financed by government represent political decisions. Rather than setting individualized tax shares (Lindahl prices) to equate marginal benefits, governments use generalized systems of taxation, and so few voters get the quantity of governmentally provided goods that they would like, given the price each person confronts in the form of tax rates. Those who prefer less output and lower taxes have little recourse, but high demanders, who prefer more services at the tax prices they pay, and those seeking an alternative type or quality of service, may turn to non-profits to supplement governmental provision. Demanders of high-quality education, for example, often send their children to private non-profit schools. Non-profit schools also accommodate diverse minority demands relating to religion or educational philosophy. Thus, it is understandable that the United States, with a population unusually diverse in religion, culture, and ethnicity, has an unusually large non-profit sector, and not only in education.

Less is known about the *supply* of non-profit services. Non-profits are often created by churches and fraternal organizations, although over their life cycle they may disassociate from their founders. Religious organizations sometimes create health, education and welfare organizations as a way of attracting new adherents, binding the faithful, and meeting the moral obligations of their faith. In addition, religious and fraternal associations form communities of repeated interaction between like-minded individuals that help overcome free-rider problems and the transactions costs involved in creating a new organization. Nonetheless, the non-profit form solves an agency problem between the organization's founders and later donors. Thus, the organization's founder rationally chooses the non-profit form if the value of donor-supported public goods exceeds the value of the option to receive a share of future profits. But the entrepreneur may also be a profit-motivated organizer who sees the patina of a non-profit organization as little more than access to the subsidies and donations non-profits receive, and weak governmental enforcement of the NDC as providing opportunities for reaping greater personal financial gains than would be possible by founding a for-profit firm.

At any point in time, the number of non-profits is also affected by inter-sectoral conversions (particularly in the hospital industry), mergers, and tax and regulatory policies. Very little work has been done on the life cycles of non-profit organizations, but there is some evidence that non-profits are slower than for-profits to enter, expand, exit and contract. A limited supply of socially oriented entrepreneurs, an organizational preference for selectivity over expansion, non-profit inattention or incompetence, a preference to hold 'excess capacity' in case of medical emergencies, a reluctance to lay off employees, or differential capital constraints could explain these findings.

The non-profit form is far from a panacea; it, too, can fail. Revenue challenges abound. Non-profits cannot solve the free-rider problem because they cannot compel payments. The NDC encourages donations, but it also eliminates equity sales as a source of finance. Moreover, while the NDC reduces non-profits' incentives to take advantage of their patrons, it also reduces incentives for productive efficiency and responsiveness to changing market demand.

Non-profits rely on a mixture of revenue sources, varying greatly across industries, but sales of goods and services – especially tuition at colleges, patient fees at hospitals, and admission fees at museums, zoos and theatres – together with government grants and private donations are the three predominant sources. Research on donations has been extensive, examining the returns to fundraising expenditures and the efficacy of various fundraising mechanisms, as well as the effect of the charitable income tax deduction for private donors, the extent of crowding-out (or in) of private donations by direct government expenditures or by governmental service contracts with non-profit organizations, and the determinants of time donations (volunteering). Laboratory and field experiments on fundraising techniques are proliferating, revealing, for example, the positive effects of raffle mechanisms and information disclosures on net funds raised. Studies of donations' crowd-out – the effect on donations of an exogenous change in other revenues – run the gamut from 'near 100 per cent' to 'small but significant' to 'crowding in' (that is, negative crowd-out).

Determining the full effects of any fundraising mechanism is complex. Government grants, for example, are not only a source of revenue but may also certify quality to other donors; private donations may be affected differently for persons who derive utility from their own giving to a non-profit – the 'warm-glow', private-good effect – and for those who do not, being indifferent between their own contribution and the same amount given by someone else; and giving of money and volunteering of time are still not clearly identified as complements or substitutes.

Sales of goods and services are the dominant overall source of non-profits' revenue, and they come in diverse forms. First, many non-profits sell services that constitute their charitable mission rather than simply generating revenue, as in tuition charged by non-profit schools or payments for health services by non-profit hospitals, whether paid by the consumer or some third party. This is often accompanied by price discrimination designed to generate revenue when doing so does not compromise the tax-exempt mission, while providing the service at low cost – even free – to the poor or other 'deserving' consumers. A second form of programme service revenue has become increasingly prominent around the world – governmental purchase-of-service contracts with non-profits for the delivery of social services. Finally, some non-profits derive income from sales of goods and service that are unrelated to their charitable purpose, denoted 'unrelated business' (UB) income. Thus, non-profit universities have become major sellers of computer software; non-profit hospitals have opened pharmacies, hotels, and fitness centres; and non-profit museums' gift shops have become major purveyors of art objects. Controversy surrounds the social-welfare impact of having tax-privileged non-profits competing with taxable for-profits in commercial markets – the 'unfair competition' issue. In addition, analysts disagree about the impact of UB commercial activity on the social missions of non-profits, as they do about interpretation of the fact that half of all non-profits engaged in UB activities report no profit at all.

The impact of each revenue source on non-profits' social mission remains an area of controversy. Conceptually, the links reflect the non-profit's need to satisfy the wants of whoever is providing revenue – consumers, corporations, governments, alumni, and so on – and the consequences of doing so for the non-profit mission. With that mission typically being quite general, there is concern about 'mission creep' – the mutating definition of mission so as to justify taking advantage of a new revenue source. All revenue sources pose this potential problem, but it is particularly acute for collaborative ventures between non-profit and for-profit organizations. The issue often arises between research universities and firms in the pharmaceutical and information sciences fields, but similar issues arise, for example, in non-profit hospital relationships with for-profit medical groups, and in food pantries' relationships with food manufacturers.

Financing non-profits involves more than monetary flows. A major resource for the non-profit sector is volunteer labour – another form of donation. Of trivial importance in the for-profit sector, volunteer labour is, in the United States, similar in value to the total amount of money donations, although controversy persists as to how such labour, with an explicit transaction price of zero, should be valued for various purposes – replacement cost to the organization, opportunity cost to the volunteer, and the average market wage being the three prominent alternatives. Not counted in official labour force statistics, and generally overlooked as a contributor to output, volunteer labour in the United States equals about five per cent of the hours worked by the entire national labour force. Research on the supply of volunteer time indicates that it is affected by the same type of price, income and income tax rate variables that affect the supply of money donations. The identification of the separate effects of volunteer supply and organization demand, however, remains largely unstudied. The mission of non-profits may be more conducive to the use of volunteers than the profit-oriented goal of private firms.

There is some evidence that even paid labour in non-profit organizations is partially volunteered, that is, workers accept a lower salary, in effect donating some of the opportunity cost of their time. However, differences between wages in the non-profit sector and in other sectors appear to be specific to particular industries and job titles.

Hundreds of studies compare the performance of non-profit organizations with similar organizations in other sectors, but severe methodological challenges remain. Reviewing the vast evidence on health care organizations with respect to economic performance, quality of care, and accessibility to unprofitable patients, Schlesinger and Gray (2006) note that some authors conclude there are no clear differences. However, they dispute this interpretation, arguing instead that the literature is persuasive that there are clear differences, but the extent and direction of such differences depend on the nature of the service provided, market conditions, and external constraints on behaviour.

Non-profit organizations sometimes convert to for-profit and vice versa, especially in three industries – hospitals, health maintenance organizations (HMOs) and higher education. When non-profits convert, they first sell their assets to the new for-profit entity, using the proceeds to create or support non-profit organizations with closely related charitable purposes. Controversy surrounds conversions because of the difficulty of establishing a fair market value for these assets, particularly in leveraged conversions by insiders. If the assets are sold too cheaply, the new owners receive windfall profits and the NDC is violated.

Both theory and quantitative evidence suggest that all forms of institutions – non-profits included – fail to be efficient or equitable under particular circumstances. The key public policy questions are: do non-profits behave in systematically different ways from proprietary organizations or governments? If so, under what conditions and in which realms of economic activity should each form be encouraged, mandated, discouraged or prohibited?

See Also

- ▶ Agency Problems
- ▶ Altruism in Experiments
- ▶ Charitable Giving
- ▶ Health Economics
- ▶ Non-governmental Organizations
- ▶ Public Goods

Bibliography

Ben-Ner, A., and T. Van Hoomissen. 1991. Non-profit organizations in the mixed economy: A demand and supply analysis. *Annals of Public and Cooperative Economics* 62: 519–550.

Bilodeau, M., and A. Slivinski. 1998. Rational non-profit entrepreneurship. *Journal of Economics and Management Strategy* 7: 551–571.

Bilodeau, M., and R. Steinberg. 2006. Donative non-profit organizations. In *Handbook on the economics of giving, altruism, and reciprocity*, ed. S.C. Kolm and J.M. Ythier, vol. 2. Amsterdam: Elsevier.

Boris, E.T., and C.E. Steuerle. 2006. Scope and dimensions of the non-profit sector. In Powell and Steinberg (2006).

Chakravarty, S., M. Gaynor, S. Klepper, and W. Vogt. 2006. Does the profit motive make Jack nimble? Ownership form and the evolution of the US hospital industry. *Health Economics* 15: 345–361.

Clotfelter, C. 1985. *Federal tax policy and charitable giving*. Chicago: University of Chicago Press.

Duncan, B. 1999. Modeling charitable contributions of time and money. *Journal of Public Economics* 72: 213–242.

Goddeeris, J., and B. Weisbrod. 2006. Why not for-profit? Conversions and public policy. In *Government and non-profit organizations: The challenges of civil society*, ed. E.T. Boris and C.E. Steuerle, Revised ed. Washington, DC: Urban Institute.

Hansmann, H. 1980. The role of non-profit enterprise. *Yale Law Journal* 89: 835–898.

Hirth, R.M. 1999. Consumer information and competition between non-profit and for-profit nursing homes. *Journal of Health Economics* 18: 219–240.

James, E. 1983. How non-profits grow: A model. *Journal of Policy Analysis and Management* 2: 350–365.

James, E. 1986. The non-profit sector in comparative perspective. In *The non-profit sector: A research handbook*, ed. W. Powell. New Haven: Yale University Press.

Krashinsky, M. 1986. Transactions cost and a look at the non-profit organization. In *The economics of non-profit institutions*, ed. S. Rose-Ackerman. New York: Oxford University Press.

Landry, C.E., A. Lange, J.A. List, M.K. Price, and N.G. Rupp. 2006. Toward an understanding of the economics of charity: Evidence from a field experiment. *Quarterly Journal of Economics* 121: 747–782.

Leete, L. 2006. Work in the non-profit sector. In Powell and Steinberg (2006).

National Research Council. 2005. *Beyond the market: Designing nonmarket accounts for the United States.* Panel to Study the Design of Nonmarket Accounts, K.G. Abraham and C. Mackie, eds., Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academies Press.

Newhouse, J. 1973. Toward a theory of non-profit institutions: An economic model of a hospital. *American Economic Review* 60: 64–73.

Pauly, M., and M. Redisch. 1973. The not-for-profit hospital as a physician's cooperative. *American Economic Review* 63: 87–99.

Payne, A. 2001. Measuring the effect of federal research funding on private donations at research universities: Is federal research funding more than a substitute for private donations? *International Tax and Public Finance* 8: 731–751.

Powell, W.W., and R. Steinberg. 2006. *The non-profit sector: A research handbook*. 2nd ed. New Haven: Yale University Press.

Ribar, D.C., and M.O. Wilhelm. 2002. Altruistic and joy-of-giving motivations in charitable behavior. *Journal of Political Economy* 110: 425–457.

Robbins, K.C. 2006. The non-profit sector in historical perspective: Traditions of philanthropy in the west. In Powell and Steinberg (2006).

Roomkin, M., and B.A. Weisbrod. 1999. Managerial compensation and incentives in for-profit and non-profit hospitals. *Journal of Law, Economics and Organizations* 15: 750–781.

Rose-Ackerman, S. 1982. Charitable giving and 'excessive' fundraising. *Quarterly Journal of Economics* 97: 195–212.

Rose-Ackerman, S. 1986. *The economics of non-profit institutions*. New York: Oxford University Press.

Salamon, L.M., S.W. Sokolowski, and R. List. 2003. *Global civil society: An overview.* Baltimore: Johns Hopkins Center for Civil Society Studies.

Schlesinger, M., and B.H. Gray. 2006. Non-profit organizations and health care: Some paradoxes of persistent scrutiny. In Powell and Steinberg (2006).

Steinberg, R. 1986. Should donors care about fundraising? In *The non-profit sector: Economic theory and public policy*, ed. S. Rose-Ackerman. New York: Oxford University Press.

Steinberg, R. 2006. Economic theories of non-profit organizations. In Powell and Steinberg (2006).

Tuckman, H.P., and C.F. Chang. 2006. Commercial activity, technological change, and non-profit mission. In Powell and Steinberg (2006).

Vesterlund, L. 2006. Why do people give? In Powell and Steinberg (2006).

Weisbrod, B.A. 1977. *The voluntary non-profit sector.* Lexington: D.C. Heath.

Weisbrod, B.A. 1988. *The non-profit economy.* Cambridge, MA: Harvard University Press.

Weisbrod, B.A. 1998a. *To profit or not to profit: The commercial transformation of the non-profit sector.* Cambridge: Cambridge University Press.

Weisbrod, B.A. 1998b. Institutional form and organizational behavior. In *Private action and the public good*, ed. W.W. Powell and E.S. Clemens. New Haven: Yale University Press.

Weisbrod, B.A., and N. Dominguez. 1986. Demand for collective goods in private non-profit markets: Can fundraising expenditures help overcome free-rider behavior? *Journal of Public Economics* 30: 83–96.

White, M. 1981. *Non-profit firms in a three sector economy.* Washington, DC: Urban Institute.

Young, D. 1983. *If not for profit, for what?* Lexington: D.C. Heath.

Non-standard Analysis

Peter A. Loeb and Salim Rashid

JEL Classifications
C0

Non-standard analysis is an area of mathematics that provides a natural framework for the discussion of infinite economies. It is more suitable in many ways than Lebesgue measure theory as a source of models for large but finite economies since the sets of traders in such models are infinite sets which can be manipulated as though they were finite sets. The number system used to describe non-standard economies is an extension of the real numbers R; it is denoted by $^{*}R$. The set $^{*}R$ contains 'infinite natural numbers' and their multiplicative inverses, which are positive infinitesimals. It was with the development in 1960 of such a number system that Abraham Robinson (1974) solved an age-old problem by making rigorous the use of infinitesimals in mathematical analysis. Robinson gave a model-theoretic approach to his theory that is relevant to any infinite mathematical structure; that approach starts by listing the basic properties of the new number system. Before taking up this approach, however, it will be helpful to consider a simple nonstandard extension of the real numbers system that is constructed from sequences of real numbers.

The real numbers can be embedded in the set of sequences by associating a constant sequence $\{c_i\}$ with each real number c so that $c_i = c$ for all i. The relation on the set of sequences defined by setting $\{r_i\} > \{s_i\}$ if $r_i > s_i$for an infinite number of indices i has the property that if $r_i = i$ and $s_i = 1/i$ for all i, then $\{r_i\} > c$ and $c > \{s_i\}$ for any positive real number c. Here $\{r_i\}$represents a positive infinite number and $\{s_i\}$represents a positive infinitesimal. The relation $>$ is not yet an ordering on a number system since, for example, if $t_i = 0$ when i is even and $t_i = 3$ when i is odd, then $\{t_i\} > 2$ and $1 > \{t_i\}$. To fix the situation, one forms an equivalence relation in the set of sequences. The above sequence $\{t_i\}$for example should either be equivalent to the constant sequence 0 or the constant sequence 3.

An equivalence relation appropriate to the formation of a simple non-standard model of the real numbers from the set of real sequences is obtained by fixing a free ultrafilter U in the natural numbers N (i.e. a collection of subsets of N such that finite intersections of sets in U are in U, but the empty set and singleton sets are not in U, and if a subset of N is not in U, its complement is in U). Two sequences of real numbers$\{r_i\}$and $\{s_i\}$are equivalent if $r_i = s_i$ for all i in an element A of U. In this case, we say that $r_i = s_i$ almost everywhere. The equivalence classes form the nonstandard real numbers $^{*}R$. As before, the constant sequence $r_i = c$ represents the standard real number c, while the sequence $r_i = i$ represents an infinite element of $^*R$ and the sequence $r_i = 1/i$ represents a nonzero infinitesimal. In general, a property holds for elements of $^*R$ if for representing sequences it holds on some set in *U;* one says that the property holds almost everywhere, or 'a.e.'. An element r of $^*R$ is finite if for some standard c in R, $|r_i|$ is smaller than c a.e., and r is infinitesimal if for every positive c in R, $|r_i|$ is smaller than c a.e.; the elements of $^*R$ that are not finite are called infinite.

Properties true for the real numbers are again true for$^*R$, but quantification over sets must be interpreted as quantification over 'internal' subsets of $^*R$. For our simple model, these subsets correspond to equivalence classes of sequences of subsets of *R;* the element of $^*R$ represented by$\{r_i\}$ is in the set represented by$\{A_i\}$if and only if r_i is in A_i a.e. Not all subsets of $^*R$ are internal. Those that are not are called external. Some internal sets, called hyperfinite sets, have all of the formal properties of finite sets. Such a set A is represented by a sequence of subsets A_i from R with A_i finite a.e. The 'internal' cardinality of A is represented by the sequence $\{\mathrm{Card}(A_i)\}$ with 0 replacing infinite cardinals in the sequence. Thus, for example, if $A_i = \{1,2,\ldots, i\}$, then A is the set of all non-standard natural numbers less than or equal to the infinite natural number γ where γ is represented by the sequence $\langle 1, 2, 3, \ldots \rangle$. The internal cardinality of A, which we denote by $|A|$ is in this case equal to γ.

In working with non-standard analysis, it is usually best to ignore any particular construction of non-standard models and think only of the properties they satisfy. For general applications, one starts with a set theoretic structure *V(S)* where S is a set containing R and *V(S)* consists of all the sets one can obtain from S in a finite number of steps using the usual operations of set theory. For example, the number 5 and the set of all Lebesgue measurable sets is in *V(S)*; so is the set of all Borel measures on R. Let L be a formal language for *V(S)*; L contains a name for each object in *V(S)*, variables, connectives (such as the symbols $\vee$ for 'and' and $\wedge$ for 'or'), quantifiers, brackets, and sentences formed with these symbols. The main result of Robinson's theory establishes the existence of a (not unique) structure $V(^*S)$ built from a set of individuals $^*S$ with the following properties; (1) Every name of an object in $V(S)$ names something of the same type (i.e. constructed with exactly the same operations) in $V(^*S)$ We write $^*A$ for the object in $\mathrm{V}(^*S)$ with the same name as A in $V(S)$; A is called standard and$^*A$, the (nonstandard) extension of A. (2) (Transfer Principle) Every sentence in L that is true for *V(S)* is true when interpreted in $V(^*S)$; quantification, however, is over 'internal' objects in $V(^*S)$. (3) If $A \in V(S)$ is a set, then there is a 'hyperfinite' set B which is a member of the extension$^*P_{\mathrm{F}}(A)$ of the set of all finite subsets of A such that for each $a \in A$; $^*a \in B$. Thus B contains the extension of each standard element of A.

The extension$^*s$of an individual s is usually denoted by s instead of $^*s$*;* one thinks of a subset A of S as being imbedded in $^*A$. Internal objects in $V(^*S)$ are those objects which are members of the extensions of standard objects; the non-internal objects in $V(^*S)$ are called external. Any object that can be described in the formal language L using the names of internal objects is itself internal. The illuminating fact that the set N of finite natural numbers forms an external set in the non-standard natural numbers $^*N$ can be established as follows: If N were an internal set, then by applying the transfer principle to the theorem that every non-empty subset of N has a first element, it would follow that there is a first infinite element of $^*N$, that is, a first element of $^*N–N$, and thus a last element of N.

Hyperfinite sets are internal sets in internal one-to-one correspondence with an initial segment of $^*N$. Such sets are useful in economics because they are treated like finite sets. To illustrate this fact and Property 3 above, we note that if γ is an infinite element of $^*N$, then the initial segment $T = \{n \in {}^*N : 1 \leq n \leq \gamma\}$ of $^*N$ is a hyperfinite set containing every element of N. Every formal property true for an initial segment of the natural numbers is true for the set T, whence T can be used as the set of traders in a 'hyperfinite economy'.

The set of non-standard real numbers $^*R$ contains infinite elements which are positive, infinite elements which are negative, and finite elements. Any finite element α of $^*R$ is infinitely close to a unique standard real number $a \in R$ that is, $\alpha - a$ is infinitesimal. We write $\alpha \approx a$ and also $\alpha - a \approx 0$ in this case; a is called the standard part of α. The standard part of α is denoted by st (α) or $^0\alpha$. The set of all points infinitely close to a is called the monad of a. A subset 0 of R is open if and only if for each point x in 0, the monad of x is contained in ^{*0}.

As an application of these ideas, we note that a real-valued sequence s_n (i.e. a mapping s from N into R) has a limit l if and only if for each infinite n, $* s_n \approx L$ where $^*s_n$ is the image of n with respect to the non-standard extension $^*s$ of the function s. A real-valued function f defined on a subset A of R is continuous at a point $x \in A$ if and only if for all $y \in {}^*A$ with $y \approx x$, $^*f(y) \approx f(x)$; is uniformly continuous on A if and only if for all $y \in {}^*A$ and $Z \in {}^*A$ with $y \approx z, {}^*f(y) \in {}^*f(y)\,(z)$. A subset A of R is compact if and only if for each $y \in {}^*A$ there is a standard x in A with $y \approx x$, whence A is compact if and only if it is closed and bounded. It is immediate that if f is continuous on a compact set A then f is uniformly continuous on A.

Brown and Robinson (1975) introduced the use of non-standard analysis in economics as a source of models for infinite exchange economies. The set of traders in non-standard economies is a hyperfinite set $T = \{1,2,\ldots, \gamma\}$ where γ is an infinite element of N^*. The preferences and endowments are internal mappings defined on T analogous to the corresponding mappings in finite economies. Each trader's commodity endowment is an infinitesimal part of the market, and so that trader's influence on the formation of prices is infinitesimal but not zero. One can show in such economies, even without the usual convexity assumptions, that approximate competitive equilibria and approximate cores exist and that these cores can be approximately decentralized by the price system.

Given a hyperfinite set T, such as the set of traders in a non-standard economy, one can apply to T all of the combinatorial methods that are available for finite sets. For example Loeb (1973) obtained a form of the Lyapunov convexity theorem that is appropriate for the hyperfinite economies described above by applying a 'packing theorem' concerning a finite set of vectors in Euclidean space. Using another construction of P.A. Loeb (1975) one can form on T a standard measure space which is rich with structures inherited from the underlying point set. This construction starts by noting that the set M of all internal subsets of T forms an algebra in the usual sense. One obtains a finitely additive probability measure P on (T, M) by setting $P\,(A)$ equal to the standard part of the ratio $|A|/|T|$ for each A in M. One may assume that any ordinary sequence $\{A_i : i \in N\}$ from M is the initial segment of an internal sequence $\{A_i : i \in {}^*N\}$ from M. This will be the case, for example, if the superstructure is constructed via an ultrafilter as indicated above. Now, if an ordinary sequence $\{A_i : i \in N\}$ from M is pairwise disjoint and $\cup A_i$ equals some element A in M, then all but a finite number of the A_i's are empty. (Extend $\{A_i : i \in N\}$ to $\{A_i : i \in {}^*N\}$ for every infinite $n \in {}^*N$ and therefore for some finite $n \in N$ A is contained in the union of the A_i's, $1 \leq i \leq n$.) The condition one checks to apply the Carathéodory extension theorem to (T, M, P) is thus vacuously satisfied. Therefore P has a σ-additive extension μ to the smallest σ-algebra $\sigma(M)$ generated by M. The space $(T, \sigma(M), \mu)$ is a standard probability space which is very close in structure to the internal hyperfinite space (T, M, P).

Rashid (1979) first established the connection between the standard measure spaces that exist on hyperfinite economies and the models of infinite economies using Lebesgue measure. Measure spaces on hyperfinite economies have the great advantage of an underlying structure that closely parallels finite economies. This parallelism has been exploited by H.J. Keisler in his forthcoming work detailing the price adjustment processes in nonstandard exchange economies. Emmons (1984) has obtained results for economies using measure spaces on hyperfinite sets of traders that are not available for general measure space economies. Nonstandard economies also have the advantage of making readily apparent regularities in the asymptotic behaviour of large but finite economies. Anderson's (1978) core equivalence theorem, for example, was obtained by translating a result originally proved with nonstandard analysis. Similarly, a translation of Khan and Rashid (1982) produced the existence theorem of Anderson et al. (1982).

A further advantage inherent in the use of the number system $^*R$ in economics is the ability to distinguish behaviour on the finite part of $^*R$ from that on the infinite part and to distinguish different orders of infinities and infinitesimals. The first type of distinction was used by K.D. Stroyan (1983) to provide an elegant non-standard characterization of myopia in the evaluation of infinite consumption streams. The second distinction was central in Brown and Loeb's (1976) short, non-standard proof of Aumann's theorem showing that the Shapley value of infinite economies under appropriate differentiability conditions coincides with the competitive equilibria.

See Also

- ▶ Existence of General Equilibrium
- ▶ Large Economies
- ▶ Lyapunov Functions
- ▶ Measure Theory
- ▶ Perfect Competition

Bibliography

Ali Khan, M., and S. Rashid. 1982. Approximate equilibria in markets with indivisible commodities. *Journal of Economic Theory* 28 (1): 82–101.

Anderson, R.M. 1978. An elementary core equivalence theorem. *Econometrica* 46 (6): 1483–1487.

Anderson, R.M., M. Ali Khan, and S. Rashid. 1982. Approximate equilibria with bounds independent of preferences. *Review of Economic Studies* 49: 473–475.

Brown, D.J., and P.A. Loeb. 1976. The values of nonstandard exchange economies. *Israel Journal of Mathematics* 25 (12): 71–86.

Brown, D.J., and A. Robinson. 1975. Nonstandard exchange economies. *Econometrica* 43 (1): 41–55.

Emmons, D.W. 1984. Existence of Lindahl equilibria in measure theoretic economies without ordered preferences. *Journal of Economic Theory* 34 (2): 342–359.

Loeb, P.A. 1973. A combinatorial analog of Lyapunov's theorem for infinitesimally generated atomic vector measures. *Proceedings of the American Mathematical Society* 39 (3): 585–586.

Loeb, P.A. 1975. Conversion from nonstandard to standard measure spaces and applications in probability theory. *Transactions of the American Mathematical Society* 211: 113–122.

Rashid, S. 1979. The relationship between measure-theoretic and non-standard exchange economies. *Journal of Mathematical Economics* 6 (2): 195–202.

Robinson, A. 1974. *Non-standard analysis*. Rev ed. Amsterdam: North-Holland.

Stroyan, K.D. 1983. Myopic utility functions on sequential economies. *Journal of Mathematical Economics* 11 (3): 267–276.

Non-substitution Theorems

Neri Salvadori

Keywords

Joint production; Non-substitution theorem; Substitution

JEL Classifications

D0

A non-substitution theorem asserts that under certain specified conditions an economy will have one particular price structure for each admissible value of the profit rate, regardless of the pattern of final demand. The theorem has two forms. As first stated, it applies to an economy with single production and therefore no fixed capital (Arrow 1951; Koopmans 1951; Samuelson 1951; Levhari 1965). In its later formulation, some special joint products are considered to take account of fixed capital (Samuelson 1961; Mirrlees 1969; Stiglitz 1970).

Consider first the single production form. The non-substitution theorem asserts that if (i) there exists one primary input (call it labour); (ii) all processes of production are perfectly divisible, with constant returns to scale, and have the same production period (this period is taken as the time unit for the analysis); (iii) each process produces one perfectly divisible commodity, making use of definite amounts of produced commodities and, perhaps, perfectly divisible labour; (iv) for each commodity there exists at least one process producing it; (v) labour is indispensable for the reproduction of commodities; (vi) the exchange of commodities takes place at the end of each period in fully competitive markets (that is the profit rate, the wage rate, and the price of each commodity are uniform); (vii) producers operate a process if and only if it is cost-reducing at current prices; then for each admissible value of the profit rate only one vector of relative prices (including the wage rate) is possible for the economy, so that relative prices are independent of demand.

In order to understand why the theorem works, let us denote with vector *p* and scalar *w* the equilibrium commodity price vector and the equilibrium wage rate, respectively, when the rate of profit equals *r* and the net output vector equals vector *d*. Hence, (a) no process is able to pay extra profit at prices *p,* wage rate w, profit rate *r;* (b) for each commodity there exists at least one operable process producing it (an operable processes is a process whose costs, including normal profits, are not larger than the price of the product); (c) operable processes can be operated in such a way to produce net output *d.*

The non-substitution theorem asserts that

(α) if the net output is $\widehat{d} \neq d$ then *p* and *w* are still an equilibrium price-vector and an equilibrium wage rate;

(β) if more than one solution exists, they are characterized by the fact that price vectors and wage rates are respectively equal each other (if the same *numéraire* is utilized).

To prove statement (α) we just need to prove that operable processes can be operated in such a way to produce net output $\widehat{d}$, since statements (a) and (b) hold. This is shown in the following way. Take one operable process for each commodity (they exist because of (b)) to arrange the material input matrix A and the labour input vector 1. It is easily shown that matrix (I – A) is invertible and $(\mathrm{I}-\mathrm{A})^{-1} \geq 0$ where I is the identity matrix of appropriate size. Hence, statement (α) is a consequence of the fact that if the operation intensities of these processes are $\widehat{d}^T(\mathrm{I}-\mathrm{A})^{-1} \geq 0$ all the others being zero, then the net output vector equals $\widehat{d}$. This procedure can also be utilized if a uniform growth rate not larger than *r* is assumed.

To prove statement (*β*) let (p_1, w_1) and (p_2, w_2) be the price vector and the wage vector relative to two equilibrium solutions, respectively. Similarly as in the proof of (α) we can arrange material input matrices A_1 and A_2 and labour input vectors 1_1 and 1_2 from the first and the second solution respectively. Hence,

$$p_1 = (1+r)A_1p_1 + w_1l_1$$
$$p_2 = (1+r)A_2p_2 + w_2l_2.$$

Moreover, axiom (vii) requires that

$$p_1 \leqq (1+r)A_2p_1 + w_1l_2$$
$$p_2 \leqq (1+r)A_1p_2 + w_2l_1$$

and since $[I-(1+r)A_\mathrm{i}]$ is invertible and $[I-(1+r)A_i]^{-1} \geq 0$ $(i = 1{,}2)$,

$$p_1 \leqq w_1[I-(1+r)A_2]^{-1l_2} = (w_1/w_2)p_2$$
$$p_2 \leqq w_1[I-(1+r)A_1]^{-1l_1} = (w_2/w_1)p_1.$$

Thus,

$$p_1 = (w_1/w_2)p_2.$$

Then, by introducing the common numeraire we obtain that

$$w_1 = w_2$$

and, therefore,

$$p_1 = p_2.$$

More recent formulations of the non-substitution theorem weaken the previously stated assumption (iii) to allow the introduction of some particular joint production cases. Assumptions are introduced to divide commodities into ‘final goods’ and ‘used machines’. Each process is assumed to produce one final good, but some joint products are allowed since used machines are produced jointly with final goods. Used machines are not transferable, that is an oven utilized once to produce bread cannot be utilized later to produce biscuits.

If machines are not used jointly, then a non-substitution theorem is stated as in the single production form. If machines can be used jointly, then the growth rate plays a role in determining prices and the wage rate, as does the profit rate. This fact has been recognized by Stiglitz (1970), who, however, failed to recognize that when this is so the uniqueness of the relative prices and wage rate does not need to hold even if prices are still independent of demand.

The label ‘non-substitution’ is appropriate to these theorems in so far as it assumed that there is only one scarce factor (primary input). Relaxation of any or all of the other assumptions, for example that of constant returns to scale, will mean that prices vary in response to changes in the structure of demand, but *will not* mean that there is ‘substitution’ in any meaningful sense.

In a neoclassical model prices are determined by the relation between demands (direct and derived) for endowment and the magnitude of the components of the endowment (typically conceived as stocks of factor services). The prices of produced commodities are equal to their costs of production, that is to the sum of rentals paid for the factor services used in their production. The possibility of *substitution* between factors, due either to substitution between commodities consumed or substitution in production, or to a combination of both, is the source of variation in derived demand, and hence in relative rentals. If, by comparison with a given situation, preferences were different, relative demands for factor services would typically be different, and hence their rentals and the prices of the commodities in the production of which they are used would be different.

But, if there is only one factor of production, no substitution is possible whatever the composition of demand or the range of technical possibilities. Hence the relative prices of produced commodities will be determined by the least amounts of the single factor by means of which (directly and indirectly) they are produced. If, as is the case in the examples discussed above, the minimum cost technique is invariant to changes in demand, then prices too will not change as demand changes. If, however, the minimum cost technique does change as demand changes, say because of increasing returns to scale, then prices will change, *but this will not be due to any substitution.* There cannot be any substitution because there is only one factor. Similarly, in those cases of joint production in which a change in the structure of demand does lead to a change in relative prices, the change derives not from substitution between factors but from a change in the minimum cost combination of production processes.

See Also

▶ Input–Output Analysis

Bibliography

Arrow, K.J. 1951. Alternative proof of the substitution theorem for Leontief models in the general case. In *Activity analysis of production and allocation*, ed. T.C. Koopmans. New York: Wiley.

Koopmans, T.C. 1951. Alternative proof of the substitution theorem for Leontief models in the case of three industries. In *Activity analysis of production and allocation*, ed. T.C. Koopmans. New York: Wiley.

Levhari, D. 1965. A non-substitution theorem and switching of techniques. *Quarterly Journal of Economics* 79: 98–105.

Mirrlees, J.A. 1969. The dynamic non-substitution theorem. *Review of Economic Studies* 36: 67–76.

Samuelson, P.A. 1951. Abstract of a theorem concerning substitutability in open Leontief models. In *Activity analysis of production and allocation*, ed. T.C. Koopmans. New York: Wiley.

Samuelson, P.A. 1961. A new theorem on non-substitution. In *Money growth and methodology.* Lund: CWK Gleerup.

Stiglitz, J.E. 1970. Non-substitution theorems with durable capital goods. *Review of Economic Studies* 37: 543–552.

Non-Tariff Barriers

John Beghin

Abstract

Non-tariff barriers (NTBs) refer to the wide range of policy interventions other than border tariffs that affect trade of goods, services and factors of production. Most taxonomies of NTBs include market-specific trade and domestic policies affecting trade in that market. Extended taxonomies include macroeconomic policies affecting trade. NTBs have gained importance as tariff levels have been reduced worldwide. Common measures of NTBs include tariff equivalents of the NTB policy(ies), and count and frequency measures of NTBs. These NTB measures are subsequently used in various trade models, including gravity equations, to assess trade and/or welfare effects of the measured NTBs.

Keywords

Antidumping; Border effects; Countertrade; Domestic content requirements; Gravity equations; Nontariff barriers; Price control; Protection; Quantity control; Tariff versus quota; Tariff-rate quotas; Tariffs; Technical barriers to trade; Trade costs

JEL Classifications

F

Nontariff barriers (NTBs) refer to the wide and heterogeneous range of policy interventions other than border tariffs that affect and distort trade of goods, services and factors of production. Common taxonomies of NTBs include market-specific trade and domestic policies such as import quotas, voluntary export restraints, restrictive state-trading interventions, export subsidies, countervailing duties, technical barriers to trade (TBTs), sanitary and phytosanitary (SPS) policies, rules of origin and domestic content requirements schemes. Extended taxonomies also include macro-policies affecting trade. No taxonomy can be complete since NTBs are defined as what they are not (Deardorff and Stern 1998). This article is complemented by related articles on antidumping, border effects, countertrade, gravity equation, tariff versus quota, and trade costs. Deardorff and Stern (1998) suggest the following taxonomy with five categories.

A first broad category covers quantitative NTBs and similar restrictions. It includes import quotas and their administration methods (licensing, auctions, and other); export limitations and bans; voluntary export restraints, a limit on imports but managed by exporters; foreign exchange controls often based on licensing; prohibitions such as embargoes; domestic content and mixing requirements forcing the use of local components in a final product; discriminatory preferential trading agreements and rules of origin; and countertrade such as barter and payments in kind.

A second category covers fees other than tariffs, and associated policies affecting imports. This category includes variable levies triggered once prices reach a threshold or target level; advanced deposit requirements on imports, antidumping and countervailing duties imposed on landing goods allegedly exported ‘below cost’ or with the help of export subsidies provided by foreign governments; and border tax adjustment such as value-added taxes potentially imposed asymmetrically on imported and domestic competing goods.

A third category is extensive. It collects various forms of government policies including a wide set of macroeconomic policies. This category covers direct governmental participation and restrictive practices in trade, such as state-trading and state-sponsored monopoly and monopsony; government procurement polices with domestic preferences; and industrial policy favouring domestic firms with associated subsidies and aids. In addition, the category extends to macroeconomic and foreign exchange policies, competition policies, foreign direct investment policies, national taxation and social security policies, and immigration policies. Where to draw on the NTB definition is context-dependent.

Two better-targeted categories deal with customs procedure and administrative practices, and technical barriers to trade, which are central to NTBs. The former covers custom valuation methods that may depart from the actual import valuation; customs classification procedures other than the international harmonized system of classification to levy further fees; and customs clearance procedures such as inspections and documentation creating trading cost. Technical barriers to trade relate to health, sanitary, animal welfare and environmental regulations; quality standards; safety and industrial standards; packaging and labelling regulations and other media/advertising regulations. With the exception of export subsidies and quotas, NTBs have become more prominent than tariffs. Tariffs on manufacturing goods have been reduced to low levels through eight successive rounds of the World Trade Organization (WTO) and its predecessor, the General Agreement on Tariffs and Trade (GATT). As of 2005, the unweighted average tariff is roughly three per cent in high-income countries, and 11 per cent in developing countries according to the World Bank, from respective levels at least three times as high in 1980. Exports subsidies have almost disappeared except in a few agri-food markets. Quotas have become less important since they have been converted into two-tier tariff schemes, the so-called tariff-rate quotas. As tariffs have been lowered, demands for protectionism have induced new NTBs, such as TBT interventions. The United Nations Conference on Trade and Development (UNCTAD) estimates that the use of NTBs based on quantity and price controls and finance measures has decreased dramatically from slightly less than 45 per cent of tariff lines faced by NTBs in 1994 to 15 per cent in 2004, reflecting commitments made during the last round of WTO negotiations, the Uruguay Round. However, the use of NTBs other than quantity and price controls and finance measures increased from 55 per cent of all NTB measures in 1994 to 85 per cent in 2004. The use of TBTs almost doubled, from 32 to 59 per cent of affected tariff lines during the same period. The use of quantity control measures associated with TBTs showed a small increase, from 21 to 24 per cent of affected tariff lines, suggesting that trade impediments within TBTs are rising. Kee, Nicita and Olarreaga compute a 9 per cent tariff equivalent of NTBs including price and quantity controls, finance measures, and TBTs, on average for all goods. The average tariff equivalent is about 40 per cent for the goods affected by these NTBs.

Increased consumer demand for safety and environment-friendly attributes have also translated into an increase in the number of TBTs. Many NTBs are regulated by the WTO agreements that came out of the Uruguay Round (the TBT Agreement, SPS Measures Agreement, the Agreement on Textiles and Clothing), and articles of the original GATT among others. NTBs in service industries have recently become more important as trade in services has been expanding (Dee and Ferrantino 2005).

Most NTBs are intrinsically protectionist whenever they do not address market failures such as externalities and information asymmetries between consumers and producers of goods being traded. Safety standards and labelling requirements are examples of the latter case. Some NTBs may restrict trade but improve welfare in the presence of negative externalities or informational asymmetries. Other NTBs can expand trade as they enhance demand and trade of a good through better information about the good or by enhancing the good's characteristics. Whether an NTB is protectionist is sometimes difficult to identify in the presence of market failure. If an NTB is equal to the measure that a social planner would implement for domestic purposes (that is, all firms are domestic firms or all agents belong to a single economy), that NTB is presumably non-protectionist (Fisher and Serra 2000).

Measuring NTBs and their effects is a challenge, because of the heterogeneity of policy instruments and lack of systematic data. A unified approach to the measuring of NTBs does not exist. Most measurement methods start from a simple partial equilibrium approach looking at a single commodity, and attempt to develop a producer, consumer or trade tax equivalent to the NTBs, that explains by how much supply, and/or demand, or trade are affected by the policy intervention. Most NTB analyses implicitly rely on a framework that accounts for three economic effects: the regulatory protection effect providing rents to the domestic sector; the 'supply shift' effect, that reflects the increased costs of enforcing compliance of the NTBs on foreign and sometime domestic suppliers; and the 'demand-shift' effect, that takes into account the fact that a regulation may enhance demand with new information or by reducing an externality.

The measurement of an NTB is hard to disentangle from the measurements of its effects on market equilibrium and trade. Most NTB measures and analyses focus on the increase in the price of imports resulting from the NTB, the resulting import reduction, the change in the price responsiveness of the demand for imports, the variability of the effects of the NTB, and the welfare cost of the NTB (Deardorff and Stern 1998; Dee and Ferrantino 2005).

Several NTBs based on a price intervention (for example, export subsidies, countervailing duties), are a tax instrument. More complex NTBs can sometimes be represented by a set of taxes, such as in the case of a domestic content requirement (Vousden 1990). These NTBs can be analysed as such taxes. To develop a tax equivalent, a basis of equivalence has to be chosen (Vousden 1990). The tax equivalent has to lead to either an equivalent protection level (same profit under the tax equivalent or the NTB), or to a price increase equivalence (a price wedge), or to consumption, production or trade equivalent. This choice of basis depends on the intended policy analysis.

However, many NTBs do not easily translate into a tax-equivalent instrument. They require more sophisticated and indirect approaches to be measured and to quantify their effects on import volume, price, and welfare. Roundabout approaches are also used because of lack of data on the direct implications of an NTB on cost of production and consumer decisions (Beghin and Bureau 2001).

Common Measurement Approaches of NTBs

The price-wedge method measures the impact of an NTB on the domestic price of a good in comparison to a reference price, often the border price of a comparable good. The aim of this method is to derive a tariff/tax equivalent to the NTB as discussed above, and use the tariff/tax equivalent in further analysis that measures implications of the NTB on resource allocation in the given markets affected by the NTB. Deardorff and Stern (1998) provide price-wedge equivalent formulas for an extended coverage of NTBs.

Conceptually, the measure compares the domestic price that would prevail without the NTB to the domestic price prevailing in the presence of the NTB, on the assumption that the price paid to suppliers remains unchanged. However, these prices are practically unobservable. Implementations of the price-wedge measure of an NTB compare the domestic and foreign prices of comparable goods in the presence of the NTB accounting for tariffs, transportation costs, and other known and observed trading costs. Adjustments can be made to recover a price estimate that would prevail in the absence of the NTB, using observed levels of quantities and prices, and own-price elasticities of demand and supply and imported goods.

The price-wedge method has several drawbacks. First, if several NTBs are jointly in place, the price-wedge measures the price effect of these policies without being informative about their respective contributions or even their nature. Second, quality differences are hard to account for precisely although they are a pivotal element of the price-wedge computation. The price-wedge estimate of an NTB is usually sensitive to the assumptions made on the substitution between the imported and domestic goods. This method has also some limitations in large empirical studies for which data are aggregated, resulting in loss of information on quality differences between import and domestic comparable goods. Finally, trading costs may be present but not accounted for and the price-wedge method may falsely attribute these trading costs to a NTB.

Inventory-based frequency measures count the number or frequency of regulations and barriers present in a given market. They are used in both quantitative and qualitative assessments of the incidence of the NTBs. Common measures include the number of regulations and policies, which can be further elaborated to indicators such as the number of pages of national regulations. Frequency of trade detentions at borders is also used, and so are survey-based frequency and number of complaints reported by exporters for perceived discriminatory regulatory practices.

When implemented, quantitative estimates often rely on catalogues of technical barriers (identification and description) using datasets such as UNCTAD’s TRAINS data-set. Measures include simple frequency of occurrence of NTBs, frequency ratios for product categories subject to an NTB, and coverage ratio based on the value of imports of products within a category subject to the NTB, expressed as a share of import value of the corresponding category. Relative measures can also be developed comparing the latter frequency measures in a given country with respect to accepted international norms or best practice, for example, for the SPS or food safety regulations. Alternatively, frequency measures can be compared across commodities or across countries to identify large deviations from average frequencies, flagging potential protectionist issues.

NTBs vary in importance across sectors and products. Even for a given NTB type, its effects may vary across products. A major drawback of the frequency measures is that a correlation between the number of NTBs and their effect on trade and welfare may be low in absolute value. International data-sets on NTBs inventories may also suffer from uneven reporting by countries and heterogeneous coverage of measures across countries and commodities. Survey-based measures focus on effective barriers rather than just an NTBs count. However, they may suffer from various reporting biases as surveys and respondents are often motivated by mercantilism to facilitate exports by the responding exporters.

Frequency measures do not identify the trade restrictiveness of NTBs but can be used in gravity equations to identify the effects of NTBs on trade flows. When trying to quantify NTBs, an obvious technique is to consider the forgone trade that cannot be explained by tariffs and known trading costs. The NTB frequency measures, or in certain cases the level of standards themselves, can help identify the trade effects of these NTBs. Provided there is enough variability across countries or over time in the measure (for example, the level of toxic residues) they can explain the variation in trade flow not explained by other explanatory variables included in the gravity equation (respective incomes of trading countries, distance, tariff, and other variables measuring border effects).

Gravity-equation techniques attempt to measure the trade impact of NTBs, not their welfare impact, and may therefore ignore some of the beneficial effect of the regulations that correct negative externalities but restrict trade. NTBs are appropriate if trade is the vector of negative externalities such as unsafe food imports or pest-infested imports. In addition, the direction of the effect of the 'NTB' variable on trade flows in the regression is not constrained. It is possible to capture a trade or demand-enhancing effect of regulations and standards. This enhancement occurs when the NTB facilitates trade and induces consumers to consume more of a product although the product's price is higher because of the NTB. Such expansion through standards has been observed in OECD food trade (Disdier et al. 2006).

Risk assessment approaches and scientific knowledge can contribute to gauging a subset of NTBs, especially safety and SPS standards and regulations. The latter approach can contribute to assessing the welfare effects and the potential protectionism of these types of NTBs. Scientific knowledge can determine if a regulation is science-based or not, or if a risk simply does not exist or is negligible. This criterion is used by the WTO in its assessment of TBT and SPS regulations. Cost–benefit calculations combined with risk assessment provide expected cost and benefits of such types of NTBs. Risk-assessment measures provide an economic criterion to gauge the desirability of an NTB and its likely protectionist nature if externalities are small and if its costs greatly exceed its benefits in expected terms. The combined use of scientific knowledge and cost–benefit assessment of an NTB is a demanding process suitable for a detailed analysis of a specific case study, rather than for large-scale multi-market analyses. Another limitation of this approach is the partial knowledge of health, environmental and other risks associated with trade and their economic significance.

NTBs measures are an essential step to computing the welfare effects of the NTBs. Beyond welfare effects, these measures are also useful for policy purposes. WTO disputes frequently arise, alleging that some NTBs impede trade more than necessary to achieve some legitimate objective, or that they are just protectionist. These NTB measures are used in the formal dispute process to estimate export market losses and price-lowering effects of the incriminated policy.

See Also

► Anti-dumping
► Border Effects

- ► Countertrade
- ► Gravity Equation
- ► Tariff versus Quota
- ► Trade Costs

Bibliography

Beghin, J., and J.-C. Bureau. 2001. Quantitative policy analysis of sanitary, phytosanitary and technical barriers to trade. *Economie Internationale* 87: 107–130.

Deardorff, A.V., and R.M. Stern. 1998. *Measurement of nontariff barriers: Studies in international economics*. Ann Arbor, MI: University of Michigan Press.

Dee, P., and M. Ferrantino. 2005. *Quantitative methods for assessing the effects of non-tariff measures and trade facilitation*. Singapore: APEC Secretariat and World Scientific.

Disdier, A.-C., Fontagné, L., and Minouni, M. 2006. The impact of regulations on agricultural trade: evidence from SPS and TBT agreements. Working paper. Paris: CEPII

Fisher, R., and P. Serra. 2000. Standards and protection. *Journal of International Economics* 52: 377–400.

Henson, S., and J.S. Wilson. 2005. *The WTO and technical barriers to trade*, Critical Perspectives on the Global Trading System and the WTO Series. Northampton, MA: Edward Elgar.

Kee, H.L., Nicita, A., and Olarreaga, M. 2006. Estimating trade restrictiveness indices. Policy Research Working Paper No. 3840. Washington, DC: World Bank.

UNCTAD (United Nations Conference on Trade and Development). 2005. Methodologies, classifications, quantification and development impacts of non-tariff barriers. Note by the UNCTAD secretariat, document TD/B/COM.1/EM.27/2, 23 June. Geneva.

Vousden, N. 1990. *The economics of trade protection*. Cambridge: Cambridge University Press.

North American Free Trade Agreement (NAFTA)

Gordon H. Hanson

Abstract

The North American Free Trade Agreement (NAFTA) eliminated trade barriers on most products between Canada, Mexico, and the United States. NAFTA included provisions to remove restrictions on cross-border investment, expand service trade, and address environmental and labour standards. Post-NAFTA increases in trade between member countries were matched by comparable decreases in their trade with the rest of the world. Freer trade has brought a shift in economic activity within Mexico and the United States towards their shared border and an increase in direct investment from the United States to Mexico. In Mexico these developments have contributed to greater wage inequality.

Keywords

Canada–United States Free Trade Agreement; Computable general equilibrium models; Foreign direct investment; International migration; North American Free Trade Agreement (NAFTA); Regional and preferential trade agreements; Rules of origin; Tariffs; Trade diversion; Trade policy, political economy of; World Trade Organization

JEL Classifications

F1

The North American Free Trade Agreement (NAFTA), which entered into force in 1994, eliminated trade barriers on most products between Canada, Mexico, and the United States. The agreement culminated a decade of liberalization in North America, which included the Canada–United States free trade agreement in 1989 and Mexico's joining the General Agreement on Trade and Tariffs (GATT) in 1986 (On the impact of the Canada–United States Free Trade Agreement, see Trefler 2005). For Mexico, NAFTA was the final step in reversing four decades of protectionist trade policies. For Canada and the United States, NAFTA completed three decades of promoting closer economic ties.

Upon its implementation, NAFTA eliminated tariffs on goods accounting for approximately one-half of trade between the three countries. Tariffs on other goods (primarily those with relatively high pre-NAFTA tariffs) were phased out over 5–15-year periods. Among the industries with the slowest tariff phase-outs were Canadian textiles, Mexican corn and US sugar. A few industries (mainly in agriculture, energy and services) were excluded from NAFTA altogether. The agreement incorporates stringent rules of origin, which apply a country's external tariff to NAFTA imports whose North American content (in terms of the share of value added) fails to meet mandated thresholds (Estevadeordal and Suominen 2005). Rules of origin prevent three-way trade in which, say, Canada imports a good from Japan at an external tariff that is below that for Mexico and then re-exports the good to Mexico at a zero NAFTA tariff. If allowed, such trade would effectively impose a common external tariff across North America equal to the minimum tariff for each good among the three countries. Content requirements vary across sectors, with those for the auto industry being among the highest.

NAFTA was broad in its scope and included provisions for removing restrictions on cross-border investment between member countries, expanding service trade and protecting intellectual property. A novel feature of the agreement was the adoption of side accords for environmental and labour standards, which created a mechanism under which citizens of member countries can adjudicate disputes over the violation of standards (which in their essence state that NAFTA members are obliged to uphold environmental and labour laws that each has on its books). While the standards were controversial at the time of NAFTA's passage, few cases of significance involving environmental or labour infractions have been resolved under the agreement.

The economic rationale for creating a regional free trade area is that it eliminates price distortions caused by tariffs, quotas and other policy barriers, which induce countries to allocate too many resources to import-competing industries and too few resources to exporting industries. In the early 1990s, results from computable general equilibrium models suggested that NAFTA would raise welfare by an amount equal to between two and four per cent of GDP in Mexico and one per cent or less of GDP in Canada and the United States (Brown et al. 1992). Low estimated gains from trade associated with NAFTA are not surprising, given that prior to the agreement Canada and the United States already had a free trade agreement in place, Canadian and US external tariffs on most products were already quite low, and Mexico had begun to unilaterally liberalize its trade following its joining the GATT.

While a free trade area creates trade between member countries, it also diverts trade between the trade bloc and the rest of the world. Between 1993 and 2004, trade between Canada, Mexico and the United States increased by 2.6 times in real terms; over the same period, trade between NAFTA countries and the rest of the world increased by only 1.9 times (Hufbauer and Schott 2005). In sectors that had the highest protection prior to NAFTA, nearly all of the increase in trade within the NAFTA region was matched by comparable decreases in trade between NAFTA members and the rest of the world (Romalis 2005), consistent with the agreement causing trade diversion.

Even where the net change in income associated with freer trade is small, gross changes in income for particular groups may be large. Because trade agreements redistribute income, they tend to provoke political conflict. The politics surrounding NAFTA were perhaps most contentious in the United States. President Clinton's support for NAFTA became an issue in the 1996 US presidential campaign, with opposition candidate Ross Perot memorably claiming that increased trade with Mexico would create a 'giant sucking sound' as US jobs moved across its southern border.

In the United States, one would expect groups allied with labour to oppose freer trade with a low-wage country and groups allied with capital-intensive industries to support it. NAFTA was narrowly approved by the US Congress, with its outcome uncertain until the final hour. Consistent with standard models of political economy, US politicians receiving donations from labour groups tended to vote against NAFTA, while those receiving donations from business groups tended to vote for NAFTA (Baldwin and Magee 2000). There was also a regional dimension to NAFTA's politics, with US politicians representing districts near the US border with Mexico being much more likely to support the agreement. This in part reflects the fact that, as US trade with Mexico has expanded, US border states have seen their manufacturing and trade-related industries grow relative to the rest of the country (Hanson 2001). In Mexico, also, NAFTA has had a varied regional impact. Following Mexico's opening to trade, Mexican states near the US border have had high growth in manufacturing employment, exports, and foreign direct investment (FDI) relative to the rest of the country. The shift in economic activity towards Mexico's border region has increased regional income differences in the country, which had been declining until Mexico began to liberalize trade (Chiquiar 2005).

Economic theory suggests that trade may either complement or substitute for factor flows, depending on the magnitude of transport costs, fixed production costs and cross-country differences in technology and factor supplies. Following NAFTA, there has been a substantial increase in FDI by the United States in Mexico. Much of the FDI has involved US multinational firms setting up export assembly plants, known as *maquiladoras*, in Mexico. FDI in assembly plants has resulted from US firms outsourcing production to Mexico and has created substantial intra-industry trade flows in which the US exports parts and components to Mexico, and Mexico exports finished goods back to the United States. Similar trade patterns have existed between the United States and Canada since the 1960s, when the two countries liberalized trade in the auto industry. By moving labour-intensive production activities out of the United States, NAFTA has decreased the relative demand for less skilled labour in the country. And by moving capital, technology and new production operations into Mexico, NAFTA has increased the relative demand for skilled labour in Mexico. Thus, United States–Mexico economic integration appears to have contributed to a widening of the skilled–unskilled wage gap in both countries (Feenstra and Hanson 1997).

At the time of NAFTA's signing, the agreement was touted as a means of reducing United States–Mexico wage differences and the incentive for workers in Mexico to migrate to the United States. By the 1980s, Mexico–United States migration had become an important policy issue on both sides of the border. NAFTA was justified in part as a way to reduce international migration flows. However, since NAFTA's implementation there has been an increase rather than a decrease in the flow of labour from Mexico to the United States (Hanson 2006). At least some of the increased migration appears associated with the collapse of the peso in 1994 and the ensuing economic contraction in Mexico (Hanson and Spilimbergo 1999). Partly as a result of the peso collapse, the difference in per capita income between the United States and Mexico was larger in 2002 than in 1990 (Tornell et al. 2003). Other evidence suggests that, whatever its long-run effects, NAFTA may have contributed to a transitory increase in Mexico-to-United States migration. By contributing to gross job destruction in agriculture and manufacturing, NAFTA may have displaced workers who then migrated to the United States.

For Canada and Mexico, NAFTA helped increase the importance of the US economy for their economic development. For the United States, NAFTA was a milestone in the country's approach to trade policy. Since 1994, the United States has concluded bilateral trade agreements with a dozen other countries, but has not succeeded in helping complete a multilateral trade agreement under the auspice of the World Trade Organization. One interpretation of this pattern is that NAFTA signalled a shift in US trade policy away from multilateralism and toward bilateralism, perhaps weakening multilateral trade institutions in the process.

See Also

- ▶ Banking Crises
- ▶ Computation of General Equilibria
- ▶ Factor Prices in General Equilibrium
- ▶ Foreign Direct Investment
- ▶ New Economic Geography
- ▶ Regional and Preferential Trade Agreements
- ▶ Supply Chains

Bibliography

Baldwin, R., and C. Magee. 2000. Is trade policy for sale? Congressional voting on recent trade bills. *Public Choice* 105: 79–101.

Brown, D., A. Deardorff, and R. Stern. 1992. North American integration. *Economic Journal* 102: 1507–1518.

Chiquiar, D. 2005. Why Mexico's regional income convergence broke down. *Journal of Development Economics* 77: 257–275.

Estevadeordal, A., and K. Suominen. 2005. Rules of origin in preferential trading arrangements: Is all well with the spaghetti bowl in the Americas? *Economia* 5: 63–69.

Feenstra, R., and G. Hanson. 1997. Foreign direct investment and relative wages: Evidence from Mexico's maquiladoras. *Journal of International Economics* 42: 371–394.

Hanson, G. 2001. U.S.–Mexico integration and regional economies: Evidence from border-city pairs. *Journal of Urban Economics* 50: 259–287.

Hanson, G. 2006. Illegal migration from Mexico to the United States. *Journal of Economic Literature* 44: 869–924.

Hanson, G., and A. Spilimbergo. 1999. Illegal immigration, border enforcement and relative wages: Evidence from apprehensions at the U.S.–Mexico border. *American Economic Review* 89: 1337–1357.

Hufbauer, G., and J. Schott. 2005. *NAFTA revisited: Achievements and challenges*. Washington, DC: Institute for International Economics.

Romalis, J. 2005. *NAFTA's and CUSFTA's impact on international trade*, Working paper no. 11059. Cambridge, MA: NBER.

Tornell, A., F. Westermann, and L. Martinez. 2003. Liberalization, growth, and financial crises: Lessons from Mexico and the developing world. *Brookings Papers on Economic Activity* 2003(2): 1–88.

Trefler, D. 2005. The long and the short of the Canada-U.S. Free Trade Agreement. *American Economic Review* 94: 870–895.

North, Douglass Cecil (Born 1920)

Avner Greif

Abstract

A pioneer of the New Institutional Economics, Douglass North has built upon the property rights and transaction cost approaches of Coase and others to explain economic growth in terms not of changes in technology and productive factors but of institutional and organizational change. His most recent work stresses the need to integrate insights from cognitive science into the examination of the interplay among belief systems, institutions, and economic performance. Institutions reduce the uncertainties that would otherwise overwhelm cognitive capacity in complex social situations, but the resulting bias in our beliefs can lead to the persistence of inefficient institutions.

Keywords

Alchian, A.; Cheung, S.; Cliometric Society; Cliometrics; Coase, R.; Commons, J.; Demsetz, H.; Feudalism; Fogel, R.; Growth and institutions; Hayek, F.; Industrial revolution; Institutionalism, old; Kuznets, S.; Mitchell, W.; New economic history; New institutional economics; North, D.; Patents; Peasants; Property rights; Social norms; Society for the New Institutional Economics; Stagnation; State, theory of; Tax transaction costs; Technical change; West, the

JEL Classifications

B31

A native of Cambridge, Massachusetts, who was born in 1920, North received his undergraduate and doctoral degrees from the University of California, Berkeley. His 1952 Ph.D. dissertation focused on the history of the American insurance industry. Most of his professional career has been spent at two institutions: the University of Washington in Seattle and, from 1983, Washington University in Saint Louis. North was among the founders of cliometrics (also known as the New Economic History). Later he was a pioneering researcher of the New Institutional Economics. In 1993 North and Robert Fogel shared the Nobel Prize in Economics. In 1997–8 he served as the first president of the Society for the New Institutional Economics.

North's publications are numerous and space limitation precludes presenting all of them or even doing justice to particular ones. Accordingly, the following discussion of his books highlights some of North's main contributions.

Neoclassical Analyses

North initially studied American economic growth. Under the influence of Simon Kuznets, he compiled the first quantitative historical series of the US balance of payments. This work, in conjunction with his studies of regional development, led to his first published book in 1961, *The Economic Growth of the United States, 1790–1860*. In this book North developed an export-based growth model to argue that the expansion of one sector (cotton plantations) in the United States stimulated development in other sectors, and led to specialization and interregional trade.

By relying on economic theory and quantitative analysis, this line of work contributed to the rise of cliometrics (or the New Economic History). In contrast to traditional economic historians who relied on narratives and non-qualitative analysis, cliometrics combines economic theory, quantitative methods, hypothesis testing, counterfactual analysis, and traditional techniques of historical analysis to explain economic outcomes, evaluate and develop economic theories, and deepen our historical knowledge. North further fostered this development by helping to found the Cliometric Society and serving as co-editor of the *Journal of Economic History* for five years.

Towards Institutional Analysis

In the late 1960s North began expanding his analysis of economic growth beyond the confines of neoclassical economics by considering the importance of organizational changes to increasing efficiency. In his 1968 article on productivity in overseas shipping, North argued that organizational changes had more important effects than technological changes in reducing transportation costs between 1600 and 1850. Market integration and growth followed due to organizational rather than technological changes.

More generally, North began to emphasize that, in order to understand growth, one had to go beyond the neoclassical framework, which at that time attributed growth to changes in technology and factors of production. In sharp contrast, North argued that changes in technology and factors of production are not the sources of growth but, in fact, constitute growth. This implies that, to understand growth, we must examine the forces that cause beneficial technological changes and increase the utilization of factors of production. North argued that institutions constitute such forces and his subsequent research focused on them.

In developing the analysis of the relationship between institutions and economic growth, North built on and expanded the property rights and transaction costs approaches advanced by Ronald H. Coase, Armen A. Alchian, Steven N.S. Cheung, Harold Demsetz, and others. His subsequent books were ambitious attempts to place institutions at the centre of economic growth analysis. Good institutions promote growth by bringing private return from economic activities closer to their social return. Economic growth transpires in response to low-cost enforcement of contracts when property rights are secured and when governments pursue growth-oriented policies rather than prey on the wealth of their subjects. Institutions that achieve these goals encourage technological innovations, foster capital accumulation, and increase labour input. Growth follows as technology improves, capital accumulates, and specialization occurs.

Institutions in the Northian framework consist of rules and regulations which, together with their enforcement mechanisms, determine the incentives faced by economic agents. Similar focus on the relationships between institutions and economic outcomes has also been the hallmark of old institutionalism (associated with such scholars as John R. Commons, Friedrich A. von. Hayek, and Wesley C. Mitchell). Old institutionalism, however, considered institutions either as exogenous and immutable or as reflecting spontaneous, uncontrolled processes. In contrast, North attempted to consider institutions as endogenous and to understand the forces that shaped their development. To do this, he particularly concentrated on the state as setting the rules of the economic game.

Institutions and American Growth

North's first book on this issue, *Institutional Change and American Economic Growth,* was co-authored with Lance Davis and published in 1971. It outlines a theoretical perspective on the role and dynamic of institutions. The main theoretical assertion is that new institutions – specifically, new property-rights assignments – arise when groups in the society perceive that there are opportunities for profit that cannot be consummated given the existing institutions, but that would be feasible if these institutions were changed. Perceptions of benefits of institutional change and the details of the political system are what determine whether socially beneficial institutional change will transpire.

The book demonstrates the merits of this assertion by considering growth in the United States during the 19th century. It advances a new interpretation of American economic growth as one that reflects the pursuit of profit opportunities by economic agents through changing politically determined rules. Commodity markets expanded, for example, because canals reduced transportation – and hence transaction – costs. Investments in canals, however, didn't occur automatically. Public investment, state- mandated changes in property rights, and changes in perceptions of the profitability of these investments were prerequisites. Similarly, political decisions and changes in property rights were crucial to other factors that directly contributed to growth: the evolution of capital markets, the rise of large corporations and of the manufacturing sector, investment in human capital and the expansion of service industries.

Institutional evolution was central to American economic growth. More generally, North's work illustrates that in order to understand economic growth, the evolution of laws and regulations governing property rights must first be analyzed. Changes in property rights are often required before individuals and societies can gain from increasing the scale and efficiency of production and exchange.

Institutions and the Rise of the West

A subsequent book (co-authored with Robert Paul Thomas) published in 1973, *The Rise of the Western World: A New Economic History*, further applied these ideas to explain the performance of various western European economies. By examining economic outcomes from the feudal period to the Industrial Revolution, the book sought answers to two questions. First, do differences in institutions account for patterns of economic growth and stagnation in European economies, and does the rise of the West reflect the efficiency of its property-rights regime? Second, what determines whether more or less efficient institutions will prevail?

The book argues that patterns of growth and stagnation in Europe reflect whether property rights were assigned efficiently and secured. The feudal system ended in economic stagnation and crises because of the misallocation of property rights to land. Peasants had few incentives to increase land productivity because they did not own it. Later, the Dutch Republic and England outpaced Spain and France because their property right's assignments were better designed to close the gap between private and social rates of return from economic activities. England's rising technological superiority, for example, reflected its effective system of patenting. In the long run, other European economies adopted similarly efficient systems of property rights.

Two forces determine institutions' relative efficiency. Institutions' degrees of efficiency respond to changes in relative prices, which, in turn, are due to changes in population and technology. As the relative price of a factor of production increases, property rights will be altered to better align incentives. The collapse of the state in medieval Europe rendered protection a valuable commodity. The feudal system, in which specialists in protection held property rights to land, reflected the relatively high value of protection. The large decline in the European population during the 14th century, however, increased the relative price of labour. This undermined the feudal system, and property rights in land were transferred to the peasants who toiled on it.

The tendency towards efficiency in institutional change is countered by the transaction costs of tax collection. Specifically, a ruler assigns property rights in a manner that maximizes his net revenue rather than efficiency. The transaction costs of tax collection place a wedge between efficient property right regimes and those that are optimal to a ruler. France's geographical scale and diversity, for example, implied that the taxation regime that was optimal to its rulers entailed a high efficiency cost. France's economic growth therefore lagged behind England's.

Given the importance of the state in this analysis, North advanced a theory of the state in his 1981 book, *Structure and Change in Economic History.* He departed from the common view of the state as an efficiency-enhancing social contract aimed at increasing security or providing other public goods, and characterized it as a ruler- predator, utilizing a bargaining framework, to consider a ruler's relationship with his subjects. In a state, citizens contract with a specialist in enforcement to provide them with protection. The terms of the deal – the extent of absolutism and predation – reflect the relative bargaining power of these parties, which, in turn, depends on such factors as military technology and the threat of entry by competing rulers. This analysis contributes to the argument that interstate competition within Europe was growth-enhancing by emphasizing that this competition may have constrained predation by rulers.

The 1981 book is also a departure from North's previous lines of analysis in that it focuses on ideology. North's previous writings noted the importance of informal institutions, such as ideology, social norms, and values, but they had not been explicitly integrated into the analysis. This book, however, claims that ideology develops as a justification for existing institutions and hence it is both endogenous to institutions and a strengthening factor. Although North's earlier analyses were rooted in history, they developed an ahistorical theory of institutions. These analyses sought a deterministic theory of institutions: a mapping from exogenous, contemporaneous conditions (such as population, technology, and geography) to institutions. Subsequently, North developed a more elaborate view of institutional change that attempted to capture how past institutions influence ensuing ones.

Recent Theoretical Developments

North's 1990 book, *Institutions, Institutional Change and Economic Performance,* develops an historical theory of institutional change. The argument revolves around the interplay between organizations and institutions. Institutions provide the incentives for establishing some organizations – for example, firms or political interest groups – but not others, and influence their activities. Through such activities, the organizations that institutions promote acquire new knowledge and information. This new knowledge enables them to recognize how they can improve their ability to advance their interests through institutional change. Therefore, these organizations act as players in the politics of setting the rules that govern economic interactions. Hence, institutional change is a path-dependent process. Institutions induce the emergence of particular organizations which later engage in institutional change. Such changes are incremental because organizations don't set out to destroy the institutions that gave rise to them. History matters.

Complementary forces that render institutional dynamics a historical process are the focus of North's 2005 book, *Understanding the Process of Economic Change.* More generally, the book emphasizes that economic stagnation emerges when and where institutions fail to adjust efficiently. The focus of the analysis is on the cognitive capabilities and limitations of individuals and how they influence institutional change. Institutions constructed by individuals reflect their understanding of reality and determine the growth of their understanding. Dissimilar initial cognitive views of reality can therefore lead societies to develop distinct institutions in the same objective situation. The different processes of individual and social learning that these initial institutions imply keep each society on a distinct institutional trajectory.

Hence, for example, the establishment of institutions in the Soviet Union was based on a particular concept of reality. Once established, however, these institutions led to particular learning processes as well as the emergence of organizations with vested interests in the institutions. The result was initial economic success followed by decades of decline because the initial concept of reality was wrong but the organizations it led to had an interest in maintaining the system.

While this book provides new answers to an important question, it more generally calls attention to the need to integrate insights from cognitive science into the examination of the interplay among belief systems, institutions, and economic performance. It particularly emphasizes the relevance of theories of connected or embedded cognition, which argue that human cognition is a social phenomenon shaped by man-made constructs. Institutions shape individual cognition by reducing the uncertainties that would otherwise overwhelm cognitive capacity in complex social situations. At the same time, the resulting bias in our beliefs about this environment can lead to a lock-in of these institutions.

See Also

- ▶ Austrian Economics: Recent Work
- ▶ Coase Theorem
- ▶ Economic History
- ▶ Growth and Institutions
- ▶ Institutionalism, Old
- ▶ Political Institutions, Economic Approaches to

Bibliography

1961. *The economic growth of the United States, 1790–1860.* Englewood Cliffs: Prentice-Hall.

1968. Sources of productivity change in ocean shipping, 1600–1850. *Journal of Political Economy* 76: 953–970.

1971. (With L. Davis.) *Institutional change and American economic growth.* Cambridge: Cambridge University Press.

1973. (With R.P. Thomas.) *The rise of the western world: A new economic history.* Cambridge: Cambridge University Press.

1981. *Structure and change in economic history.* New York: Norton.

1990. *Institutions, institutional change and economic performance.* Cambridge: Cambridge University Press.

2005. *Understanding the process of institutional change.* Princeton: Princeton University Press.

North, Dudley (1641–1691)

Douglas Vickers

Keywords

Free trade; Loanable funds; McCulloch, J. R.; Monetary circulation; Money supply; North, D.

JEL Classifications

B31

Sir Dudley North, knighted for his service as a sheriff of London in 1682, was born at Westminster in 1641, the third of five sons of the fourth Baron Guilford. He died at Covent Garden on the last day of December 1691. After a highly successful merchant career in the Levant, he returned to England in 1680 and was appointed a Commissioner of the Customs in 1683. He was promoted to Commissioner of the Treasury in 1685, and when that Commission was dissolved a few months later he returned to the Customs where he remained until the Revolution of 1688.

North's place in the history of economic theory is due to his essay *Discourses upon Trade,* published in 1691 (or early 1692). His clear-sighted advocacy of free trade principles, his opposition with John Locke, to the proposals advocated by Sir Thomas Culpeper and Sir Josiah Child for a legal maximum rate of interest, and his advanced views of the beneficial effects of monetary circulation make the *Discourses* a high-water mark in the pre-classical literature.

The *Discourses,* first published anonymously, were summarized in the biography of Sir Dudley published by his brother Roger in 1744. The Preface to the *Discourses*, the concluding paragraph of the second Discourse, and the final paragraph of the Postscript appear to be the work of Roger. The work was rediscovered and evaluated very highly by the classical economists and J.R. McCulloch published a reprint of the *Discourses* in 1822.

Applying a general supply and demand theory of prices to the determination of interest rates, North argued that a law to restrict the interest rate to a specified maximum level would be ineffective. The market rate of interest depended heavily on the availability of loanable funds which depended on the savings made out of income, a 'surplus' that provides an accumulation of investable 'stock'. A fourfold proposition follows. First, ' as more buyers than sellers raiseth the price of a commodity, so more borrowers than lenders will raise interest'. Second, 'as the landed man letts his land, so these still let their stock; ... thus to be a landlord or a stock-lord is the same thing'. Third, 'it is not low interest that makes trade, but trade increasing, the stock of the nation makes interest low.' Fourth, as the largest part of the demand for loanable funds was for consumption purposes (leading to a *prodigality* and thrift theory of interest, rather than one of *productivity* and thrift) 'an ease of interest will rather be a support to luxury than to trade'.

North argued that it was not so much that trade depended on money as that the money supply depended on trade. For 'nations which are very poor, have scarce any money, and in the beginnings of trade have often made use of something else, ... as wealth increased, gold and silver hath been introduced and drove out the other'. A money supply adequate to the needs of trade would be assured, moreover, by the 'ebbing and flowing of money', the coining, melting, and recoining of bullion. 'The buckets work alternately.' Emphasizing the significance of monetary expenditure and circulation, and not simply the money supply as such, complaints against a shortage of money were met by the argument that the remedy for a depressed economy was not 'the increase of specific money' but a disposition to spend rather than hoard.

'The nation ... never thrives better than when riches are tost from hand to hand.'

See Also

▶ Mercantilism

Selected Works

1691. *Discourses upon trade.* London. Ed. J.H. Hollander. Baltimore: Johns Hopkins University Reprint, 1907.

Bibliography

Letwin, W. 1963. *The origins of scientific economics: English economic thought 1660–1776*. London: Methuen.

McCulloch, J.R. 1824. *A discourse on the rise, progress, peculiar objects, and importance of political economy.* Edinburgh: Constable.

McCulloch, J.R. 1856. *Early English tracts on commerce*. London: Political Economy Club.

North, R. 1744. *The life of the Honourable Sir Dudley North ... and of Dr. John North, master of Trinity College*, ed. M. North. London.

Vickers, D. 1959. *Studies in the theory of money 1690–1776*. Philadelphia: Chilton.

N

North–South Economic Relations

Ravi Kanbur

North–South is the title of a book which became popularly known as 'The Brandt Report'. Published in 1980, the book had an immediate impact in terms of popular coverage and appeal. There was a conference of world leaders at Cancun to discuss the report, amidst great publicity. In 1983 there was a follow- up report which received less publicity, and in any case it can be argued that the ardour over North–South relations had cooled somewhat by then. Did the Brandt Report simply introduce a new phrase in international dialogue, or did its achievements go beyond that? In order to answer this question we have to take a step back and consider the nature of North–South economic relations.

Rather as ‘the Third World’ was used to signal the problems of nations that belonged neither to the developed countries of the West nor to the centrally planned economies of the East, ‘North–South’ is meant to signal divisions between rich nations and poor nations, in contradistinction to the ‘East–West’ divide. In fact, the position of the centrally planned economies in the North–South divide is ambiguous. The Brandt Report clearly wished to categorize them with the rich countries of the North, but the centrally planned economies have themselves rejected such a classification, preferring to see the poverty of the South as the result of the imperalist past of the western capitalist countries, with the attendant economic structures that are still argued to be in place today. These reservations on the part of the eastern bloc countries led to considerable discussion in the late 1970s, when a group of developing countries attempted to set an agenda for the achievement of what they termed a ‘New International Economic Order’.

Even leaving aside the issues raised by the existence of the centrally planned economies, the economic relations between North and South are complex and manifold. Trade is the most obvious form of economic interaction, but associated with trade are capital flows. In the latter category are private capital flows, including investment by multinational corporations, as well as official flows of aid. The official flows category can be further subdivided into bilateral aid and aid channelled through multilateral agencies. Associated with capital flows is the question of technology transfer and the question of repatriation of profits earned in the South back to the parent company in the North.

The simplest stylized model of North–South trade is one where the South specializes in the production of primary commodities while the North specializes in the production of manufactures. This ‘Argentina–England’ model has become less significant as many poor countries have diversified their output and their exports to include light manufactures (such as the often referred to success stories of Korea, Taiwan, Hong Kong and Singapore) or even heavy manufactures after a period of import substitution (such as India or Brazil). However, it would nevertheless be true to say that the primary commodities/manufactures divide is the one most analysts use as a framework for thinking about North–South relations. Hence the concern in the Brandt Report with the fluctuations of primary product prices. Schemes to stabilize these prices were given great emphasis before, during and after the period in which the Report came out. UNCTAD’s Integrated Program for Commodities was designed to be a major buffer stock scheme to stabilize the prices of several primary commodities. It was argued that demand fluctuations in the North impose a cost in terms of real income variability in the South, and the same was true of uncontrollable climatic factors in the South. Variability in one region was transmitted to the other through the channel of trade, and it was suggested that international cooperation was needed to overcome the costs of this particular aspect of North–South economic relations.

An even stronger claim is that the price of primary products relative to that of manufactures is on a downward secular trend. There is considerable debate regarding this ‘Prebisch-Singer’ hypothesis. One simple model of why there might be such a trend is that the demand for food is income inelastic while the demand for manufactures is income elastic. Thus with given supply conditions as world income grows the shift in demand in favour of manufactures raises their price. The problem with this argument is of course that it neglects supply conditions. Even within the framework adopted if the supply of food is relatively price elastic then the effects of a shift in demand will be mitigated. The question then turns on the elasticity of food supply, an issue which is complicated by the fact that many of the Northern countries (e.g. the USA) are major producers and exporters of food. A further complication is that food is only one component of primary commodity output. Other countries produce and export such natural products as rubber, copper and bauxite. Here it is technological innovations which are important in shifting demand away from the exports of less developed countries.

An alternative line of argument is sometimes used to theorize about the possible long-term decline in the terms of trade of the poor countries of the South. This is that while the production of primary commodities is undertaken primarily by peasant smallholders, the production of manufactures is in the hands of large oligopolistic corporations in the North who use their market power to resist downward movements in the price of manufactures relative to primary commodities. Kaldor (1976) makes such an argument. Given the myriad of factors influencing the North–South terms of trade, it is perhaps not surprising that the empirical results of testing for a secular decline against the South are by no means unequivocal. Spraos (1980) summarizes the debate, which will undoubtedly continue.

Private capital flows, particularly direct investment by Northern multinational corporations in the South, have been a topic of considerable controversy. Those in favour of such investment argue that the reason why such investment is good for the South is precisely the reason why such investment is considered profitable by the multinationals. The South is labour abundant and capital scarce. Wages are low and hence investment by multinationals is profitable. But such investment should be encouraged in a capital scarce economy, since this is a way of building up capital stock and hence raising wages. Those against multinational investment argue that the technology which is transferred to developing countries in this way has been developed in the context of developed countries and hence inappropriate to the conditions prevailing in the former. In particular, it is too capital-intensive relative to the employment creation needs of developing countries. Moreover, the types of product manufactured are inappropriate to the poor in developing countries, relying rather on the demands of the rich. The technology, apart from being capital-intensive, is skill-intensive and creates a class of highly paid workers in contrast to the mass of low-paid workers elsewhere. It is therefore argued that this particular channel of North–South economic relations relies on inequality within developing countries, and perpetuates it.

Other than capital flows for investment, there is also short-term debt that the developing countries have built up, particularly in adjusting to the two oil shocks of the 1970s, and the world recession of the 1980s. If the oil price rises and commodity booms of the 1970s played their part in bringing North–South economic relations in the forefront of debate, it is the ‘debt crisis’ which has played that role in the 1980s. The debt levels of less developed countries as a whole have reached historic highs, and the picture has been equally dramatic for particular countries such as Brazil and Mexico. More importantly, it is clear that many of the leading banks and financial institutions in the USA and in the West generally have allowed themselves to be exposed to risk of default. Given the interlocking nature of financial institutions and of the financial system in general, events in the South have taken on a new meaning for policy-makers in the North. In days gone by default by an entire nation could be and was taken care of by physical force. This is no longer possible, and sovereign default is a real possibility for developing countries, and a real worry for developed ones.

The solvency of a nation should be assessed by whether or not, over the long haul, it can service its debt out of growth in income. The real question then concerns not short-term liquidity but the long-term growth prospects for developing countries. But in the short term there seem to be considerable impediments to developing countries being able to export enough to service their debt. The deep recession in the West in the early 1980s meant that demand for their exports was low. Another consequence of unemployment problems in the OECD countries has been the growing demand for protectionist measures. The deep and abiding interactions between North and South are clearly highlighted in the debt crisis. If the North adopts protectionism then the South cannot export. But if the South cannot export it cannot service its debt, which will lead to default. A default on Southern debt spells disaster for the financial system of the North, and hence for output and employment in North and South. This short-term impasse, in which the global recession is certainly playing its part, has long-term consequences as investment falls and hence future potential output is curtailed. Yet another strand in this complex weave is added when one takes into account the effect of high interest rates in the North on the Southern debt burden.

However, it is as well as this stage to note that a detailed look at specific countries reveals a more varied picture of North–South economic relations than a single and simple label might suggest. On the one hand are the fast growing newly industrializing countries such as Korea, Taiwan, Hong Kong and Singapore, which have had protectionist measures directed against their manufactured exports, and on the other hand are countries in Africa which do not have any manufactured exports. Their exports are still primarily agricultural in nature, and are suffering from a slump in demand. A stemming of the protectionist tide will not help them, and it is here that the most dire poverty in the world is to be seen.

Let us turn, then, to an answer to the question posed at the start. What was achieved by the Brandt Report, and by the push in the past decade for a New International Economic Order between North and South? One important contribution of the Report will have been to highlight the complex web of relations between North and South which make one region interdependent on the other. The global events of the past few years, in particular the debt crisis, have only served to underline this factor. But more is perhaps revealed by what the metaphor of North–South excludes than by what it includes. As noted at the start, the position of the Communist bloc in this categorization is not clear, and any attempt to place them in the Northern group has been resisted by the block itself. Some have suggested that these centrally planned economies occupy a middle position in terms of trade – they import primary products from the South and export medium technology manufactures to them. From the North they import high technology manufactures and export primary products and medium technology manufactures. They themselves have chosen to characterize North–South economic relations as emanating from a colonial past, and have excluded themselves from the categorization altogether.

But perhaps the greatest difficulty in making sense of a global concept such as North–South economic relations is the great diversity of the South. It includes labour surplus countries in Asia and land surplus countries in Africa; highly industrialized countries in Far East Asia and Latin America, and primarily agricultural ones in Africa; light manufactures exporters and heavy manufactures exporters; countries which themselves have multinationals in other Southern countries; countries in which organized labour is strong and countries in which it has been brutally suppressed; countries which have elected governments and countries which have always been ruled by dictatorships. While the North–South metaphor has proved useful in crystallizing certain features of the divide between rich and poor, attention must now turn to the details of the case under consideration.

See Also

- ▶ Immiserizing Growth
- ▶ Periphery
- ▶ Strategic Reallocations of Endowments
- ▶ Terms of Trade

Bibliography

Brandt, W. 1980. *North-South. A program for survival*. London: Pan.

Brandt, W. 1983. *Common crisis. North-South: Co-operation for world recovery*. London: Pan.

Kaldor, N. 1976. Inflation and recession in the world economy. *Economic Journal* 86: 703–714.

Spraos, J. 1980. The statistical debate on the net barter terms of trade between primary commodities and manufactures. *Economic Journal* 90: 107–128.

Novalis [Georg Friedrich Philipp Von Hardenberg] (1772–1801)

Murray Milgate

Abstract

Born on 2 May 1772 in Saxony, Novalis ranks among the finest of the German Romantic writers. His works mark the transition from early Romanticism to that more politically oriented movement (not always of reaction) that rose to prominence in the 19th century. Novalis was educated in Jena and Leipzig, attending Schiller's lectures on history and becoming closely associated with Friedrich Schlegel. After graduating in law (and while serving as a minor government official in Arnstadt) he embarked upon a systematic study of Fichte's *Wissenschaftslehre.* Himself a traditionalist and an admirer of Edmund Burke, Novalis promulgated an 'organistic' view of society and called into question mechanistic and utilitarian conceptions which in economics had been the hallmark of the Enlightenment. Perhaps the best example of that critique is the complex tale which comprises the ninth chapter of his unfinished *Heinrich von Ofterdingen.*

Born on 2 May 1772 in Saxony, Novalis ranks among the finest of the German Romantic writers. His works mark the transition from early Romanticism to that more politically oriented movement (not always of reaction) that rose to prominence in the 19th century. Novalis was educated in Jena and Leipzig, attending Schiller's lectures on history and becoming closely associated with Friedrich Schlegel. After graduating in law (and while serving as a minor government official in Arnstadt) he embarked upon a systematic study of Fichte's *Wissenschaftslehre.* Himself a traditionalist and an admirer of Edmund Burke, Novalis promulgated an 'organistic' view of society and called into question mechanistic and utilitarian conceptions which in economics had been the hallmark of the Enlightenment. Perhaps the best example of that critique is the complex tale which comprises the ninth chapter of his unfinished *Heinrich von Ofterdingen.*

While the trend in Novalis' writing is correctly regarded today as conservative, if not authoritarian, it should not be forgotten that his early questioning of the notion that economic progress consisted in the acquisition of material wealth in a society regulated only by self-interest, has reverberated in manifold ways right down to the present day. His 'organistic' conception of society (similar to that of Coleridge) was taken up directly by Adam Heinrich Müller and through this channel influenced economists as diverse as Rodbertus, List and Othmar Spann. Novalis died on 25 March 1801; he would have been 29 at his next birthday. His prose poem *Hymnen an die Nacht* is the work by which he is best known.

Selected Works

Aphorisms and fragments. In *German romantic criticism,* ed. A. Leslie Wilson, New York: Continuum, 1982. *Heinrich von Ofterdingen.* Trans. P. Hilty, New York: Ungar, 1962. *Hymnen an die Nacht.* Trans. in *German poetry: 1750–1900,* ed. R.M. Browning, New York: Continuum, 1984. *Schriften.* Ed. P. Kluckhorn and R. Samuel, Leipzig, 1929.

Bibliography

Pinson, K.S. 1933. Novalis. In *Encyclopaedia of the social sciences*, ed. E.R.A. Seligman. London: Macmillan.

Nove, Alexander (Alec) N. (1915–1994)

P. J. D. Wiles

Keywords

Nove, A. N.; Planning; Socialism; Stalinism, political economy of

JEL Classifications
B31

Born in Leningrad, of a Polish–Jewish family Novakovski, Alec Nove would say he felt like a Russian. His father was a Menshevik and his uncle a Bolshevik. He used as a small boy to listen to them arguing, and respected his uncle more. His native language was Russian. The family emigrated to London shortly after the Revolution.

Nove was educated at the London School of Economics (B.Sc. Econ. 1936).

His first civilian job was at the Board of Trade and he entered academic life in 1958 as Reader at the London School of Economics. He was Professor Emeritus at Glasgow University at the time of his death.

Impatience with orthodox theory and its whole implementarium did not conceal a sharp economic mind. This was applied mainly to Sovietology and to socialism generally. Nove comes after the great pioneers of Sovietological economics: Sergei Prokopovich, Naum Jasny, Solomon Schwarz and (slightly younger) Abram Bergson. Less Soviet or at least less Russian than the first three, he was also less Western than the last, and this from personal choice, since his whole education was Western. But he always cultivated an understanding of the system in its own terms, and this fit in with his anti-neoclassical bent. A flaw here however was his extreme reluctance to master Marxist ideology in its many varieties: it was a strictly practical view of the Soviet system that was taken.

His methodology can only be called breadth of mind, energy and intuition: foraging through the wasteland of the current Russian literature and making new and important insights. Nove was the first to write seriously about the *variety* of the success indicators imposed upon planned enterprises; the first to note that in about 1980 Soviet economists were producing and almost publishing their own price indices (these rose far quicker than the official ones); one of the first to spot the brave scholars who were revising the harvest figures during the collectivization and the famine.

Much of his work was political economy: Trotsky and socialism; Stalinism and planning; the decision as to when to collectivize agriculture; glasnost. There is a also a cornucopia of minor contributions on Soviet literature; pre-revolutionary Russian opera; Poland; Hungary; and the misuse of economic criteria by the British public sector (this sideline however was flawed by Scottish nationalism, if that is a correct name for the disgruntlement of a Glaswegian globetrotter who finds he must go everywhere via London).

Here as everywhere inspired common sense and strong empirical knowledge produced work that was occasionally wrong-headed, usually brilliant, very seldom dull, never unclear.

Selected Works

1961. *The Soviet economy.* London: George Allen & Unwin; 3rd ed., 1969.

1965. *Was Stalin really necessary?* London: George Allen & Unwin.

1969. *An economic history of the USSR*. London: Allen Lane/The Penguin Press, 3rd ed., 1993.

1973. *Efficiency criteria for nationalised industries*. London: Allen & Unwin.

1977. *The Soviet economic system*. London: Allen & Unwin.

1979. *Political economy and Soviet socialism*. London: Allen & Unwin.

1983. *The economics of feasible socialism*. London: Allen & Unwin; revision *The economics of socialism revisited.* London: Routledge, 1992.

1986. *Socialism, economics and development*. London: Allen & Unwin.

1989. *Glasnost in action: Cultural renaissance in Russia*. London: Routledge.

1991. *Studies in economics and Russia*. London: Palgrave MacMillan.

1998. *Alec Nove on communist and post-communist countries. Previously unpublished writings*, 2, ed. I. Thatcher. Cheltenham: Edward Elgar.

1998. *Alec Nove on economic theory. Previously unpublished writings*, 1, ed. I. D. Thatcher. Cheltenham: Edward Elgar.

Novozhilov, Viktor Valentinovich (1892–1970)

Holland Hunter and Robert W. Campbell

Keywords

Capital intensity; Depreciation; Inversely related expenditures; Kantorovich, L. V.; Labour theory of value; Nemchinov, V. S.; Novozhilov, V. V.; Opportunity cost; Resource allocation; Value theory

JEL Classifications

B31

Novozhilov was born in Khar'kov, and died in Leningrad. He was instrumental, along with the mathematician Leonid Vital'evich Kantorovich, in reviving a mathematical approach to economic theory in the USSR after Stalin's death, and in laying a basis for a modern theory of value and allocation.

Educated at Kiev University before the revolution, Novozhilov taught at several institutions in the Ukraine, but from 1922 lived in Leningrad, teaching and working in research institutes. From 1935 he taught at the Leningrad Polytechnical Institute, and from 1944 until 1952 was also professor and head of the Department of Statistics at the Leningrad Engineering–Economics Institute. His work with projectmaking institutes involved Novozhilov in the issue of capital intensity choices, which became the basis for his doctoral dissertation. In illuminating the question of effective allocation of capital among competing projects, he developed a more general theory for allocation of all resources, the centrepiece of which was the concept of 'inversely related expenditures' (*zatraty obratnoi sviazi*) equivalent to opportunity cost. His analytic framework was dynamic, incorporating capital allocation over time, as well as the impact of depreciation and obsolescence.

His original and elegant theoretical ideas were presented in papers published in 1939, 1941, 1946 and 1947 that were largely ignored. The most comprehensive exposition of Novozhilov's ideas is a book he was finally able to publish in 1967, which illustrates his ideas on investment choices and the time factor in economics, places his innovative approach in its doctrinal context, and defends it against domestic and foreign critics. His economic theory is expounded within the limits of political orthodoxy. Novozhilov took the structure of demand as given (by the Party), which enabled him to spell out resource-allocating criteria for the Soviet economy very similar to those familiar in the West, except that with the demand blade of the scissors held fixed, only the supply side cut the paper. By casting the resource allocation problem in terms of minimizing labour input (direct and indirect) he sought to preserve Marx's labour theory of value. Both his contribution and the absence of an explanation of demand were soon recognized abroad (see Grossman 1953; Campbell 1961).

In the mid-1950s, when V.S. Nemchinov organized a revival of serious economic analysis in the USSR, Novozhilov, along with Kantorovich, was a central figure in training a new generation of economists. The three men were awarded Lenin Prizes in 1965. As a result of the pioneering work of Novozhilov and Kantorovich, the basis for a correct and comprehensive theory of value has already been to hand for several decades. Additional biographic and bibliographical details, and interpretations of Novozhilov's work may be found in Campbell (1961), Ellman (1973), Grossman (1953), Holubnychy (1982), and Petrakov (1972).

See Also

- ▶ Economic Calculation in Socialist Countries
- ▶ Kantorovich, Leonid Vitalievich (1912–1986)

Selected Works

1939. Metody soizmereniia narodnokhoziaistvennoi effektivnosti planovykh i proektnykh variantov [Methods of commeasuring the economic effectiveness of variants in planning and project-making]. *Trudy Leningradskogo industrial'nogo institute* [Papers of the Leningrad Industrial Institute] No. 4. Leningrad.

1941a. Metody izmereniia narodnokhoziaistvennoi effektivnosti proektnykh variantov [Methods of measuring the national economic effectiveness of project variants]. Dissertation for the doctoral degree, awarded in 1941.

1941b. Praktikuemye metody soizmereniia sebestoimosti i vlozhenii [Methods used in practice for co-measuring current outlays and investments]. *Trudy Leningradskogo politekhnicheskogo instituta* [Papers of the Leningrad Polytechnical Institute] No. 1. Leningrad.

1946. Metody nakhozhdeniia minimuma zatrat v sotsialisticheskom khoziaistve [Methods of finding the minimum expenditure in a socialist economy]. *Leningradskii politekhnicheskii institut imeni M.I. Kalinina*: *Trudy* [The Leningrad Kalinin Polytechnical Institute Papers] No. 1. Leningrad.

1947. Sposoby nakhozhdeniia maksimuma effekta kapitalovlozhenii *v* sotsialisticheskom khoziaistve [Methods for finding the maximum effect of capital investments in the socialist economy]. *Trudy Leningradskogo finansovo-ekonomicheskogo instituta* [Papers of the Leningrad Financial Economic Institute], Vypusk III [Issue 3]. Leningrad.

1967. *Problemy izmereniia zatrat i rezul'tatov pri optimal'nom planirovanii* [Problems of cost–benefit analysis in optimal planning]. Moscow. 2nd ed., 1972. Trans. (with title as shown), White Plains: International Arts and Sciences Press, 1970.

1972. *Voprosy razvitiia sotsialisticheskoi ekonomiki* [Questions of the development of socialist economics]. Moscow.

Bibliography

Campbell, R.W. 1961. Marx, Kantorovich, and Novozhilov: Stoimost' versus reality. *Slavic Review* 20: 402–418.

Ellman, M. 1973. *Planning problems in the USSR; The contribution of mathematical methods to their solution, 1961–1971*. Cambridge: Cambridge University Press.

Grossman, G. 1953. Scarce capital and Soviet doctrine. *Quarterly Journal of Economics* 67: 311–343.

Holubnychy, V. 1982. V.V. Novozhilov's theory of value. In *Soviet regional economics: selected works of Vsevolod Holubnychy*, ed. I.S. Koropeckyj. Edmonton: Canadian Institute of Ukrainian Studies, University of Alberta.

Petrakov, N.I. 1972. Nauchnaia i pedagogicheskaia deiatel'nost' V.V. Novozhilova [The scientific and pedagogical work of V.V. Novozhilov]. In Novozhilov (1972).

Numeraire

Michael Allingham

Keywords
Money; numeraire

JEL Classifications
D0

In general equilibrium theory the price of one good in terms of another is interpreted as the amount of the second which can be exchanged for a given amount of the first. There is thus no essential role for a standard of value, or *numéraire,* though it is frequently helpful to introduce this. Such a *numéraire* is a commodity in terms of which, by convention, other commodities are valued.

The concept seems to have been introduced by Steuart (1767), albeit with some confusion between the properties of 'money' and 'units of account'. Walras (1874–7) clarified the concept, and showed how prices expressed in terms of one *numéraire* could be translated into prices in terms of another, without any introduction of 'money'. In the present discussion we commence with a justification of the use of a *numéraire.* We then discuss the choice of a *numéraire* and some problems which may arise through the use of this.

We may represent an economy with n commodities by the excess demand function f: $S \to R^n$ where $S = R^n_+ - 0$. The interpretation is that f (p) is the vector of aggregate excess demands (positive) or excess supplies (negative) expressed at the price system p. A basic property of f is that it is homogeneous of degree zero, that is $f(tp) = f(p)$ for all positive t.

It is this property which justifies the use of a *numéraire.* We can, for example, take commodity n to be *numéraire,* that is, ensure that $p_n = 1$, by setting the scalar t appropriately. Thus the price system q can be replaced by the *numéraire* price system p with $p_n = 1$ by multiplying q by $t = 1/q_n$; nothing real changes, since $f(p) = f(q)$. However, this is only possible if we can ensure that q_n is positive; since q is restricted only to S this may prove difficult.

The problem of the price of a chosen *numéraire* possibly being zero may be avoided by using a composite *numéraire*, that is a basket of goods. The scalar t may then be set as $u \cdot q$ where u is the unit vector in R^n*:* this has the effect of restricting p to the unit simplex in R^n. Alternatively, a nonlinear normalization may be used, for example setting $t = q \cdot q$*:* this has the effect of restricting p to the surface of a sphere in R^n_+.

However, in reality prices are usually quoted in terms of some single unit of account, or *numéraire*, and it may be useful for the model of the economy to recognize this. Provided that all commodities are desirable, in the sense that $f_i(p)$ is infinite if $p_i = 0$, there is no possibility of any price being zero in equilibrium, that is some p where $f(p) = 0$. But there may be a problem of p_i being zero on some adjustment path of prices. Whether this is indeed a problem will depend on both the nature of f and on the adjustment process governing this path. For example, if the adjustment process is given by $\dot{p} = h(f(p))$ where h is a continuous sign-preserving function (and a dot indicates differentiation with respect to time) and if f has the above desirability property, then there is no problem. Alternatively, if the adjustment process is $\dot{p}_i = 0$ if $p_i \leq 0$ and $f_i(p) < 0$, while $\dot{p}_i = h_i(f_i(p))$ otherwise, then again there is no problem, provided of course that initial prices are positive (Arrow and Hahn 1971). However, if these properties do not apply, and particularly if the adjustment process is discrete, there may be a problem.

Provided we can use a simple *numéraire* it is clear that if equilibrium is unique in terms of one *numéraire* then it will be unique in terms of another. However, the choice of *numéraire* may be relevant to considerations of stability: that is, for some given adjustment process involving a *numéraire* the economy may be stable for some *numéraire* but not for some other. Some sufficient conditions for stability, such as the condition that f have the revealed preference property, are clearly independent of any choice of a *numéraire,* while others are not (Hahn 1982). For example, the diagonal dominance condition that all commodities are normal and that there are some units in which commodities can be measured such that each of their excess demands is more sensitive to a change in its own price than it is to a change in all other *non-numéraire* prices combined, is clearly dependent on the choice of *numéraire;* indeed, because of homogeneity it makes no sense to attempt to extend it to include the *numéraire.* An economy may have this property, which is sufficient for stability, for one *numéraire* but not for some other. Since this condition is not necessary for stability it does not follow that the economy will be unstable with the second *numéraire,* but neither can stability be guaranteed.

The reason why uniqueness, for example, does not depend on the choice of *numéraire* while stability may, is that stability depends on the adjustment process. Strictly speaking, a change of *numéraire* is simply a change of adjustment process: it is quite natural that the economy may be stable under one adjustment process but not under another.

The question of a *numéraire* has a practical as well as a theoretical importance. In many cases 'money' is the natural *numéraire* – though the introduction of money in an essential sense, as opposed to simply as a unit of account, introduces its own problems (Clower 1967).

See Also

▶ Walras, Léon (1834–1910)

Bibliography

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

Clower, R.W. 1967. A reconsideration of the microfoundations of monetary theory. *Western Economic Journal* 6: 1–8.

Hahn, F. 1982. Stability. In *Handbook of mathematical economics*, ed. K.J. Arrow and M.D. Intriligator, vol. 2. Amsterdam: North-Holland.

Steuart, Sir J. 1767. *Principles* (Book 1). London.

Walras, L. 1874–7. *Eléments d'économie politique pure.* Definitive ed., Lausanne: Corbaz, 1926. Trans. W. Jaffé as *Elements of pure economics*. London: George Allen & Unwin, 1954.

Numerical Determination of the Laws of Utility

F. Y. Edgeworth

Abstract

Numerical Determination of the Laws of Utility is the title given by Jevons (*Theory of Political Economy,* 2nd edn, p. 158) to an operation which he, like Gossen, regards as possible – the ascertainment of the form of demand curves by statistics of prices and consumption. It may be objected to this phrase, that laws of utility cannot be deduced from laws of price, except on the assumption that price is the measure of utility – the Marginal Utility of money being constant (see "▶ Final Degree of Utility). But, even upon this assumption, there are great difficulties in the way of the statistical operation. First, the utility derived from a set of articles is in general not the simple sum, but some unknown function, of the utilities derived from each. Thus the amount consumed of any one article will vary with the prices of others – especially of those which are substitutes for the one under consideration, as tea is for coffee, or complementary to it, as bats are to balls. Accordingly, to observe the changes in the demand for an article corresponding to the changes in its price is apt to be nugatory unless it can be assumed that the prices of all other articles are constant. Again, utility is not only a complicated function of the amounts consumed, but a variable one, changing its form with every vicissitude of taste and fashion. Professor Marshall has pointed out these and other difficulties (*Principles,* bk. iii, ch. iii), and attempted to evade them (ibid., last section).

Numerical Determination of the Laws of Utility is the title given by Jevons (*Theory of Political Economy,* 2nd edn, p. 158) to an operation which he, like Gossen, regards as possible – the ascertainment of the form of demand curves by statistics of prices and consumption. It may be objected to this phrase, that laws of utility cannot be deduced from laws of price, except on the assumption that price is the measure of utility – the Marginal Utility of money being constant (see "▶ Final Degree of Utility). But, even upon this assumption, there are great difficulties in the way of the statistical operation. First, the utility derived from a set of articles is in general not the simple sum, but some unknown function, of the utilities derived from each. Thus the amount consumed of any one article will vary with the prices of others – especially of those which are substitutes for the one under consideration, as tea is for coffee, or complementary to it, as bats are to balls. Accordingly, to observe the changes in the demand for an article corresponding to the changes in its price is apt to be nugatory unless it can be assumed that the prices of all other articles are constant. Again, utility is not only a complicated function of the amounts consumed, but a variable one, changing its form with every vicissitude of taste and fashion. Professor Marshall has pointed out these and other difficulties (*Principles,* bk. iii, ch. iii), and attempted to evade them (ibid., last section).

Numerical Optimization Methods in Economics

Karl Schmedders

Abstract

Optimization problems are ubiquitous in economics. Many of these problems are sufficiently complex that they cannot be solved analytically. Instead economists need to resort to numerical methods. This article presents the most commonly used methods for both unconstrained and constrained optimization problems in economics; it emphasizes the solid theoretical foundation of these methods, illustrating them with examples. The presentation includes a summary of the most popular software packages for numerical optimization used in economics, and closes with a description of the rapidly developing area of mathematical programs with equilibrium constraints, an area that shows great promise for numerous economic applications.

Keywords

Central path; Computable general equilibrium models; Computational algebraic geometry; Constrained optimization; Global minimum; Homotopy continuation methods; Interior-point methods; Karush-Kuhn-Tucker conditions; Kuhn–Tucker conditions; Lagrange multipliers; Line search methods; Linear independence constraint qualification (LICQ); Linear programming; Local minimum; Local solution; Logarithmic barrier method; Mathematical programs with equilibrium constraints; Metaheuristics; Multidimensional optimization problems; Newton's method; Newton–Raphson methods; Nonlinear programming; Numerical optimization methods; Optimality conditions; Optimization algorithms; Penalty methods; Polynomial functions; Quadratic programs; Sequential quadratic programming; Simplex method for solving linear programs; Simulated annealing; Smale's global Newton method; Steepest descent method; Taylor's th; Trust region methods; Unconstrained optimization

JEL Classifications

D5

Optimizing agents are at the centre of most economic models. In our models we typically assume that consumers maximize utility or wealth, that players in a game maximize payoffs, that firms minimize costs or maximize profits, or that social planners maximize welfare. But it is not only the agents in our models that optimize. Econometricians maximize likelihood functions or minimize sums of squares. Clearly optimization is one of the key techniques of modern economic analysis.

The optimization problems that appear in economic analysis vary greatly in nature. We encounter finite-dimensional problems such as static utility maximization problems with a few goods. An optimal solution to such a problem is a finite-dimensional vector. We analyse infinite-dimensional problems such as infinite-horizon social planner models or continuous-time optimal control problems. Here the solution is an infinite-dimensional object, a vector with countably infinitely many elements or even a function over an interval. Our agents may face constraints such as budget equations, short-sale restrictions or incentive-compatibility constraints. There are also unconstrained problems such as nonlinear least-square problems. Decision variables may even be restricted to be discrete. Agents' objective functions may be linear or nonlinear, convex or nonconvex, many times differentiable or discontinuous. Finally, an economic optimization problem may be deterministic or stochastic.

Unless we consider stylized models in theoretical work or make very stringent and often quite unrealistic assumptions in applied models, the optimization problems that we encounter cannot be solved analytically. Instead we need to resort to numerical methods. The numerical methods that we employ to solve economic optimization models vary just as much as the optimization problems we encounter. It is therefore impossible to cover the wide variety of numerical optimization methods that are useful in economics in a short article. For the purpose of the exposition here we focus on deterministic finite-dimensional nonlinear optimization problems including linear programs. This is a natural choice because such problems are ubiquitous in economic analysis. Moreover, the techniques for these problems play also an important part in many other numerical methods such as those for solving economic equilibrium and infinite-dimensional problems. The interested reader should consult computation of general equilibria (new developments), computational methods in econometrics and dynamic programming.

We first indicate some of the fundamental technical difficulties that we need to be aware of when we apply numerical methods to our economic optimization problems. We then highlight the basic theoretical foundations for numerical optimization methods. The popular numerical optimization methods have strong theoretical foundations. Unfortunately, current textbooks in computational economics, with the partial exception of Judd (1998), neglect to emphasize these foundations. As a result some economists are rather sceptical about numerical methods and view them as rather ad hoc approaches. Instead, a good understanding of the theoretical foundations of the numerical solution methods gives us an appreciation of the capabilities and limitations of these methods and can guide our choice of suitable methods for a specific economic problem. We outline the most fundamental numerical strategies that form the basis for most algorithms. All presented numerical strategies are implemented in at least one of the those computer software packages for solving optimization problems that are most popular in economics. We close our discussion with a look at mathematical programs with equilibrium constraints (MPECs), a promising research area in numerical optimization that has useful applications in economics.

Newton’s Method in one Dimension

We start with the one-dimensional unconstrained optimization problem

$$\min_{x \in \mathbb{R}} f(x). \tag{1}$$

Perhaps the first (if any) numerical method that most of us learnt in our calculus classes is Newton’s method. Newton’s method attempts to minimize successive quadratic approximations to the objective function f in the hope of eventually finding a minimum of f. To start the computations we need to provide an initial guess $x^{(0)}$. The quadratic approximation $q(x)$ of $f(x)$ at the point x^0 is

$$q(x) = f\left(x^0\right) + f'\left(x^{(0)}\right)\left(x - x^{(0)}\right) + \frac{1}{2} f''\left(x^{(0)}\right)\left(x - x^{(0)}\right)^2$$

where f' and f'' denote the first and second derivative of the function f, respectively. Solving the first-order condition

$$q'(x) = f'\left(x^{(0)}\right) + f''\left(x^{(0)}\right)\left(x - x^{(0)}\right) = 0$$

on the assumption that f' $(x^{(0)}) \neq 0$ yields the solution.

$$x^{(1)} = x^{(0)} - \frac{f'\left(x^{(0)}\right)}{f''\left(x^{(0)}\right)}.$$

Now we repeat this process using a quadratic approximation to f at the point $x^{(1)}$. The result is a sequence of points, $\{x^{(k)}\} = x^{(0)}, x^{(1)}, x^{(2)}, \ldots, x^{(k)}, \ldots$, that we hope will converge to the solution of our minimization problem. This approach is based on the following theoretical result.

Theorem Suppose x^* is the solution to the minimization problem (1). Suppose further that f is three times continuously differentiable in a neighborhood of x* and that f' $(x^*) \neq 0$. Then there exists some $\delta > 0$ such that if $|x^* - x(0)| < \delta$, then the sequence $\{x^{(k)}\}$ converges quadratically to x*, that is,

$$\lim_{k\to\infty} \frac{|x^{(k+1)} - x^*|}{|x(k) - x^*|^2} = \kappa$$

for some finite constant κ.

We illustrate this theorem with a simple example.

Example 1 A consumer has a utility function $u(x, y) = \ln(x) + 2\ln(y)$ over two goods. She can spend $1 on buying quantities of these two goods, both of which have a price of $1. After substituting the budget equation, $x + y = 1$, into the utility function the consumer wants to maximize $f(x) = \ln(x) + 2\ln(1 - x)$. Setting the first order condition equal to 0 yields the solution $x^* = \frac{1}{3}$ (This quantity is globally optimal because the function f is strictly concave.)

Suppose we start Newton's method with the initial guess $x^{(0)} = 0.5$. Then the first Newton step yields

$$x^{(1)} = 0.5 - \frac{f'(0.5)}{f''(0.5)} = 0.5 - \frac{-2}{-12} = \frac{1}{3}.$$

Newton's method found the exact optimal solution in one step. This (almost) never happens in practice. Much more usual is the behaviour we observe when we start with $x^{(0)} = 0.8$. Then Newton's method delivers as its first five steps

$$0.63030303, 0.407373702, 0.328873379,$$
$$0.333302701, 0.333333332.$$

We observe that the sequence rapidly converges to the optimal solution. The corresponding errors $|x^{(k)} - x*|$,

$$0.2969697, 0.07404037, 0.00445995, 3.0632\cdot 10^{-5}, 1.4078\cdot 10^{-9}$$

converge to but never exactly reach zero. The rate of convergence is called quadratic since $|x^{(k+1)} - x^*| < L|x^{(k)} - x^*|^2$ for some constant L once k is sufficiently large.

But, of course, contrary to this simple example, we typically do not know x^* and so cannot compute the errors $|x^{(k)} - x^*|$. Instead, we need a stopping rule that indicates when the procedure terminates. The requirement that $f'(x^{(k)}) < \delta$ may appear to be an intuitive stopping rule. But that rule may be insufficient for functions that are very 'flat' near the optimum and have large ranges of non-optimal points satisfying this rule. Therefore, a safer stopping rule requires both $f'(x^{(k+1)}) < \delta$ and $|x^{(k+1)} - x^{(k)}| < \varepsilon(1 + |x^{(k)}|)$ for some pre-specified small error tolerance $\varepsilon, \delta > 0$. So the Newton method terminates once two subsequent iterates are close to each other and the first derivative almost vanishes.

Observe that Newton's method found a maximum, and not a minimum, of the utility function. The reason for this fact is that this method does not search directly for an optimizer. Note that the key step in the algorithm is finding a stationary point of the quadratic approximation $q(x)$, that is, a point satisfying $q'(x) = 0$. Before we can claim to have found a maximum or minimum of f we need to do more work. In this example the strict concavity of the utility function ensures that a stationary point of f yields a maximum. So an assumption of our economic model assures us that the numerical method indeed finds the desired maximum.

Example 2 Consider the simple polynomial function $f(x) = x(x - 2)^2$. Starting with $x^{(0)} = 1$ leads to the sequence

$$0.5, 0.65, 0.666463415, 0.666666636, \ldots$$

converging to $\frac{2}{3}$. Starting with $x^{(0)} = 1.5$ leads to the sequence

$$2.75, 2.198529412, 2.022777454, 2.000376254$$

converging to 2. Neither of these two points yields a global optimum, the function f is actually unbounded above and below. The point $\frac{2}{3}$ is a local maximizer ($f''(2/3) = -4 < 0$) while 2 is a local minimizer ($f''(2) = 4 > 0$). The stationary point that we find greatly depends on our initial guess.

Our simple observations about the behaviour of Newton's method for one-dimensional optimization problems apply in practice to higher-dimensional nonlinear optimization problems and to almost all optimization methods. We will almost always face these fundamental issues in our economic applications. First, most practical optimization methods for unconstrained problems search only for a stationary point (with possibly additional favourable properties). They do not directly attempt to compute an optimizer. Second, as a result, most practical methods may terminate with a non-optimal point. To ensure global optimality we need to perform additional checks. Third, it is rather unusual in practice to explicitly solve for an exact solution. Usually we can only hope for a sequence of points $\{x^{(k)}\}$ generated by an iterative process that converges to a limit having some desired property. Therefore, we need a stopping rule that indicates when the iterative process stops. Fourth, the algorithm may not terminate and diverge even if a globally optimal solution exists.

Newton's method is a special instance of a family of methods for solving multidimensional optimization problems. Before we examine more general methods we provide some basic intuition for the theoretical underpinnings of these solution methods.

Theoretical Foundation: Taylor's Th

The gradient of the function f at a point $x = (x_1, x_2, \ldots, x_n)$ is the column vector

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}(x), \ldots, \frac{\partial f}{\partial x_n}(x)\right)^\top$$

of partial derivatives of f with respect to the variables $x_1, x_2, \ldots, x_n$. The Hessian of f at x is the $(n \times n)$-matrix

$$H(x) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(x)\right)_{i,j=1}^{n}$$

of the second derivatives $\frac{\partial^2 f}{\partial x_i \partial x_j}(x)$ of f. The inner product of two (column) vectors $x, y \in \mathbb{R}^n$ is denoted by $x^\top y$.

Many numerical methods rely on linear or quadratic approximations of the function f. Taylor's theorem provides a justification for this approach. Here we give a simple version of this theorem for functions with Lipschitz continuous derivatives. Consider a function $F: X \rightarrow Y$ for open sets $X \subset \mathbb{R}^n$ and $Y \subset \mathbb{R}^m$. Then F is Lipschitz continuous at $x \in X$ if there exists a constant $\gamma(x)$ such that

$$||F(y) - F(x)|| \quad \leq \gamma(x)||y - x||$$

for all $y \in X$, where $|| \cdot ||$ denotes the standard Euclidean norm.

Theorem Suppose the function $f: X \rightarrow$ R is continuously differentiable on the open set $X \subset \mathbb{R}^n$ and that the gradient function ∇f is Lipschitz continuous at x with Lipschitz constant $\gamma^l(x)$. Also suppose that for $s \in \mathbb{R}^n$ the line segment $x + \theta s \in X$ for all $\theta \in [0,1]$. Then, the linear function l with $l(s) = f(x) + \nabla f(x)^\top s$ satisfies

$$|f(x + s) - l(s)| \quad \leq \frac{1}{2}\gamma^l(x)||s||^2.$$

Moreover, if f is twice continuously differentiable on X and the Hessian H is Lipschitz continuous at x with Lipschitz constant $\gamma^q(x)$, then the quadratic function q with $q(s) = f(x) + f(x)^\top s + \frac{1}{2}s^\top H(x)s$ satisfies

$$|f(x + s) - q(s)| \quad \leq \frac{1}{6}\gamma^q(x)||s||^3.$$

Unconstrained Optimization

The multidimensional generalization of the unconstrained optimization problem (1) is given by

$$\min_{x \in \mathbb{R}^n} f(x). \qquad (2)$$

Solving this optimization problem entails finding a global minimizer x^* satisfying $f(x^*) \leq f(x)$ for all $x \in \mathbb{R}^n$. With the exception of a few algorithms for problems that are either very small or have very special structure, there are no algorithms that are guaranteed to find a global minimum. Thus, we need to think in terms of local minima. A local minimizer is a point x^* that satisfies $f(x^*) \leq f(x)$ for all $x \in \mathscr{N}(x^*)$ where $\mathscr{N}(x^*)$ denotes a neighborhood of x^*. The point x^* is called an isolated local minimizer if it is the only local minimizer in $\mathscr{N}(x^*)$.

All these definitions by themselves are not all that helpful for finding a minimum. Instead, just as Newton's method in one dimension does, all practical numerical methods for unconstrained optimization problems rely on optimality conditions to find candidates for local minima. For functions with sufficient differentiability properties these are the following well-known conditions.

Theorem [Optimality conditions for unconstrained minimization].

1. If f is continuously differentiable and x^* is a local minimizer of f, then $\nabla f(x^*) = 0$.
2. If f is twice continuously differentiable and x^* is a local minimizer of f, then $\nabla f(x^*) = 0$ and $s^\top H(x^*)s \geq 0$ for all $s \in \mathbb{R}^n$.
3. If f is twice continuously differentiable and if x^* satisfies $\nabla f(x^*) = 0$ and $s^\top H(x^*)s > 0$ for all $s \in \mathbb{R}^n$, $s \neq 0$, then x^* is an isolated local minimizer of f.

But when can we be assured that a local minimizer of f is actually a solution to the unconstrained optimization problem (2)? The perhaps easiest sufficient condition is that the function f is convex, that is, $s^\top H(x)s \geq 0$ for all $x \in \mathbb{R}^n$ if f is twice differentiable. Then any local minimizer x^* is a solution to problem (2), in fact, any stationary point x^* is a solution to (2).

The optimality conditions provide the foundation for all practical unconstrained optimization methods. The focus of all these algorithms is to find (actually, to approximate) a stationary point of f, that is, a solution to $\nabla f(x) = 0$. They do so by generating a sequence of iterates $\{x^{(k)}\}$ that ideally terminates once a stopping rule is satisfied indicating that an approximate solution has been found. The key step for these methods is to generate a new iterate $x^{(k+1)}$ from a current iterate $x^{(k)}$. A vast majority of optimization routines uses one of two basic strategies for moving from $x^{(k)}$ to $x^{(k+1)}$, a line search approach or a trust region method.

Line Search Methods

The general set-up of a line search method is as follows. From a point $x^{(k)}$ (with $\nabla f(x^{(k)}) \neq 0$) we look for a search direction $s^{(k)}$ that leads us to lower function values for f. Using the linear approximation l with $l(s) = f(x^{(k)}) + \nabla f(x^{(k)})^\top s$ we determine a descent direction $s^{(k)}$ satisfying

$$\nabla f(x)^\top s^{(k)} < 0,$$

which in turn implies $l(s^{(k)}) < f(x^{(k)})$. Because of Taylor's theorem we hope that along a step in the direction $s^{(k)}$ the function value $f(x)$ will be reduced. We calculate a suitable step length $\alpha_k > 0$ to ensure that $f(x^{(k+1)}) < f(x^{(k)})$ where

$$x^{(k+1)} = x^{(k)} + \alpha_k s^{(k)}.$$

Observe that at a given point $x^{(k)}$ and for a descent direction $s^{(k)}$ finding the optimal value of α_k requires us to solve a one-dimensional optimization problem. In principle we could apply Newton's method to this problem. In practice, however, this one-dimensional problem does not need to be solved exactly because repeatedly finding the optimal step length is both unnecessary for convergence of line search methods and computationally rather inefficient. Instead modern line search methods prefer to use inexact line searches that just pick a step length that leads to a sufficient decrease in the objective function value. One such approach is the backtracking Armijo line search, which requires that

$$f\left(x^{(k)} + \alpha_k s^{(k)}\right) \leq f\left(x^{(k)}\right) + \alpha_k \beta \nabla f\left(x^{(k)}\right)^\top s^{(k)}$$

for some $\beta \in (0,1)$. The idea of this requirement is to link the step size α_k to the decrease in f. The longer the step the larger the decrease must be. Starting with an initial guess for α_k, say 1, we can now stepwise reduce the value of α_k until the above condition is satisfied. At that point we set $x^{(k+1)} = x^{(k)} + \alpha_k s^{(k)}$.

While the basic line search method seems very intuitive, it can fail if the search direction and the gradient tend to a point where they are orthogonal to each other, that is, the product $\nabla f(x^{(k)})^\top s^{(k)}$ tends to zero without the gradient itself approaching zero. This kind of failure can be avoided by a proper choice of search direction.

Method of Steepest Descent

The perhaps most intuitive choice for a descent direction is

$$s^{(k)} = -\nabla f\left(x^{(k)}\right),$$

because this search direction gives the greatest possible decrease in the linear approximation l (for a fixed step length). It is thus called the steepest descent direction. And indeed, a line search with the steepest descent direction has very nice theoretical properties.

Theorem Suppose that f is continuously differentiable and that ∇f is Lipschitz continuous on $\mathbb{R}^n$. Then for the sequence $\{x^{(k)}\}$ of iterates generated by a line search method using the steepest descent direction and the backtracking Armijo line search one of the following three conditions must hold.

(C1) $\nabla f(x^{(k)}) = 0$ for some $k \geq 0$.
(C2) $\lim_{k\to\infty} \nabla f(x^{(k)}) = 0$.
(C3) $\lim_{k\to\infty} f(x^{(k)}) = -\infty$.

The method of steepest descent has the global convergence property, that is, independent of the starting point the sequence of gradients will converge to a stationary point (but that does not mean that the sequence $x^{(k)}$ converges, think of $-\ln(x^{(k)})$!) or the function values diverge and indicate that no minimum exists.

Example 3 A consumer has a utility function $u(x_1, x_2, x_3) = \sqrt{x_1} + 2\sqrt{x_2} + 3\sqrt{x_3}$ over three goods. She can spend \$1 on buying quantities of these three goods, all of which have a price of \$1. After substituting the budget equation, $x_1 + x_2 + x_3 = 1$, into the utility function the consumer wants to maximize $\sqrt{x_1} + 2\sqrt{x_2} + 3\sqrt{1 - x_1 - x_2}$. (We can trivially solve this problem with pencil and paper and find the optimal solution $\left(\frac{1}{14}, \frac{4}{14}, \frac{9}{14}\right)$.) We solve the consumer's optimization problem by minimizing the function $f(x_1, x_2) = -\left(\sqrt{x_1} + 2\sqrt{x_2} + 3\sqrt{1 - x_1 - x_2}\right)$ with a steepest descent method (using the optimal step length in each step). Figure 1 indicates some of the early steps and Table 1 lists details of some of the steps. (To show convergence of variable values and the optimal function value we report six digits for these terms. The search direction and norm of the gradient are converging to zero and so for simplicity we report fewer and not always the same number of digits. We abbreviate numbers like $6.7 \cdot 10^{-8}$ by 6.7(−8).)

The steepest descent method makes good progress in the first few iterations but then slows down considerably. Note the comparatively little change in the values of $x^{(k)}$ during the last 10 to 15 iterations. The figure shows a lot of 'zigzagging' from iterate to iterate.

The behaviour of the steepest descent method in the example is quite typical. As a result the convergence of the method is rather slow. And so, despite having the global convergence property, it is useless in practice. The slow convergence (see Nocedal and Wright 2006, ch. 3) of this method renders it impractical. The convergence problems are essentially due to the reliance on a first-order approximation, which ignores the curvature properties of f. Newton's method takes advantage of a second-order approximation.

Newton Methods

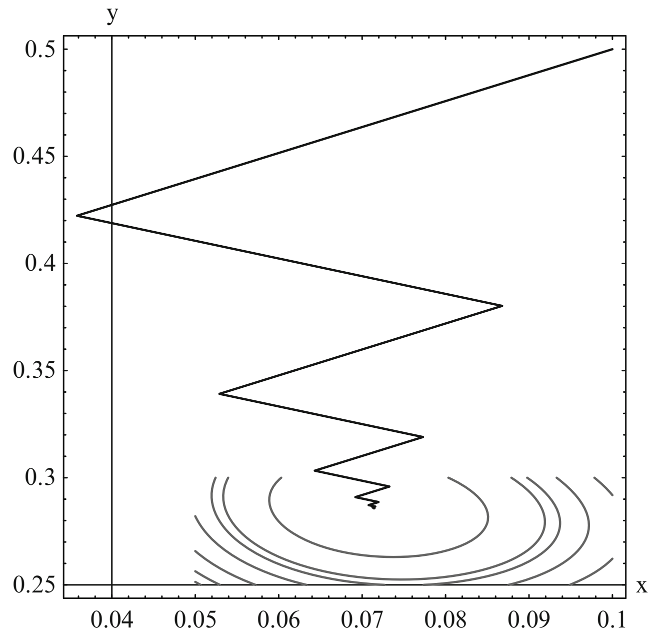
The quadratic approximation q of the objective function f at an iterate $x^{(k)}$ is given by.

$$q(s) = f\left(x^{(k)}\right) + \nabla f\left(x^{(k)}\right)^\top s + \frac{1}{2} s^\top H\left(x^{(k)}\right) s.$$

The first-order condition $q'(s) = 0$ yields the search direction

$$s^{(k)} = -H\left(x^{(k)}\right)^{-1} \nabla f\left(x^{(k)}\right).$$

Only under very strong conditions is Newton's method globally convergent.



Numerical Optimization Methods in Economics, Fig. 1 First steps of a steepest descent method

Numerical Optimization Methods in Economics, Table 1 Steps of a steepest descent method

k	$x_1^{(k)}$	$x_2^{(k)}$	$s^{(k)}$		$\|\nabla f(x^{(k)})\|$	$f(x^{(k)})$
0	0.1	0.5	– 0.7906	– 0.9575	1.2417	– 3.62781
1	0.0358229	0.422272	0.6041	– 0.4988	0.7834	– 3.69734
2	0.0867861	0.380194	– 0.3573	– 0.4328	0.5612	– 3.71804
3	0.0528943	0.339146	0.2503	– 0.2066	0.3245	– 3.73387
4	0.0772951	0.318999	– 0.1321	– 0.1600	0.2075	– 3.73858
5	0.0643195	0.303284	0.0853	– 0.0704	0.1106	– 3.74074
6	0.0732734	0.295891	– 0.0414	– 0.0502	0.0651	– 3.74136
7	0.0691862	0.290940	0.0257	– 0.0212	0.0334	– 3.74157
⋮						⋮
10	0.0715805	0.286543	– 0.0034	– 0.0041	0.0054	– 3.74166
⋮						⋮
15	0.0714140	0.285747	1.64 (– 4)	– 1.35 (– 4)	2.12 (– 4)	– 3.74166
⋮						⋮
20	0.0714288	0.285716	– 5.94 (– 6)	– 7.19 (– 6)	9.33 (– 6)	– 3.74166

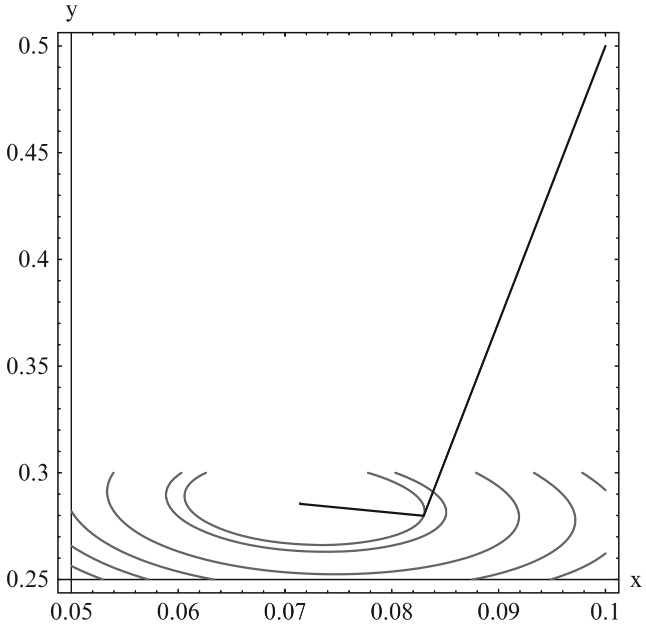
Theorem Suppose that f is continuously differentiable and that ∇f is Lipschitz continuous on $\mathbb{R}^n$. If for the sequence $\{x^{(k)}\}$ of iterates generated by a line search method using the Newton direction and the backtracking Armijo line search the Hessian matrices $H(x^{(k)})$ are positive definite with eigenvalues that are uniformly bounded away from zero, then one of the conditions (C1), (C2), (C3) must hold.

Example 4 We revisit the consumer's optimization problem from Example 3 and minimize the function $f(x_1; x_2) = -\left(\sqrt{x_1} + 2\sqrt{x_2} + 3\sqrt{1 - x_1 - x_2}\right)$ with a Newton method (using the optimal step length in each step). Table 2 lists all the steps of this method and Fig. 2 displays some of the early steps.

Newton's method converges very rapidly. Unlike the steepest descent method it does not

Numerical Optimization Methods in Economics, Table 2 Steps of a Newton method

k	$x_1^{(k)}$	$x_2^{(k)}$	$s^{(k)}$		$\|\nabla f(x^{(k)})\|$	$f(x^{(k)})$
0	0.1	0.5	– 0.0161	– 0.2078	1.2417	– 3.62781
1	0.0829896	0.280	– 0.0128	0.0062	0.1440	– 3.74077
2	0.0714128	0.285450	1.58 (– 5)	2.64 (– 4)	0.0014	– 3.74166
3	0.0714286	0.285714	– 2.27 (– 8)	1.10 (– 8)	3.15 (– 7)	– 3.74166
4	0.0714286	0.285714			5.46 (– 15)	– 3.74166



Numerical Optimization Methods in Economics, Fig. 2 First steps of a Newton method

slow down near the solution, instead we see a quadratic rate of convergence just like in the one-dimensional problem in Example 1.

The condition that the Hessian matrix $H(x^{(k)})$ is positive definite for the entire sequence $\{x^{(k)}\}$ is rarely satisfied for general problems. But if the Hessian is not positive definite then the search direction $s^{(k)}$ may be an ascent instead of a descent direction. The modified Newton methods address this problem by modifying the Hessian matrix $H(x^{(k)})$. These methods choose a search direction

$$s^{(k)} = -\left(H\left(x^{(k)}\right) + M\left(x^{(k)}\right)\right)^{-1} \nabla f\left(x^{(k)}\right),$$

where the matrix $M(x^{(k)})$ is chosen so that $H(x^{(k)}) + M(x^{(k)})$ is 'sufficiently' positive definite. If $H(x^{(k)})$ is sufficiently positive definite itself then, of course, $M(x^{(k)}) = 0$. A proper choice of $M(x^{(k)})$ is crucial for the effectiveness of this approach; see Gould and Leyffer (2002) and Nocedal and Wright (2006) for many more details.

The most tedious task in Newton's method is the computation of the Hessian matrix $H(x^{(k)})$. Therefore, for decades it was fashionable to develop methods, the so-called quasi-Newton methods, that rely on approximations of the exact Hessian matrix. Interest in these methods has somewhat diminished due to the development of automatic differentiation techniques. These techniques allow a very fast and reliable computation of derivatives and so make the task of calculating the Hessian feasible even for large problems. Nocedal and Wright (2006, ch. 6) discuss quasi-Newton methods in detail.

Before we continue our discussion of optimization algorithms we pause for a quick comment on some potential name confusion. In addition to Newton methods for unconstrained optimization there is also a Newton method for solving nonlinear systems of equations. To avoid confusion and for historical reasons the root-finding methods for nonlinear systems of equations are sometimes called Newton–Raphson methods; see Judd (1998) and references therein. In particular, Newton methods for solving unconstrained optimization problems should not be confused with so-called global Newton methods. In economic theory the term 'Smale's global Newton method' appears to be well known. This term refers to a solution method for solving nonlinear systems of equations (see Smale 1976) which is closely related to homotopy continuation methods. Clearly, we could use methods for nonlinear equations to solve the first-order conditions $\nabla f(x) = 0$. This approach, however, does not use other information from the underlying optimization problem and thus is often inefficient. Here we do not discuss methods for solving nonlinear equations, and refer to Allgower and Georg (1979), Judd (1998) and Miranda and Fackler (2002).

Trust Region Methods

Line search methods use an approximation of the objective function f to generate a search direction. Subsequently they determine a suitable step length along this direction. Trust region methods also rely on an approximation of f, but they first define a region around the current iterate in which they trust the approximation to be adequate. Then they simultaneously choose the direction and step length.

For the purpose of our discussion here we consider a quadratic approximation of f around $x^{(k)}$,

$$q_k(s) = f\left(x^{(k)}\right) + \nabla f\left(x^{(k)}\right)^{\top} s + \frac{1}{2} s^{\top} B\left(x^{(k)}\right) s,$$

where $B(x^{(k)})$ is a symmetric approximation of the Hessian matrix $H(x^{(k)})$. Trust region methods do not require the Hessian matrix of the function q_k to be positive definite. Therefore, we could use $B(x^{(k)}) = H(x^{(k)})$. In that case, the algorithm is called a trust region Newton method. Given a trust region radius $\Delta_k > 0$ in each iteration, the algorithm seeks an (approximate) solution to the trust region sub-problem

$$\min_{s \in \mathbb{R}^n} q_k(s) \text{subject to } ||s|| \leq \Delta_k.$$

Before we discuss how we may solve this sub-problem we need to decide on a proper choice for the trust region radius. Note that $q_k(0) - q_k(s^{(k)})$ is the predicted reduction for a step $s^{(k)}$. Similarly, $f(x^{(k)})$– $f(x^{(k)} + s^{(k)})$ is the actual decrease in the objective. The ratio

$$\rho_k = \frac{f\left(x^{(k)}\right) - f\left(x^{(k)} + s^{(k)}\right)}{q_k(0) - q_k\left(s^{(k)}\right)}$$

gives an indication on how well the quadratic approximation predicts the reduction in the function value. Ideally we would like the step $s^{(k)}$ to yield a value of ρk of close to or larger than 1. In that case we accept the step and may possibly increase the radius for the next iteration. If, however, ρk is close to zero or even negative, then we would decrease the trust region radius, set $x^{(k+1)} = x^{(k)}$, and attempt to solve the sub-problem again.

Recall that line search methods do not require the step length to be chosen optimally in order to be globally convergent. Similarly, it is unnecessary and in fact computationally inefficient to solve the trust region sub-problem exactly. Instead, it suffices to search for a step giving a sufficient reduction in q_k. Such a sufficient reduction is achieved by requiring a decrease that is at least as large at that obtained by a step in the direction of steepest descent. The solution to

$$\min_{\alpha \in \mathbb{R}} q_k\left(-\alpha \nabla f\left(x^{(k)}\right)\right) \text{ subject to } ||-\alpha \nabla f\left(x^{(k)}\right)|| \leq \Delta_k$$

yields the Cauchy point

$$s_k^C = -\tau_k \Delta_k \frac{\nabla f\left(x^{(k)}\right)}{||\nabla f\left(x^{(k)}\right)||}$$

where the constant $\tau_k \in (0,1]$ depends on the curvature of q_k and the radius Δ_k; see Nocedal and Wright (2006) for a closed-form solution. The approximate solution $s^{(k)}$ of the trust region subproblem must now satisfy $q_k\left(s^{(k)}\right) \leq q_k\left(s_k^C\right)$.

Theorem Let q_k be the second-order approximation of the objective function *f* at $x^{(k)}$ and let s_k^C be its Cauchy point in the trust region defined by $\|s\| \leq \Delta_k$. Then

$$q_k(0) - q_k\left(s_k^C\right) = f\left(x^{(k)}\right) - q_k\left(s_k^C\right)$$
$$\geq \frac{1}{2}\left\|\nabla f\left(x^{(k)}\right)\right\| \min\left\{\frac{\|\nabla f\left(x^k\right)\|}{1 + \|B(x^{(k)})\|}, \Delta_k\right\}.$$

The theorem has the typical flavour of results on trust region methods. It relates the reduction in the quadratic approximation, $q_k(0) - q_k\left(s_k^C\right)$, to $\|\nabla f(x^{(k)})\|$, which is a measure for the distance to optimality. Once again a global convergence result holds.

Theorem Consider the sequence $\{x^{(k)}\}$ of iterates generated by the described trust region method. Suppose that *f* is twice continuously differentiable and both the Hessian of *f* and the quadratic approximation q_k are bounded for all *k*. Then one of the conditions (C1), (C2), (C3) must hold.

The trust region method based on the Cauchy point is effectively a steepest descent (line search) method where the choice of the step length is bounded by the trust region radius. Therefore, this method also suffers from very poor convergence in practice. Better algorithms start from the Cauchy point and try to improve upon it. There is a variety of such methods that take advantage of additional properties of *f*; see Gould and Leyffer (2002) and Nocedal and Wright (2006). For a comprehensive treatment of trust region methods, see Conn et al. (2000).

Constrained Optimization

Now we consider the constrained optimization problem

$$\begin{array}{lll} \min_{x \in \mathbb{R}^n} & f(x) & \\ \text{s.t.} & g_i(x) \geq 0 i \in I & \text{(NLP)} \\ & h_j(x) = 0 j \in E. & \end{array}$$

We define the feasible region $\mathcal{F}$ of this optimization problem to be the set of all points that satisfy the constraints, so

$$\mathcal{F} = \left\{x \in \mathbb{R}^n | \, g_i(x) \geq 0, \quad i \in I; h_j(x) = 0, j \in E\right\}.$$

Just as for the unconstrained optimization problem, we can define global and local solution. Of course, a desired optimal solution x^* to this optimization problem satisfies $f(x^*) \leq f(x)$ for all $x \in \mathcal{F}$. A point x^* is a local minimizer if it satisfies $f(x^*) \leq f(x)$ for all $x \in \mathcal{N}(x^*) \cap \mathcal{F}$ for some neighbourhood $\mathcal{N}(x^*)$ of x^*. The vector x^* is an isolated local minimizer if there exists a neighbourhood $\mathcal{N}(x^*)$ in which it is the only local minimizer.

The conditions of these definitions, just like their counterparts for unconstrained optimization problems, are pretty much useless for the computation of optimal solutions – with one major exception. The simplex method for solving linear programming problems relies on the comparison of objective function values at some special points in the feasible region. Most other practical numerical methods, however, rely again on optimality conditions. Penalty methods transform the problem (NLP) into (a sequence of) unconstrained optimization problems and then rely on their respective first-order conditions. Many methods rely directly on optimality conditions for constrained optimization. These optimality conditions require that certain degenerate behaviour does not occur at potential minimizers. Conditions that rule out such degenerate points are called ‘constraint qualifications’. These conditions are important but do not always get the proper attention in economics, but see Simon and Blume (1994) for a rigorous treatment. Numerous such constraint qualifications exist; here we just mention one such condition.

The set of constraints that hold with equality at a feasible point $x \in \mathcal{F}$ is called the active set $\mathcal{A}(x)$. Formally,

$$\mathcal{A}(x) = \{i \in I | \, g_i(x) = 0\} \cup E.$$

The linear independence constraint qualification (LICQ) holds at a point $x \in$ F if the gradients of all active constraints are linearly independent. Now we can state the well-known first-order necessary conditions, which most of us learnt as Kuhn–Tucker or Karush-Kuhn-Tucker (KKT) conditions.

Theorem Suppose x^* is a local solution of the problem (NLP) that satisfies the (LICQ). Then there exist unique Lagrange multipliers $v_i^*, i \in I$, and $\lambda_j^*, j \in E$, such that the following conditions are satisfied.

$$\nabla f(x^*) - \sum_{i \in I} v_i^* \nabla g_j(x^*) - \sum_{i \in E} \lambda_i^* \nabla h_j(x^*) = 0, \tag{3}$$

$$g_i(x^*) \geq 0, \quad \text{for all } i \in I, \tag{4}$$

$$h_j(x^*) = 0, \quad \text{for all } i \in E, \tag{5}$$

$$v_i^* g_i(x^*) = 0, \quad \text{for all } i \in I, \tag{6}$$

$$v_i^* \geq 0, \quad \text{for all } i \in I. \tag{7}$$

Again we may ask when we can be assured that a solution to the KKT conditions is actually a solution to the nonlinear optimization problem (NLP). If the feasible region $\mathcal{F}$ is a convex set (see Simon and Blume 1994), and the objective function f is convex on $\mathcal{F}$, then the problem (NLP) is called a convex programming problem, and any local solution is also a (global) solution of (NLP). For example, if the functions h_j, $j \in E$, are all linear and the functions $-g_i$, $i \in I$, are all convex, then $\mathcal{F}$ is a convex set. In this case, if f is convex, too, indeed any solution to the KKT conditions is a solution to (NLP).

Many of the most popular numerical methods for solving nonlinear constrained optimization problems take advantage of the KKT conditions in one form or another. First, however, we describe the basic version of the simplex method for linear programming which does not rely on first-order conditions.

The Simplex Method

When the objective function f and the constraint functions g_i, $i \in I$, and h_j, $j \in E$, are all linear functions in the variables $x \in \mathbb{R}^n$, then the constrained optimization problem is a linear programming problem, or 'linear program' for short. Linear programs have a standard form,

$$\begin{array}{lll} \min_{x \in \mathbb{R}^n} & c^\top x & \\ \text{s.t.} & Ax = b & \text{(LP)} \\ & x \geq 0 & \end{array}$$

where $c \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, and A is an $m \times n$ matrix. We can easily transform any linear programming problem with arbitrary linear inequalities and unbounded variables into this standard form.

The development of the simplex method in the late 1940s (Dantzig 1949) for solving linear programs is generally regarded as the beginning of the modern era of optimization (Nocedal and Wright 2006). The simplex method is, however, not only of historical importance but to this day the perhaps most widely used tool in optimization outside economics. Here we describe the fundamental idea of the basic version of the simplex method.

The system of equality constraints, $Ax = b$, has m equations in the n decision variables. For the LP to be an interesting optimization problem it must be the case that $m < n$. If $m > n$ then either the linear system is overdetermined and so the feasible region is empty and the LP has no solution, or the system can be simplified so that the number of equations does not exceed the number of variables. The same conclusions apply for the case $m = n$ if the matrix A is singular. If $m = n$ and A has full rank, then the feasible region consists of at most one point and the LP is trivial. We can therefore assume that the system of equality constraints is underdetermined, that is, it has fewer equations than variables. Modern computer implementations of the simplex method start with a pre-processing phase, which transforms a given linear programming problem by removing redundancies and possibly even also eliminating some variables.

We can easily calculate some of the solutions to the system $Ax = b$. If we choose m of the n variables and set the remaining $n - m$ variables to zero, then the system reduces to a square system of m linear equations, which can be solved via

Gaussian elimination. The chosen variables for which we solve the system are called ‘basic variables’, while those variables that we set to zero are called ‘non-basic variables’. Solving the m linear equations in the m basic variables can lead to three possible outcomes. First, we may detect that the system has no solution. Second, a solution, called basic solution, may exist and it also satisfies the remaining constraints of the LP, namely the sign restrictions $x \geq 0$. In this case the solution is called a ‘basic feasible solution’. Third, the solution to the linear system may entail a negative value for at least one variable and thus violate the sign restriction. Such a solution is called ‘basic infeasible’. Two basic solutions are called adjacent if their respective sets of basic variables have all but one element in common. The next theorem explains why the basic feasible solutions are of central importance to the linear program.

Theorem If the problem (LP) has a non-empty feasible region, then there is at least one basic feasible solution. If the problem (LP) has an optimal solution then it has the following properties.

1. At least one optimal solution is a basic feasible solution.
2. If (LP) has a unique solution, then this optimal solution is basic feasible.
3. If a basic feasible solution x^* has an objective function value that is not larger than the objective function values at all its adjacent basic feasible solutions, then x^* is a solution of (LP).
4. If the feasible region is bounded and a basic feasible solution x^* has an objective function value that is strictly less than the objective function value at all its adjacent basic feasible solutions, then x^* is the unique solution of (LP).

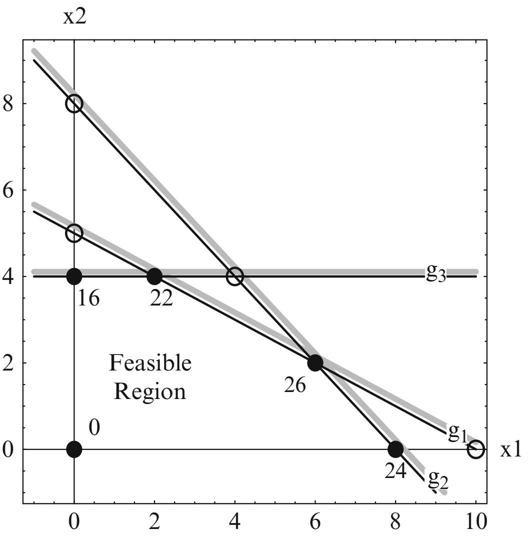
This theorem provides the foundation for the basic approach of the simplex method. According to the first statement of the theorem, if an optimal solution exists then there must be a basic feasible solution that is optimal. Thus, for solving the problem (LP) it suffices to only examine basic feasible solutions. In principle we could now find a solution to the problem (LP) by simply calculating all its basic solutions and then choosing a basic feasible solution with the smallest objective function value. We would not want to do this in practice, however, since the number of possible basic solutions is $\binom{n}{m}$ and thus is huge for many applications. The simplex method prescribes a smart way of searching through the basic feasible solutions. Starting from some basic feasible solution, the simplex searches for another basic feasible solution with a lower objective function value. From a computational standpoint it is much quicker to examine only adjacent basic feasible solutions. The information we have from having solved a linear system in, for example, the variables x_1, x_2, x_3, greatly simplifies finding a solution in the variables x_2, x_3, x_5. Therefore, the simplex considers only adjacent feasible solutions and chooses one of them by exchanging one basic variable against one non-basic variable and solving the resulting system of linear equations. On most (but not all) steps of the method the objective function value decreases. This process repeats itself until the method reaches a basic feasible solution without any adjacent basic feasible solutions having a lower objective function value. The third statement of the theorem (which is a special version of the convex programming property for linear programs) then ensures that the simplex method has found an optimal solution.

We illustrate the basic ideas underlying the simplex method in the following example.

Example 5 Consider the following linear programming problem.

$$\begin{array}{rr} \max_{x_1,x_2} & 3x_1 + 4x_2 \\ \text{s.t.} & x_1 + 2x_2 \leq 10 \\ & x_1 + x_2 \leq 8 \\ & x_2 \leq 4 \\ & x_1 \geq 0 \\ & x_2 \geq 0 \end{array}$$

Linear programming problems with two variables allow a beautiful graphical representation, which greatly helps us to gain some intuition for the simplex method. Figure 3 shows the feasible region of this linear programming problem.



Numerical Optimization Methods in Economics, Fig. 3 Feasible region of the (LP)

This problem has three inequality constraints and two sign restrictions. The introduction of three so-called slack variables transforms the inequalities into equations. This introduction of new variables is just one of several simple transformations that allow us to rewrite any linear programming problem into a linear program in standard form; see Dantzig (1963) or many other linear programming books. Here we obtain the following linear program.

$$\begin{array}{lccccccccccl}
\min_{x_1,x_2,x_3,x_4,x_5} & -3x_1 & - & 4x_2 & & & & & & & & \\
\text{s.t.} & x_1 & + & 2x_2 & + & x_3 & & & & & & =10 \\
 & x_1 & + & x_2 & + & & & x_4 & & & & =8 \\
 & & & x_2 & + & & & & & x_5 & & =4 \\
 & x_1, & & x_2, & & x_3, & & x_4, & & x_5 & & \geq 0
\end{array}$$

This linear program has $n = 5$ variables and $m = 3$ constraints.

Table 3 lists all $\binom{5}{3} = 10$ possible combinations of basic variables, the corresponding basic solution (if it exists), whether this solution is feasible, and the objective function values $z = -3x_1 - 4x_2$ for the basic feasible solutions. For example, the basic solution (4, 4, − 2, 0, 0) is obtained by setting $x_4 = x_5 = 0$ and then solving the remaining three equations

$$x_1 + 2x_2 + x_3 = 10, x_1 + x_2 = 8, x_2 = 4,$$

in the three basic variables x_1, x_2, x_3. This basic solution is infeasible since $x_3 = -2$ violates the non-negativity constraint on this variable. The basic variables x_1, x_3, x_4 lead to the three equations

$$x_1 + x_3 = 10, \quad x_1 + x_4 = 8, \quad 0 = 4,$$

which obviously have no solution. We can relate the nine basic solutions to points in the graph of the feasible region in Fig. 3. The five feasible solutions are represented by disks while the four infeasible solutions are given by circles. We can easily identify the coordinates of the nine indicated points with the values of the original variables x_1 and x_2 in the nine basic solutions. But where are the later introduced slack variables? The values of these variables at a basic solution show us where the corresponding point in the figure is in relation to the three constraints. The basic solution (4,4, − 2,0,0) is represented by the point (4,4) in the graph. This point lies on the lines representing the second and third constraints, since $x_4 = x_5 = 0$, and outside the first constraint, since $x_3 < 0$.

The simplex method quickly solves this problem. Starting from the basic feasible solution that corresponds to the origin in Fig. 3 it moves through the basic feasible solutions ('BFS') listed in Table 4 to find the optimal basic feasible solution (6,2,0,0,2). Figure 4 illustrates the steps of the simplex method. Starting from the point (0,0) it moves upwards to the point (0,4) with an objective function value of $z = -16$, then to (2,4) with $z = -22$ and finally to (6,2) with $z = -26$. The basic feasible solution corresponding to this last point has a strictly lower objective function value than both its adjacent basic feasible solutions at (2,4) and (8,0) and hence it must be the unique optimal solution. In Fig. 4 only the visited points are indicated by disks and the iso-objective function lines for the values $-z$ of the original objective function (from the maximization problem) at these points.

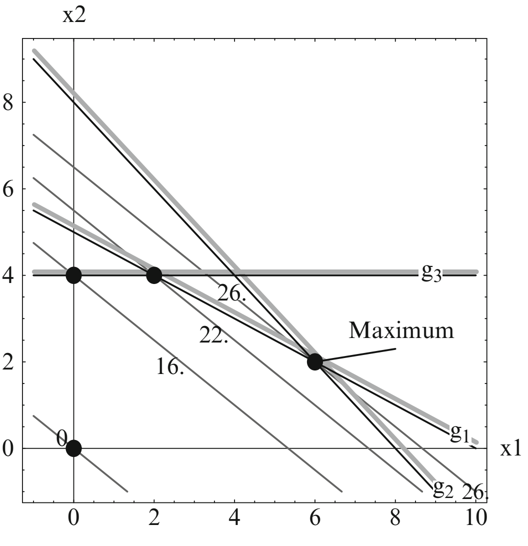
We have conveyed only the basic principle of the simplex method for solving linear programming problems. Of course, an efficient and robust implementation of the simplex algorithm must address many technical details; see

Numerical Optimization Methods in Economics, Table 3 All basic solutions

Basic variables	Basic solution	Property	z
x_1 , x_2 , x_3	(4, 4, – 2, 0, 0)	Infeasible	–
x_1 , x_2 , x_4	(2, 4, 0, 2, 0)	Feasible	– 22
x_1 , x_2 , x_5	(6, 2, 0, 0, 2)	Feasible	– 26
x_1 , x_3 , x_4	–	No solution	–
x_1 , x_3 , x_5	(8, 0, 2, 0, 4)	Feasible	– 24
x_1 , x_4 , x_5	(10, 0, 0, – 2, 4)	Infeasible	–
x_2 , x_3 , x_4	(0, 4, 2, 4, 0)	Feasible	– 16
x_2 , x_3 , x_5	(0, 8, – 6, 0, – 4)	Infeasible	–
x_2 , x_4 , x_5	(0, 5, 0,3, – 1)	Infeasible	–
x_3 , x_4 , x_5	(0, 0, 10, 8, 4)	Feasible	0

Numerical Optimization Methods in Economics, Table 4 Iterates of the simplex method

Basic Variables	BFS	z
x_3 , x_4 , x_5	(0, 0, 10, 8, 4)	0
x_2 , x_3 , x_4	(0, 4, 2, 4, 0)	– 16
x_1 , x_2 , x_4	(2, 4, 0, 2, 0)	– 22
x_1 , x_2 , x_5	(6, 2, 0, 0, 2)	– 26



Numerical Optimization Methods in Economics, Fig. 4 Solving the (LP)

Fletcher (1987) or once again Nocedal and Wright (2006). The classical reference for the theory of the simplex method is the book by Dantzig (1963).

The simplex method is highly efficient on virtually all practical problems, but there do exist pathological problems on which it shows very poor performance. In these worst-case problems the running time of the simplex method grows exponentially in the dimension of the problems. (In a nutshell, the method visits far too many basic feasible solutions until it finds the optimal one.) Therefore, the simplex method is of exponential complexity. Although these examples are irrelevant for practical applications, they generated interest in the development of different algorithms that would show better worst-case running times, in particular, that would have running times that grow only polynomially in the size of the problems. The first linear programming algorithm with polynomial complexity was the ellipsoid method of Khachiyan (1979). Although this method has polynomial complexity it is useless for actual computations, and apparently there has never been a serious practical implementation. The projective algorithm of Karmarkar (1984) started what is nowadays called the 'interior-point revolution'. This algorithm both has polynomial complexity and is of practical use, although the initial claims about its supposedly stellar practical performance were shown to be outrageous. The projective algorithm has long been superseded by more efficient methods, and the field of interior-point methods remains an active area of research to this day.

The Idea of Interior-Point Methods

Primal-dual methods are an important subclass of interior-point methods for solving constrained optimization problems. Here we give a basic outline of such a method for solving linear programs. The Karush–Kuhn–Tucker conditions for a linear programming problem in standard form are as follows.

$$A^{\top}\lambda + s = c \tag{8}$$

$$Ax = b \tag{9}$$

$$x_i s_i = 0, \quad i = 1, 2, \ldots, n \tag{10}$$

$$x \geq 0 \tag{11}$$

$$s \geq 0 \tag{12}$$

These first-order conditions characterize both the optimal solution of the given linear program and of its dual. (See Dantzig 1963, or any book on linear programming for the definition of the dual of a linear program.) That fact motivates the name 'primal-dual' method.

Interior-point methods (approximately) solve a sequence of perturbed problems. Consider the following perturbation of the first-order conditions.

$$A^{\top}\lambda + s = c \tag{13}$$

$$Ax = b \tag{14}$$

$$x_i s_i = \mu, \quad i = 1, 2, \ldots, n \tag{15}$$

$$x > 0 \tag{16}$$

$$s > 0 \tag{17}$$

Observe that the complementarity condition (10) has been replaced by the Eqs. (15) for some positive scalar μ >0. Assuming that a solution ($x^{(0)}$,$\lambda^{(0)}$,$s^{(0)}$) to this system is given for some initial value of $\mu^{(0)} > 0$, interior-point methods decrease the parameter μ and thereby generate a sequence of points ($x^{(k)}$, $\lambda^{(k)}$, $s^{(k)}$) that satisfy the non-negativity constraints on the variables strictly, $x^{(k)}$> 0 and $s^{(k)} > 0$. This property led to the name 'interior-point' method. In the limit, as μ is decreased to zero, a point satisfying the original first-order conditions is reached. The set of solutions to the perturbed system,

$$C = x\{(\mu), \lambda(\mu), s(\mu) | \, \mu > 0\}$$

is called the central path.

The method is rather intuitive at this point. Given an iterate ($x^{(k)}$, $\lambda^{(k)}$, $s^{(k)}$) for some parameter value $\mu^{(k)}$ decrease the parameter to $\mu^{(k+1)} < \mu^{(k)}$ and determine the next iterate ($x^{(k+1)}$, $\lambda^{(k+1)}$, $s^{(k+1)}$). Implementing this method requires handling of many details. For example, it is often difficult to find a feasible starting point ($x^{(0)}$,$\lambda^{(0)}$, $s^{(0)}$) of the perturbed system. The most important step in the method is to solve the system (13)–(15) in each iteration (while maintaining the inequalities (16, 17)). Observe that this system consists of $2n + m$ linear and bilinear equations in as many variables. We can apply a nonlinear equations solver to this model. A popular approach is to use Newton's method for solving nonlinear systems of equations; see Judd (1998) or Miranda and Fackler (2002). The difficulty is to maintain the strict non-negativity constraints on the variables $x^{(k+1)}$ and $s^{(k+1)}$. An alternative approach for solving the parameterized system of equations is the application of path-following methods; see Nocedal and Wright (2006). Intuitively we can think of interior-point methods to be closely related to homotopy continuation methods for solving nonlinear systems of equations; see Allgower and Georg (1979).

Example 6 We revisit the linear program from Example 5. The perturbed first-order conditions (13)–(17) for this (LP) are as follows.

Numerical Optimization Methods in Economics, Table 5 Solutions x*(μ) for small $\boldsymbol{\mu}$

μ	$x_1(\mu)$	$x_2(\mu)$	$x_3(\mu)$	$x_4(\mu)$	$x_5(\mu)$
1	5.9775	1.5451	0.9323	0.4774	2.4549
0.5	6.0305	1.7311	0.5073	0.2384	2.2689
0.1	6.0029	1.9478	0.1014	0.0492	2.0522
0.01	6.0000	1.9950	0.0100	0.0050	2.0050
0	6	2	0	0	2

$$\begin{aligned}
y_1 + y_2 + s_1 + 3 &= 0\\
2y_1 + y_2 + y_3 + s_2 + 4 &= 0\\
y_1 + s_3 &= 0\\
y_2 + s_4 &= 0\\
y_3 + s_5 &= 0\\
x_1 + 2x_2 + x_3 - 10 &= 0\\
x_1 + x_2 + x_4 - 8 &= 0\\
x_2 + x_5 - 4 &= 0\\
x_i \cdot s_i = \mu, \quad i &= 1, 2, \ldots, n\\
x_1, x_2, \ldots, x_5 &> 0\\
s_1; s_2, \ldots, s_5 &> 0
\end{aligned}$$

Table 5 displays the values for the variables x_1, x_2,. . ., x_5 at some points on the central path for small values of μ. We observe how the central path moves through the interior of the feasible region, see Fig. 3, and converges to the optimal solution as $\mu \rightarrow 0$.

By now the conceptual differences between the simplex method and interior-point methods are transparent. In geometric terms, the simplex method moves on specific points around the boundary of the feasible region until it finds a corner point corresponding to an optimal basic feasible solution. Interior-point methods move through the interior (or some methods even through the exterior) of the feasible region but they do not move within the boundary. Instead they approach the boundary only in the limit. In computational terms, the typical iteration of an interior-point method is relatively expensive to compute but can make significant progress towards the solution. Conversely, an iteration of the simplex method is relatively inexpensive but the method often requires a larger number of iterations.

Obviously the question arises of which of these two basic approaches is better for solving linear programs in practice. The answer depends very much on the nature of the problem. Currently the best available computer programs are efficient implementations of the dual simplex method (a special variant of the described standard simplex method) and primal–dual interior-point methods. Simplex method computer programs are usually faster on problems of small or medium size (say, of fewer than a million variables and constraints) while interior-point methods tend to do better on many but certainly not all large problems. If the user has significant prior information about the optimal solution, such as a good initial guess for an optimal basic feasible solution, then the simplex method is often much faster. The reason for this is that the simplex method is much easier to 'warm-start' than interior-point methods. In summary, interior-point methods and the simplex method are both important and useful algorithms for solving linear programs in practice.

Before we turn to interior-point methods for nonlinear optimization problems, we outline the basic concepts of another class of optimization algorithms. Penalty methods are quite intuitive and some of their ideas are relevant for interior-point methods but they are also of interest on their own.

Penalty Methods

The basic idea of penalty methods is to replace the constrained optimization problem (NLP) by an unconstrained optimization problem and to solve the new problem instead. The objective function for the new unconstrained problem is the original objective plus a new term for each constraint. The new term is zero when the original constraint is satisfied but is positive if the original constraint is violated. The simplest and perhaps most intuitive penalty function is the quadratic penalty function.

To start we consider a nonlinear optimization problem with only equality but no inequality constraints,

$$\begin{aligned} &\min_{x \in \mathbb{R}^n} f(x) \\ &\text{s.t.} h_j(x) = 0 \quad j \in E. \end{aligned}$$

For such a problem we can define a penalty function

$$Q(x;\mu) = f(x) + \mu\left(\sum_{j \in E} h_j^2(x)\right)$$

with a penalty parameter $\mu > 0$. The idea of the penalty function method is to minimize the function Q for increasing values of μ. Observe that the function Q inherits its differentiable properties from the functions f and h_j, $j \in E$, of the original problem, and so we can use unconstrained optimization methods for minimizing $Q(x;\mu)$. In addition, as we generate a sequence $\mu^{(k)}$, $k = 0,1,2,\ldots$, we can use the previously calculated minimizer $x^{(k)}(\mu^{(k)})$ as initial guesses for the problem with $\mu^{(k+1)}$. This intuitive approach has a strong theoretical foundation, as the following theorem reveals; see Nocedal and Wright (2006).

Theorem Consider a sequence $\{\mu^{(k)}\}$ of penalty parameters with $\mu^{(k)} \to \infty$. Suppose that $x^{(k)}$ is the exact global minimizer of $Q(x;\mu_k) = f(x) + \mu \left(\sum_{j \in E} h_j^2(x)\right)$. Then every limit point x^* of the sequence $\{x^{(k)}\}$ is a global solution of the (NLP).

Although this result is nice from a theoretical viewpoint, it does not directly apply to practical applications. Of course, we typically cannot determine the exact minimizer of the penalty function and have to account for errors in the numerical approximation. The discussion in Nocedal and Wright (2006) shows that things get more complicated in practice once we allow for approximation errors. In addition, the penalty function may have many other stationary points that are not global or even local minimizers. The penalty function may even be unbounded if the penalty parameter μ is too small. At the other extreme, for very large values of μ the unconstrained minimization problem becomes more difficult, and the Hessian of Q gets ill-conditioned. All kinds of numerical problems arise that need to be carefully addressed in robust computer implementations of the quadratic penalty method; see Nocedal and Wright (2006).

For the general problem (NLP) with inequality and equality constraints we can define the penalty function as

$$\begin{aligned} &Q(x;\mu) = f(x) \\ &+ \mu\left(\sum_{i \in I} (\max(-g_i(x), 0))^2 + \sum_{j \in E} h_j^2(x)\right). \end{aligned}$$

Now, however, things get more complicated since Q will typically not be twice differentiable. As a result the new unconstrained problem becomes more difficult to solve.

In addition to the quadratic penalty method several other such approaches exist and are used in practice. Nocedal and Wright (2006) describe non-differentiable penalty functions and the augmented Lagrangian method. Here we finish our discussion with an illustration of the quadratic approach.

Example 7 Consider a simple example of the classical portfolio optimization problem (Markowitz 1952). An investor wants to allocate her entire wealth across three securities with respective expected returns of 4 per cent, 8 per cent and 12 per cent. If she invests the respective portions x_1, x_2, x_3 in the three assets, then the variance of such a portfolio is $x_1^2 + 5x_2^2 + 3x_2x_3 + 10x_3^2$. The investors wants to minimize this variance under the condition that the expected return of her portfolio is at least 9 per cent. To simplify this illustration of the quadratic penalty method we exploit the fact that at the optimal solution the lower bound on the expected return is binding and thus write the investor's portfolio allocation problem as a nonlinear optimization problem with only equality constraints.

$$\begin{aligned} &\min x_1, x_2, x_3 \;\; x_1^2 + 5x_2^2 + 3x_2x_3 + 10x_3^2 \\ &\text{s.t.} \quad x_1 + x_2 + x_3 - 1 = 0 \\ &4x_1 + 8x_2 + 12x_3 - 9 = 0 \end{aligned}$$

Numerical Optimization Methods in Economics, Table 6 Solutions $x^*(\mu)$ for large $\boldsymbol{\mu}$

μ	$x_1^*(\mu)$	$x_2^*(\mu)$	$x_3^*(\mu)$
1	0.78547	0.30659	0.26008
10	0.42484	0.35361	0.36870
100	0.21880	0.37632	0.42571
1000	0.18852	0.37962	0.43403
10,000	0.18535	0.37996	0.43490
∞	0.185	0.38	0.435

The quadratic penalty function for the investor's portfolio optimization problem is

$$\begin{aligned}Q(x;\mu) = {} & x_1^2 + 5x_2^2 + 3x_2x_3 + 10x_3^2 \\ & + \mu\Big((x_1+x_2+x_3-1)^2 + (4x_1+8x_2+12x_3-9)^2\Big).\end{aligned}$$

We can easily solve the unconstrained problem with the basic Newton method as described in section "Newton methods". Table 6 shows the solution to the unconstrained minimization of the penalty function for increasing values of μ.

Observe that the nonlinear optimization problem in this example has a quadratic objective function and linear constraints. Such optimization problems constitute a special and important subclass of problems called quadratic programs. Their special properties give rise to efficient solution methods, and we would not want to solve large quadratic programs with a penalty method. Nocedal and Wright (2006) present several algorithms for quadratic programming. Since solving quadratic programs is comparatively easy, an integral part of some algorithms for more general nonlinear optimization problems, such as the sequential quadratic programming methods, is to repeatedly solve quadratic programs that are derived as approximations for the more general problem.

The Logarithmic Barrier Method

Logarithmic barrier methods are a particular type of interior-point methods for the solution of nonlinear optimization problems. We illustrate the basic idea of these methods for an inequality-constrained minimization problem.

$$\begin{array}{ll}\min_{x\in\mathbb{R}^n} & f(x) \\ \text{s.t.} & g_i(x) \geq 0 \quad i \in I.\end{array}$$

We can combine the objective function and the constraints and define a penalty function for this optimization problem by

$$P(x;\mu) = f(x) - \mu \sum_{i \in I} \ln g_i(x),$$

where $\mu > 0$ is called the barrier parameter and the expression $\sum_{i \in I} \ln g_i(x)$ is called a logarithmic barrier function. Each logarithmic term – $\ln g_i(x)$ tends to infinity as x approaches the boundary given by $g_i(x) \geq 0$ from the interior of the feasible region. This effect of the logarithmic terms will decrease as the barrier parameter μ becomes smaller. The idea of the logarithmic barrier method is now to let the parameter μ converge to zero. Under some conditions the optimal solution $x^*(\mu)$ of the unconstrained optimization problem $\min_{x \in \mathbb{R}^n} P(x;\mu)$ converges to the optimal solution of the original constrained optimization problem as μ tends to zero. Note that the logarithm ensures that $g_i(x^*(\mu)) > 0$ for all $\mu > 0$, that is, the solution to the unconstrained minimization problem is in the strict interior of the original constraints. This property represents a crucial distinction between this variant of an interior-point method and an active-set method such as the simplex method, which always tracks the set of binding constraints at a given iterate.

Observe that the first-order conditions for the penalty function problem are given by

$$\nabla_x P(x;\mu) = \nabla f(x) - \sum_{i \in I} \frac{\mu}{g_i(x)} \nabla g_i(x) = 0.$$

Now define for all $i \in I$

$$v_i(\mu) := \frac{\mu}{g_i(x)}.$$

Numerical Optimization Methods in Economics, Table 7 Solutions $x^*(\mu)$ for small μ

μ	$x_1^*(\mu)$	$x_2^*(\mu)$	$x_3^*(\mu)$
1	0.0421124	0.168450	0.379012
0.5	0.0547198	0.218879	0.492478
0.1	0.0677112	0.270845	0.609401
0.01	0.0710478	0.284191	0.639430
0.005	0.0712379	0.284952	0.641141

Note that since $\mu > 0$ by definition we have that $v_i(\mu)>0$. Thus, at a stationary point of the penalty function the following conditions hold.

$$\nabla f(x) - \sum_{i\ \in I} v_i \nabla g_i(x) = 0 \; g_i(x) - s_i = 0 \quad \text{for all } i \in \ I$$
$$v_i s_i = \mu \text{ for all } i \in \ I v_i > 0 \quad \text{for all } i \in \ I s_i > 0 \quad \text{for all } i \in \ I.$$

This set of conditions is just the primal-dual interior-point conditions for our original problem. We see that conditions (13)–(17) are just the specialization of these conditions for the linear programming model. And just like in the illustration of the section "The idea of interior-point methods" we are interested in taking the parameter μ to zero. Unfortunately we do not have the space here to properly state a formal theorem. To make a long story short, under a few additional technical conditions, most notably second-order conditions of optimality, the following statements hold for a local solution x^* at which the KKT conditions are satisfied for some Lagrange multipliers v^*.

1. The local minimizer $x^*(\mu)$ of $P(x;\mu)$ in some neighbourhood of x^* with $\lim_{\mu\downarrow 0} x^*(\mu) = x^*$ uniquely defines a continuously differentiable vector function $x^*(\mu)$ for all sufficiently small μ.
2. The function $x^*(\mu)$ yields Lagrange multipliers $v(\mu)$ satisfying $\lim_{\mu\downarrow 0} v(\mu) = v^*$ where $v^* g_i(x^*) = 0$.

An algorithm for solving the constrained problem is apparent now. For a given value of μ solve the unconstrained optimization problem with the objective function P. Then reduce μ stepwise to zero and follow the path of solutions $x^*(\mu)$. In the limit we can find the local solution x^* of the original problem. While this approach works in principle, it entails various difficulties. For example, the Hessian matrix of P becomes ill-conditioned for small values of μ. For this and many other technical issues see Gould and Leyffer (2002). Here we just illustrate the fundamental idea with an example.

Example 8 We revisit the consumer's utility maximization problem from Example 3 once again. The consumer has a utility function $u(x_1, x_2, x_3) = \sqrt{x_1} + 2\sqrt{x_2} + 3\sqrt{x_3}$over three goods and faces the budget constraint $x_1 + x_2 + x_3 \leq 1$. We formulate the consumer's problem as the constrained minimization problem

$$\begin{array}{cc} \min_{x_1, x_2, x_3} & -\left(\sqrt{x_1} + 2\sqrt{x_2} + 3\sqrt{x_3}\right) \\ \text{s.t.} & 1 - x_1 - x_2 - x_3 \geq 0. \end{array}$$

The unconstrained function including a logarithmic barrier function for this minimization problem is

$$\begin{aligned} P(x_1, x_2, x_3; \mu) = &-(\sqrt{x_1} + 2\sqrt{x_2} + 3\sqrt{x_3}) \\ &- \mu \ \ln \ (1 - x_1 - x_2 - x_3). \end{aligned}$$

Table 7 displays solutions to this unconstrained problem for a few values of μ. Note that as $\mu \to 0$ the optimal solution approaches the optimal solution of the original utility maximization problem.

In all our examples so far we ignored the sign restriction of the variables. We could do that since the utility functions exhibit an Inada property, that is, $\lim_{x_i \to 0} \frac{\partial u}{\partial x_i} = +\infty$, and so we hope that a solver starting at a strictly positive solution will only iterate through such solutions (although we have to be careful in practice). But, of course, we can easily take the non-negativity constraints explicitly into account and consider the following problem.

Numerical Optimization Methods in Economics, Table 8 Solutions $x^*(\mu)$ for small μ

μ	$x_1^*(\mu)$	$x_2^*(\mu)$	$x_3^*(\mu)$
1	0.219696	0.270563	0.331757
0.5	0.195664	0.278956	0.389721
0.1	0.124696	0.286370	0.543846
0.01	0.0789861	0.285758	0.630009
0.005	0.0753165	0.285727	0.636309

$$\begin{aligned} \min_{x_1,x_2,x_3} \quad & -\left(\sqrt{x_1}+2\sqrt{x_2}+3\sqrt{x_3}\right) \\ \text{s.t.} \quad & 1-x_1-x_2-x_3 \geq 0. \\ & x_1 \geq 0 \\ & x_2 \geq 0 \\ & x_3 \geq 0 \end{aligned}$$

Note that the condition (LICQ) is always satisfied since not all four constraints can be satisfied simultaneously. As long as three constraints are binding (LICQ) holds. The unconstrained function including a logarithmic barrier function for this minimization problem is

$$\begin{aligned} P(x_1,x_2,x_3;\mu) = & -\left(\sqrt{x_1}+2\sqrt{x_2}+3\sqrt{x_3}\right) \\ & -\mu(\ln(1-x_1-x_2-x_3)+\ln(x_1)+\ln(x_2)+\ln(x_3)). \end{aligned}$$

Table 8 displays solutions to this unconstrained problem for a few values of μ. Again we observe that $x^*(\mu) \rightarrow x^*$ as $\mu \rightarrow 0$.

Strangely enough, some of the modern and best interior-point algorithms are based on work predating Karmarkar (1984). For example, Frisch (1955) had already proposed an interior-point method based on logarithmic barrier functions for solving linear programs. A full early history with many results on barrier functions is Fiacco and McCormick (1968).

Sequential Quadratic Programming

Sequential quadratic programming (SQP) methods are among the most effective constrained optimization techniques, particularly when nonlinear constraints are present. These algorithms belong to the class of active-set methods that keep track of the binding constraints at each step. For a description of the basic ideas we consider a minimization problem with only equality constraints. (But these methods are much more widely applicable.)

$$\min_{x\in\mathbb{R}^n} f(x) \quad \text{s.t.} \quad h_j(x)=0, \quad j\in E. \tag{18}$$

The KKT conditions for this problem are as follows.

$$\nabla f(x) - \sum_{j\in\ E} \lambda_j \nabla h_j(x) = 0, \tag{19}$$

$$h_j(x)=0, \quad j\in E. \tag{20}$$

These conditions are a system of $n + m$ nonlinear equations in the n variables x and the m Lagrange multipliers λ. Newton's method for solving nonlinear equations is now a natural approach for solving this system. The Jacobian of the left-hand side of the system (19)–(20) is given by

$$\begin{bmatrix} H(x) - \sum_{j\in E}\lambda_j H_j(x) & -A(x)^\top \\ A(x) & 0 \end{bmatrix},$$

where the matrix $A(x)^\top = [\nabla h1\ (x),\ldots, \nabla h_J\ (x)]$ is the collection of the gradient vectors of all constraints $h(x) = (h_j(x))_{j\ \in\ E=\{1,2,\ldots,\ J\}}$. The matrix $H_j(x)$ denotes the Hessian matrix of the constraint function h_j at the point x. For a given point $(x^{(k)},\lambda^{(k)})$ the Newton step is then determined by the linear system

$$\begin{bmatrix} H\left(x^{(k)}\right) - \sum_{j\in E}\lambda_j^{(k)} H_j\left(x^{(k)}\right) & -A\left(x^{(k)}\right)^\top \\ A\left(x^{(k)}\right) & 0 \end{bmatrix} \begin{bmatrix} s_x^{(k)} \\ s_\lambda^{(k)} \end{bmatrix} = -\begin{bmatrix} \nabla f\left(x^{(k)}\right) - \sum_{j\ \in\ E}\lambda_j^{(k)} \nabla h_j\left(x^{(k)}\right) \\ h\left(x^{(k)}\right) \end{bmatrix}$$

resulting in the new iterate $\left(x^{(k+1)},\lambda^{(k+1)}\right) = \left(x^{(k)}+s_x^{(k)},\lambda^{(k)}+s_\lambda^{(k)}\right)$. Note that this last system is equivalent to the following linear system.

$$\begin{bmatrix} H\left(x^{(k)}\right) - \sum_{j\,\in\,E}\lambda_j^{(k)} H_j\left(x^{(k)}\right) & A\left(x^{(k)}\right)^{\top} \\ A\left(x^{(k)}\right) & 0 \end{bmatrix} \begin{bmatrix} s_x^{(k)} \\ -\lambda^{(k+1)} \end{bmatrix}$$
$$= -\begin{bmatrix}\nabla f\left(x^{(k)}\right) \\ h\left(x^{(k)}\right)\end{bmatrix}$$

Now consider the following quadratic optimization problem (QP).

$$\min_{s\in\mathbb{R}^n} \frac{1}{2} s^{\top}\left(H(x) - \sum_{j\,\in\,E} \lambda_j H_j(x)\right) s + \nabla f^{\top}(x) s$$
$$\text{s.t.}\quad A(x)s + h(x) = 0$$

The left-hand side of the constraints are a first-order (Taylor) approximation of the constraint function h of the original optimization problem. The objective function of (QP) is a second-order approximation of the difference $f(x+s) - f(x)$. The KKT conditions for the problem (QP) are as follows.

$$\left(H(x) - \sum_{j\,\in\,E} \lambda_j H_j(x)\right) s + \nabla f(x) - A^{\mathrm{T}}(x) v = 0 \tag{21}$$

$$A(x)s + h(x) = 0 \tag{22}$$

Observe that these KKT conditions at a point $(x^{(k)}, \lambda^{(k)})$ are equivalent to (21). Solving the first-order conditions of the original optimization problem with Newton's method is, under some technical conditions, equivalent to solving the quadratic optimization problem (QP). A Newton step at a given point $(x^{(k)}, \lambda^{(k)})$ is the same as solving the (QP) at this point. The idea of SQP methods is now to repeatedly solve this quadratic problem to generate a sequence of iterates that converges to a local solution of the original problem. Various good methods for solving quadratic optimization problems exist and can be applied to the problem (QP). Moreover, when combined with line search or trust region methods the approach has useful global convergence properties. Gould and Leyffer (2002) and Nocedal and Wright (2006) discuss details of line search and trust region SQP methods.

Global Optimization

We emphasized repeatedly that most practical algorithms for solving nonlinear optimization problems search for a solution only to the (necessary) first-order conditions, that is, they search for a local solution. Unless we are solving a convex programming problem or an unconstrained minimization problem of a convex function, we often cannot be sure that a computed local solution is indeed an approximate solution to the problem at hand; recall Example 2. Only occasionally other additional knowledge, perhaps some particular property of an underlying economic model, may assure us that we found an optimal solution. Obviously it would be helpful to have methods for general non-convex problems that may not, or are at least less likely to, get stuck in only locally optimal solutions. Here we lay out two approaches for global optimization. We describe the basic ideas of some popular metaheuristics and, subsequently, the very promising area of research in polynomial optimization, which is likely going to produce powerful tools for economic problems.

Metaheuristics

Metaheuristics provide a general framework and basic guidelines for developing specific heuristics for solving optimization problems. While the underlying principles are very general, typically a method must be carefully tailored in order to obtain an effective algorithm for the special problem at hand. Most metaheuristics were originally developed for solving discrete optimization problems, such as integer or combinatorial problems. Their principal ideas can also be applied to come up with heuristics for continuous nonlinear optimization problems.

The central problem of most nonlinear optimization methods is the possibility of getting stuck at a locally optimal solution. Many methods allow only for iterative steps that lead to an improvement in the objective function value, but, for an exception, see the discussion on nonmonotone techniques in Conn et al. (2000) and Nocedal and Wright (2006). Such methods cannot get away from a locally optimal solution. In order to escape from such a local solution we must allow our search procedure, at least sometimes, to move into a non-improving search directions; that is, temporarily the objective function value of the sequence of iterates may increase (in a minimization problem). Three metaheuristics that are supposed to escape local solution are tabu search, simulated annealing, and genetic algorithms. The latter two methods are examples of stochastic approaches for optimization. Here we give a description of the basic ideas underlying these three methods and refer to Brandimarte (2006), Judd (1998) and the citations in those books for details.

Tabu Search

The choice of non-improving search directions must be carefully managed to avoid repeatedly returning to a previously found optimal solution. Such cycling may occur if, after a non-improving step away from a local solution, the algorithm takes an improving step and immediately returns to the previously found local solution. A tabu search procedure imposes at every iteration a list of search directions that the algorithm is not allowed to pursue. For example, if the method just took a step in the direction $s^{(k)}$ then it may not be allowed to examine a neighbourhood of search directions around $-s^{(k)}$ for the next few iterations. In order to avoid memory problems in practical implementations, the tabu list usually consists only of the most recent steps taken. Of course, many technical issues need to be addressed to obtain a robust and efficient algorithm. The treatment of these issues usually depends greatly on the specific problem.

Simulated Annealing

Simulated annealing is another metaheuristic that helps an algorithm to escape from locally optimal solutions. Instead of choosing only iterates that decrease the objective function in a minimization problem, simulated annealing methods also accept with some probability new iterates that increase the objective. The probability of accepting an iterate $x^{(k+1)}$ if $f(x^{(k+1)}) > f(x^{(k)})$ is

$$e^{-\frac{f\left(x^{(k+1)}\right)-f\left(x^{(k)}\right)}{T}}$$

with a parameter $T>0$. Simulated annealing methods typically start out with a fairly large value for T and then decrease it to 0. Observe that for large values of T the heuristic is likely to accept non-decreasing iterates, and so it allows the method to explore the feasible region. As T decreases the probability of acceptance of non-decreasing iterates of a fixed size also decreases. In the limit $T \to 0$ the method allows only iterates that decrease the objective function value. The perhaps simplest rule for reducing T is to start from a high value T_0 and then to set

$$T_{l+1} = \alpha T_1 \quad \text{for some } 0 < \alpha < 1.$$

The basic ideas of simulated annealing are derived from an analogy of minimization with the physical annealing process of slowly cooling metals in order to reach a strong low-energy solid state. This analogy motivates the particular probability function for accepting increasing iterates and explains why the parameter T is called the temperature of the process. The rule of decreasing T is analogously called the cooling schedule. The earliest applications of simulated annealing were combinatorial problems; see Kirkpatrick et al. (1983) as well as Cerny (1985).

Genetic Algorithms

Genetic algorithms are derived from the analogy of finding better and better solutions with the theory of biological evolution of selecting fitter and fitter members of a species. As a result the literature on genetic algorithms uses terminology from evolutionary biology. Iterates in tabu search and simulated annealing algorithms are a single point. Contrary to that, genetic algorithms work with a set (‘generation’) of several current points. A genetic algorithm constructs a sequence of such sets. In a given iteration the objective function is evaluated at the points in the set (‘fitness of a member’). The method then chooses elements of the set in a probabilistic fashion in order to build new elements for the next set. Usually the probability of an element being chosen is the higher the better its objective function value. Several ways to construct new elements exist. A standard operation is the so-called crossover. Given two elements ('parents') $x^{(k)}$ and $y^{(k)}$ in the set the crossover operation leads to

$$x^{(k+1)} = \left(x_1^{(k)}, \ldots, x_l^{(k)},\ y_{l+1}^{(k)}, \ldots, y_n^{(k)}\right), \tag{23}$$

$$y^{(k+1)} = \left(y_1^{(k)}, \ldots, y_l^{(k)},\ x_{l+1}^{(k)}, \ldots, y_n^{(k)}\right), \tag{24}$$

where the method chooses some arbitrary break point l in the n-dimensional vectors. The idea behind crossover is to preserve some parts of the original elements and at the same time generate quite arbitrarily new elements ('children') that are far away from the original ones, and thereby to escape local solution. Another type of operation aimed at achieving this goal is to randomly exchange an element in a member $x^{(k)}$ by another value ('mutation'). While these approaches have proven useful in combinatorial optimization, it is quite apparent that they may run into severe difficulties for constrained problems. Many technical details must, therefore, be resolved before these ideas yield a useful heuristic approach for solving an optimization problem.

The monograph by Holland (1975) popularized genetic algorithms. The basic ideas of computer simulations of evolution are much older.

Any heuristic method derived from a metaheuristic will always be an ad hoc approach to the problem at hand. Just like the standard methods of nonlinear optimization presented in this article, they are not guaranteed to find the solution of a problem. And, while such heuristics have proven useful in discrete optimization, they are generally regarded as inferior to the modern standard optimization techniques for continuous optimization. An economist's first choice of a solution method for a continuous optimization problem, particularly when nonlinear constraints are present, should always be one of the standard methods.

Polynomial Functions

A substantial number of prominent economic models involves only polynomial functions, equations or inequalities. Even problems that at first appear to be non-polynomial can sometimes be transformed into having only polynomial expression. For example, the first-order conditions for the standard log-utility maximization problem

$$\max_{x \in \mathbb{R}^n} \sum_{i=1}^{n} \ln(x_i) \quad \text{s.t.} \quad \sum_{i=1}^{n} p_i(x_i - \omega_i) = 0$$

for prices $p_{t.}$ and endowments $\omega_i\ i = 1,\ldots, n$, can be written in polynomial form,

$$1 - \lambda p_i x_i = 0, \quad i \ \in \{1, \ldots, n\}, \tag{25}$$

$$\sum_{i=1}^{n} p_i(x_i - \omega_i) = 0, \tag{26}$$

where λ denotes the Lagrange multiplier. Polynomial functions and equations can be analysed using tools from computational algebraic geometry (Cox et al. 1997). Global optimization with polynomials is an active field of research in mathematics; see, for example Lasserre (2001), Parrilo and Sturmfels (2003), and the book by Sturmfels (2002) and the citations therein. It is possible (at least in theory) to compute all local minima of polynomial functions. Similarly, it is possible to compute all solutions to a polynomial system of equations. With further expected advances in the theory of polynomial optimization and ever increasing speed of modern computers, these tools will soon have an impact in economics. For first results see computation of general equilibria (new developments).

Popular Optimization Software in Economics

This section covers software packages and modelling languages that are frequently used in economics to solve optimization problems. This list is by no means exhaustive, and many other software products for solving optimization problems exist.

Perhaps the most popular software for numerical work in economics is MATLAB (MATLAB is a registered trademark of The MathWorks, Inc.). Computational economics and finance textbooks such as Brandimarte (2006), Kendrick et al. (2006) and Miranda and Fackler (2002) use MATLAB to solve economic problems. Other popular packages include GAUSS (GAUSS is a registered trademark of Aptech Systems, Inc.) and Mathematica (Mathematica is a registered trademark of Wolfram Research, Inc.) All three languages offer solvers for nonlinear optimization problems, which are continuously enhanced to solve larger and more difficult problems. Here we just list a few features of these software packages.

MATLAB has an optimization toolbox containing routines for solving both unconstrained and constrained nonlinear optimization problems. Methods for unconstrained problems include quasi-Newton and trust region techniques. The solvers for constrained optimization include an SQP method. MATLAB also has specialized methods for nonlinear least square problems; however, most of these solvers are considered to be of only mediocre quality. Much better solvers in MATLAB are available through the NAG toolbox (NAG is a registered trademark of The Numerical Algorithms Group, Inc.) The NAG Foundation Toolbox provides access to the large set of numerical routines contained in the Fortran-based NAG Foundation Library, which contains routines for constrained and unconstrained optimization.

The high-level matrix programming language GAUSS includes an applications module for constrained optimization that uses an SQP method in conjunction with several line search methods or a trust region method. GAUSS has some specialized modules for constrained maximum likelihood problems. For Mathematica there exists a global optimization package, which contains various functions for optimization. These functions are designed to search for global optima for problems with hundreds of variables. The monograph by Bhatti (2000) comes with an optimization toolbox for Mathematica that includes all the methods presented in this article.

These high-level languages are popular in economics because they are easy to learn and quickly facilitate solving problems of moderate size. For larger problems with thousands or even hundreds of thousands of variables, however, they are not reliable and certainly too slow. Economists interested in solving large problems need to use alternative software. An excellent alternative is the use of algebraic modelling languages.

The General Algebraic Modeling System (GAMS) is a high-level modelling language designed for mathematical programming and optimization; see Rosenthal (2006) for a user's guide. GAMS consists of a language compiler and a family of integrated high-performance solvers. GAMS is tailored for complex, large-scale modelling applications, and allows the user to build large models. It has a long history of successful applications in economics, particularly in solving large-scale computable general equilibrium (CGE) models. AMPL (Fourer et al. 2003) is an algebraic modelling language for mathematical programming, which allows users to set up and solve a great variety of optimization problems. The user has access to many popular and sophisticated solvers.

An exciting environment for solving optimization problems is the Network-Enabled Optimization System (NEOS); see Czyzyk et al. (1998) and Ferris et al. (2000). NEOS is an optimization site that allows users to submit optimization problems over the Internet. The user does not need to download any solver but can just send optimization problems to NEOS and choose from a list of solvers. NEOS has access to many of the most current and powerful optimization routines. NEOS returns a solution and some runtime statistics to the user. Unfortunately, NEOS has been largely ignored by many economists.

Mathematical Programs with Equilibrium Constraints

Mathematical programs with equilibrium constraints (MPECs) are currently at the frontier of numerical analysis. Economic models that can be classified as 'leader-follower' games are examples of MPECs. Suppose that the economic variables can be partitioned into x, those chosen by the 'leader' (government, employer, market maker, mechanism designer, and so on), and y, those chosen by the 'followers' (taxpayers, employees, traders, and so on) or determined in equilibrium (such as price). Suppose that the leader's payoff is $f(x, y)$ and that the equilibrium value y given x is represented by a combination of inequality conditions, $c(x,y) \geq 0$, and complementarity constraints, $0 \leq y \perp F(x, y) \geq 0$, where $0 \leq y \perp F(x, y) \geq 0$ if and only if $0 \leq y$, $F(x,y) \geq 0$, and $y^T F(x, y) = 0$. Equality constraints can be added without difficulty. The constraints correspond to, for example, budget and incentive constraints, and the complementarity constraints model the optimality conditions of the followers including any Lagrange multipliers. Then the leader's problem and the corresponding equilibrium are given by the solution to the MPEC

$$\begin{array}{ll} \max_{x,y} & f(x,y) \\ \text{s.t.} & c(x,y) \geq 0 \\ & 0 \leq y \perp F(x,y) \geq 0. \end{array}$$

MPECs present many mathematical challenges; the constraints are non-convex and reformulations as standard nonlinear optimization problems violate fundamental stability assumptions. Despite these facts, nonlinear optimization methods applied to such reformulations have been successful at solving some MPECs. For example, Chen et al. (2006) solve MPECs derived from some large-scale electricity market models. But they also show the limitations of the nonlinear optimization approach, and advocate the development of robust algorithms for solving MPECs that directly exploit the structure of the complementarity constraints. The development of such methods is under way. The ability to solve large and complicated MPECs will greatly enhance economic modelling in many areas and will likely make MPECs a key tool of computational economic analysis in the future.

See Also

- ▶ Computation of General Equilibria
- ▶ Computation of General Equilibria (New Developments)
- ▶ Computational Methods in Econometrics
- ▶ Dynamic Programming
- ▶ Linear Programming
- ▶ Non-linear Programming
- ▶ Operations Research
- ▶ Simplex Method for Solving Linear Programs

Acknowledgment *I am grateful for helpful discussions with Sven Leyffer and am indebted to Ken Judd, Annette Krauss, and in particular Che-Lin Su for detailed comments on earlier drafts. I also thank the editors Larry Blume and Steven Durlauf for a careful review of my initial submission.*

Bibliography

Allgower, E.L., and K. Georg. 1979. *Introduction to numerical continuation methods*. New York: John Wiley & Sons. Reprinted by SIAM Publications, 2003.

Bhatti, M.A. 2000. *Practical optimization methods: With mathematica applications*. New York: Springer-Verlag.

Brandimarte, P. 2006. *Numerical methods in finance and economics: A MATLAB-based introduction*. New York: John Wiley & Sons.

Cerny, V. 1985. A thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *Journal of Optimization Theory and Applications* 45: 41–51.

Chen, Y., B.F. Hobbs, S. Leyffer, and T.S. Munson. 2006. Leader–follower equilibria for electric power and NO_x allowances markets. *Computational Management Science* 4: 307–330.

Conn, A.R., N.I.M. Gould, and P.L. Toint. 2000. *Trust-region methods*. Philadelphia: SIAM.

Cox, D., J. Little, and D. O'shea. 1997. *Ideals, varieties, and algorithms: An introduction to computational algebraic geometry and commutative algebra*. New York: Springer-Verlag.

Czyzyk, J., M.P. Mesnier, and J. Morè. 1998. The NEOS server. *IEEE Journal on Computational Science and Engineering* 5: 68–75.

Dantzig, G.B. 1949. Programming of inter-dependent activities II, mathematical model. *Econometrica* 17, 200–211. Also in Koopmans, T.C., ed. *Activity analysis of production and allocation*. New York: John Wiley & Sons, 1951.

Dantzig, G.B. 1963. *Linear programming and extensions*. Princeton: Princeton University Press.

Ferris, M.C., M.P. Mesnier, and J. Moré. 2000. NEOS and Condor: Solving nonlinear optimization problems over the Internet. *ACM Transactions on Mathematical Software* 26: 1–18.

Fiacco, A.V., and G.P. McCormick. 1968. *Nonlinear programming: Sequential unconstrained minimization* *techniques*. New York: John Wiley & Sons, Inc. Reprinted by SIAM Publications, 1990.

Fletcher, R. 1987. *Practical methods of optimization*. Chichester: John Wiley & Sons.

Fourer, R., D.M. Gay, and B.W. Kernighan. 2003. *AMPL: A modeling language for mathematical programming*. Pacific Grove: Brooks/Cole–Thomson Learning.

Frisch, R.A.K. 1955. The logarithmic potential method of convex programming. Technical Report, University Institute of Economics, University of Oslo, Norway.

Gould, N.I.M., and S. Leyffer. 2002. An introduction to algorithms for nonlinear optimization. In *Frontiers in numerical analysis*, ed. J.F. Blowey, A.W. Craig, and T. Shardlow. Berlin/Heidelberg: Springer-Verlag.

Holland, J.H. 1975. *Adaptation in natural and artifical systems*. Ann Arbor: University of Michigan Press.

Judd, K.L. 1998. *Numerical methods in economics*. Cambridge, MA: MIT Press.

Karmarkar, N. 1984. A new polynomial-time algorithm for linear programming. *Combinatorics* 4: 373–395.

Kendrick, D.A., P.R. Mercado, and H.M. Amman. 2006. *Computational economics*. Princeton: Princeton University Press.

Khachiyan, L.G. 1979. A polynomial algorithm in linear programming. *Soviet Mathematics Doklady* 20: 191–194.

Kirkpatrick, S., C.D. Gelatt, and M.P. Vecchi. 1983. Optimization by simulated annealing. *Science* 220: 671–680.

Lasserre, J.B. 2001. Global optimization with polynomials and the problem of moments. *SIAM Journal of Optimization* 11: 796–817.

Markowitz, H.M. 1952. Portfolio selection. *Journal of Finance* 7: 77–91.

Miranda, M.J., and P. Fackler. 2002. *Applied computational economics and finance*. Cambridge, MA: MIT Press.

Nocedal, J., and S.J. Wright. 2006. *Numerical optimization*. New York: Springer.

Parrilo, P.A., and B. Sturmfels. 2003. Minimizing polynomial functions. *Algorithmic and Quantitative Real Algebraic Geometry, DIMACS Series in Discrete Mathematics and Theoretical Computer Science* 60: 83–99.

Rosenthal, R.E. 2006. *GAMS – A user's guide*. Washington, DC: GAMS Development Corporation. Online. Available at http://www.gams.com/docs/gams/GAMSUsersGuide.pdf. Accessed 7 Feb 2007.

Simon, C.P., and L. Blume. 1994. *Mathematics for economists*. New York: W. W. Norton.

Smale, S. 1976. A convergent process of price adjustment and global Newton methods. *Journal of Mathematical Economics* 3: 107–120.

Sturmfels, B. 2002. *Solving systems of polynomial equations,* vol. 97, CBMS Regional Conference Series in Mathematics, Providence: American Mathematical Society.

Nurkse, Ragnar (1907–1959)

Kaushik Basu

Keywords

Balanced growth; Development economics; Full employment; International policy coordination; International trade; Nurkse, R.; Poverty traps

JEL Classifications

B31

Nurkse was born on 5 October 1907 on an estate where his father was an overseer, near the village of Viru in Estonia. His father was Estonian and his mother of Swedish origin. Ragnar Nurkse was educated in Tallinn, Tartu, Edinburgh and Vienna. From 1934 to 1945 he worked as an economist with the League of Nations and from 1945 until his death he was a professor at Columbia University. He wrote on international currency questions, trade, vicious circles of poverty and on balanced growth. In 1959 he delivered the Wicksell Lectures in Stockholm. Exhausted by the lectures, he went to Geneva and while taking a stroll on Mont Pèlerin he collapsed and died of a heart attack or stroke on 6 May 1959. The Wicksell Lectures were published posthumously (Nurkse 1961).

One of Nurkse's two most important books was *International Currency Experience: Lessons of the Inter-War Period* (1944). It was published by the League of Nations, and though it did not carry the name of any author, this was (excepting chapter 6) the work of Nurkse. From this and several other of his writings, what comes out most clearly is Nurkse's pragmatism. Though he was one of the originators of the doctrine of balanced growth, he never minimized the role of international trade. However, he believed that the scope for trade-based expansion for Third World countries was much less in the 20th century than it was in the 19th century. Balanced growth could supplement this and even enlarge the scope for trade. Balanced growth and international trade, Nurkse argued, ‘are really friends, not enemies’ (Haberler and Stern 1961, p. 257).

Nurkse had a deep concern for full employment. He viewed exchange rate adjustments and trade restrictions as legitimate measures for preventing balance of payments difficulties from translating into unemployment and domestic instability. He stressed that trade restrictions ought to be used as temporary measures. With the emergence of Keynesian macroeconomics, Nurkse came to have faith in effective- demand management as a tool for maintaining employment in the face of trade adversities. This also led him to argue for some international coordination of domestic policies.

Nurkse’s other important (and, in my opinion, more important) book was *Problems of Capital Formation in Underdeveloped Countries* (1953). Here he developed the important idea that though the producer of each commodity may find an expansion unprofitable because of limitations of the market, a coordinated expansion of all productive activities could be profitable for all producers. Hence, atomistic behaviour on the part of producers could trap an economy *within* its production possibility frontier. This idea had been discussed earlier – most notably by Rosenstein-Rodan (1943) and more distantly by Young (1928) – but Nurkse took it further. While this work has been the basis of several debates in development economics (for critiques and formalizations, see Flemming 1955; Findlay 1959), it has the scope for further research, especially in the light of recent advances in non-Walrasian equilibrium analysis (see Basu 1984).

The lack of formalization in Nurkse’s work led to much misunderstanding – handsomely contributed to by Nurkse himself – about the policy implications of the poverty-trap doctrine. Nurkse tried to clarify these in his Ankara lectures in 1957 and his posthumously published note in *Oxford Economic Papers* (1959), both reprinted in Haberler and Stern’s (1961) collection. The potential of this branch of development economics remains large.

See Also

▶ Balanced Growth

Selected Works

1944. *International currency experience: Lessons of the interwar period.* Princeton: League of Nations.

1947. International monetary policy and the search for economic stability. *American Economic Review, Papers and Proceedings* 35: 569–580.

1953. *Problems of capital formation in underdeveloped countries.* Oxford: Basil Blackwell.

1954. International investment today in the light of nineteenth-century experience. *Economic Journal* 64: 744–758.

1959. Notes on ‘unbalanced growth’. *Oxford Economic Papers* 11(3): 295–297.

1961. *Patterns of trade and development. The wicksell lectures*. Oxford: Basil Blackwell.

Bibliography

Basu, K. 1984. *The less developed economy: A critique of contemporary theory.* Oxford: Basil Blackwell.

Findlay, R.V. 1959. International specialisation and the concept of balanced growth: Comment. *Quarterly Journal of Economics* 73: 339–346.

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

Haberler, G., and R.M. Stern, eds. 1961. *Equilibrium and growth in the world economy: Economy essays by Ragnar Nurkse*. Cambridge, MA: Harvard University Press.

Rosenstein-Rodan, P.N. 1943. Problems of industrialization of Eastern and South-Eastern Europe. *Economic Journal* 53: 202–211.

Young, A. 1928. Increasing returns and economic progress. *Economic Journal* 38: 527–542.

Nursing Homes

Edward C. Norton

Abstract

Nursing homes are healthcare providers for persons, often elderly, who need assistance living with chronic illness. This article

describes the main economic issues of supply and demand for nursing home care, including quality of care and long-term care insurance.

Nursing home care is an important area for health economics because it represents the largest share of long-term care expenditure. The potential for needing nursing home care affects economic decisions for individuals over a lifetime and across generations (Norton 2000). For example, an elderly widow anticipating a need for long-term care may decrease her savings or increase her bequests to qualify for means-tested public insurance, or may demand informal care from a working daughter, even though she ultimately never enters a nursing home.

Long-term care differs from acute medical care in four fundamental ways (Norton 2000). First, long-term care is care for chronic illness or disability instead of treatment of an acute illness. Medical expenses accumulate unrelentingly. Second, the nursing home industry is dominated by for-profit facilities sometimes facing excess demand, in contrast to the hospital industry which is dominated by non-profit facilities with an excess supply of beds. Third, nursing homes have many close substitutes, including informal care. Informal care may affect the caregiver's labour supply or may influence bequests, if such bequests are used to elicit attention and informal caregiving by children. Fourth, in contrast to relatively comprehensive acute care insurance for elderly, few people purchase private long-term care insurance and most public insurance is means-tested, with high co-payments. Thus, long-term care is usually the greatest out-of-pocket expenditure risk faced by the elderly.

This article summarises the theoretical and empirical economic research on nursing homes and long-term care. In addition to discussing supply and demand, particular attention is focused on quality of care and the market for long-term care insurance.

Keywords

Aging; Assisted living; Chronic care; Elderly; Informal care; Insurance; Long-term care; Nursing homes

JEL Classifications

I110; I130

Taxonomy

Long-term care covers a continuous spectrum, from infrequent informal care provided by a neighbour to institutional care with around-the-clock nursing. The nursing home industry is an appropriate starting point for a review of long-term care because of its size and cost. Many elderly, and a few disabled nonelderly, people enter a nursing home when they are no longer able to live independently. In the USA, on any given day 5% of persons aged 65 and older are nursing home residents. Lengths of stay in nursing homes vary widely, from short stays of a day or two, to lengthy stays of several decades. Nursing home care is expensive, and insurance is far from complete.

There are many imperfect substitutes for nursing homes for long-term care. The choice depends on the individual's physical and mental health, finances, and family situation. Despite the visibility of nursing homes, most care for the elderly is provided informally. Informal care is most often provided by spouses and children. Informal care can reduce nursing home use and expenditures because it is a substitute (Van Houtven and Norton 2004, 2008). Other forms of long-term care that are partial substitutes for nursing home care include home healthcare, board and care homes, adult foster care, adult day care, hospice care (Hamilton 1993) and continuing care retirement communities (CCRC). In summary, the market has developed a variety of solutions to the problem of giving care to chronically ill persons with widely varying physical and mental health status, finances and family situations.

For research on long-term care outside the USA, see Carmichael and Charles (2003), Forder and Netten (2000), Laine et al. (2005), Lindeboom et al. (2002), Noguchi and Shimizutani (2006), O'Neill et al. (2000), and Portrait et al. (2000).

Supply

The nursing home market has many properties of a competitive market (Bishop 1988). Barriers to entry are low, capital costs per bed are much lower for a nursing home than for a hospital, and new nursing homes can enter with little owner equity. Nursing homes hire relatively unskilled labour and do not need highly specialised equipment. Administrative and licensing costs are also low. Furthermore, there are few, if any, economies of scale. Nursing homes can enter with few beds. Therefore, barring regulation of entry, nursing homes of all sizes should be able to enter the market easily, based on entry costs.

Despite these attributes of a competitive market, the nursing home market is not competitive in many ways. Many nursing homes have waiting lists and operate at, or near, full capacity. The waiting lists may imply that demand exceeds supply, which would not happen in a freely competitive market in equilibrium. Part of the constraint on the market in the USA is that a majority of residents are covered by Medicaid or Medicare and pay regulated rates. Medicare covers short-term stays for persons expected to recover. Another reason may be due to direct constraints on supply due to Certificate-of-Need (CON) regulations. Although data from the late 1960s through the early 1980s argued that CON was a binding constraint for Medicaid beds (Scanlon 1980), recent research has shown different results. Grabowski et al. (2003) found that states that repealed CON and moratoria laws did not experience an increase in Medicaid expenditures relative to states that did not repeal these laws. Similarly, Gulley and Santerre (2003) found that the CON laws did not affect access to nursing home care. The national trend towards lower occupancy rates is consistent with the idea that CON and moratoria are no longer binding constraints in most nursing home markets.

A competitive market also requires informed consumers. Unlike acute medical care, the demand for nursing home care is often not time-sensitive. Potential residents may have days or weeks in which to search. Potential residents can obtain help from hospital discharge planners, relatives and social workers. Nursing home services are not technical and can be evaluated more easily by consumers than, say, surgical skill. There is a wide range of close substitutes, creating competition. However, in practice most consumers are not well informed. Elderly people who need nursing home care are disproportionately the ones with no close family to help them search, and end up in a nursing home because they have fewer options than other elderly. Those with close family often postpone searching for a nursing home because the thought of institutionalisation is unpleasant. Then, when a decision becomes necessary, location is often the overriding criterion. Elderly persons may have no choice if there are waiting lists and they are covered by Medicaid.

Quality of Care

The early literature on nursing home quality of care was largely based on Scanlon's model (1980), in which nursing homes face two markets. One market is for private residents with downward sloping demand, and the other is for Medicaid residents who are insensitive to price. Scanlon presented evidence that the Medicaid residents faced excess demand nationally. Certificate-of-Need regulations and construction moratoria policies had constrained growth in the supply of nursing home beds, and nursing homes preferred to admit higher-paying private patients. As a result, when a bed shortage existed, it was the Medicaid patients who would be excluded.

Many policymakers argued that nursing home quality could be improved by raising Medicaid reimbursement rates. By incorporating a quality variable into Scanlon's model, Nyman (1985) showed that raising Medicaid rates in a market with excess demand would result in nursing homes facing a reduced incentive to use quality of care to compete for the private patients. Several papers confirmed this inverse relationship between Medicaid reimbursement level and quality of care (Nyman 1985; Gertler 1989; Dusansky 1989; Gertler 1992). Nyman (1988, 1989) proposed that this outcome would be spurious if tight markets eliminated an observable measure of quality – the occupancy rate – to inform consumers. Norton

(1992) showed that cost-mix adjusted reimbursement with incentives for quality improvement lead to improved health outcomes.

More recently the market has changed due to the decline in nursing home occupancy rates and the repeal of CON laws in some states. Recent studies have generally found a modest positive relationship between state Medicaid payment rates and nursing home quality, unlike the earlier research. Higher payment rates have been found to be associated with fewer pressure ulcers (Grabowski and Angelelli 2004), more staffing (Grabowski 2001b), fewer hospitalisations (Intrator and Mor 2004), fewer physical restraints (Grabowski et al. 2004), less feeding tube use and fewer government-cited deficiencies (Grabowski 2004). In terms of the size of the effect, these studies indicate a payment–quality elasticity in the range 0.1–0.7, depending on the quality measure. Importantly, the most recent studies provide little support for a negative relationship between the Medicaid payment level and quality.

In an attempt to bridge the two generations of this literature, Grabowski (2001a) replicated the data, methods and quality measures from Gertler (1989) to identify the underlying source of the different findings. When the methods and quality measures from the earlier study were applied to more recent data, Medicaid payment was found to be positively associated with quality. Changes in the marketplace – not alternative data or methods – explain the different findings across the two generations of studies. However, using national data from the earlier time period, Grabowski also found that Gertler's New York results did not generalise to the entire USA. Thus, the earlier result may have been only been relevant for a minority of states or markets where CON laws were particularly binding.

Public Quality Information

Economists have long studied the problem of asymmetric information in the healthcare market (Arrow 1963). Without accurate information on nursing home quality, the market matching patients to providers will result in poor matches. Healthcare is partly an experience good. In principle, a patient could eventually discern a nursing home's quality, but most patients only seek care once or a few times.

There are now published report cards and performance measures in the USA for nursing homes (and also for physicians, hospitals and home healthcare providers). The idea of Nursing Home Compare is to pool information on the experience of recent patients and make that information available to all potential patients. By pooling collective experience, healthcare can be an experience good.

Clearly, accurate timely information could help consumers choose higher-quality providers and induce providers to compete on quality (Werner and Asch. 2005). Even with good information, there are many potential problems and unintended effects (Werner and Asch 2005). These problems may be exacerbated with elderly patients, who are usually less able to handle complex comparative information.

Empirical results are quite mixed on the effect of Nursing Home Compare on quality of care. Werner et al. (2009) found that two of three published quality measures improved, while a third, no delirium, did not improve significantly but was already at high levels. The unreported measure of hospitalisation, however, worsened. Hospitalisations are not merely an indication of a poor health outcome. They can also be used strategically to improve a nursing home's score (Konetzka et al. 2013). In contrast to CABG patients, where all patients are included in quality outcomes, for nursing homes only patients who stay at least 14 days are included. Konetzka et al. (2013) explain that this gives nursing homes a different kind of selection mechanism. They can discharge sicker patients back to the hospital just prior to the 14-day limit to keep poor-prognosis patients from adversely affecting their scores. Konetzka and colleagues find evidence of the hypothesised behaviour. This indicates that the concern about selection in performance measures is complicated in nursing homes.

A key assumption for advocates of report cards is that consumers will respond to quality information. If consumers are not responsive, then the case for publicly provided information falls.

Therefore it is important to show that consumers respond to quality information (Werner et al. 2012). Werner and colleagues lay out the argument on both sides of the debate for how response to web-based information may differ by education. On the one hand, those with more education may be better able to process the complex information and use it to make decisions. On the other hand, people with more resources may always have been able to find out about quality of care, so providing it publicly may actually level the playing field, especially with social workers and discharge planners offering advice. Werner et al. (2012) found that nursing homes with higher reported quality of care for pain control increased their market share for post-acute care, indicating that consumers are responsive to information about certain kinds of quality, although the magnitude of the effect was small. For education, they found that those with higher education had a slightly higher response, and the difference was statistically significant.

Quality of care also depends on the market conditions. Building moratoria and Certificate-of-Need restrictions reduce supply from free market levels, leading to excess demand in the nursing home market. In these cases, nursing homes may not compete as well on quality of care. Not surprisingly, nursing homes in competitive markets responded more to quality incentives by improving quality after Nursing Home Compare than nursing homes with greater market power (Grabowski and Town 2011). Previous work by Grabowski (2002) showed that in excess demand markets more dependent residents had access problems, but that quality of care remained unchanged with the introduction of case-mix reimbursement.

Ownership Type

In contrast to the hospital industry, two-thirds of all nursing homes are for-profit. In both industries, the mix of for-profit and non-profit firms has led to studies of how ownership affects costs, quality and access to care. In nursing homes, the primary concern is the existence of asymmetric information about quality. Arrow (1963) hypothesised that non-profit providers are common in markets for complex personal services because they have less incentive than for-profit providers to under-provide quality to poorly informed consumers (see also Hirth 1999). Consumers, especially frail elderly people with no close family support, may have trouble discerning quality within a nursing home, and may not have the ability to shop among nursing homes (Spector et al. 1998).

Several papers promote the idea that non-profit status is a signal of quality. Chou (2002) looked at differences in quality of care, measured by death and adverse health outcomes, between for-profit and non-profit nursing homes and between residents who had close family. She found that the differences between ownership types were greater when there was asymmetric information, meaning that no spouse or child visited within 1 month of admission. Grabowski and Hirth (2003) looked at the related issue of how the share of non-profit nursing homes in the market affected quality of care. They argue that a greater percentage of non-profit nursing homes would have competitive spillover effects, which is what they found after controlling for the endogeneity of non-profit market share.

Demand

Demand for nursing home care depends primarily on health status and the out-of-pocket price relative to the price of close substitutes. Those in worse health demand more long-term care. Those with fewer substitutes, or whose substitutes are higher-priced, demand more long-term care. Demand curves slope downward, and health shocks shift the demand curve outward.

The primary determinant of demand for nursing home care is health status – both physical and mental health. Persons in worse health status are more likely to go to a nursing home. As physical or mental health deteriorates, a person is less able to live independently and less able to perform the basic activities that most persons take for granted. Demand for long-term care is also related to other demographic characteristics, such as age, gender and race, because these variables are proxies for health status. Health status generally declines with age. Gender is related to nursing home use, but much of the effect of gender is due to health status and marital status. Married persons are more likely to receive informal care from their spouse. Married persons are also more likely to have children, another important source of informal care. Because women tend to outlive their husbands, women near the end of their life are less likely to be married and therefore are more likely to demand nursing home care. Another consequence is that men have worse health status at admission than women because they are more likely to have been able to stay at home with a spouse.

Race is significant in nearly every empirical study of nursing home use. Whites are more likely to use nursing home care than black, Hispanic or Asian people. Black people are more likely than white people to be on Medicaid, have severe illness and not have long-term care insurance coverage – all factors that hinder admission to a for-profit nursing home (White-Means 1997). Differences persist in empirical work, even after controlling for observable differences in insurance and health status. The difference in nursing home use may be related to cultural differences in preference for location of care, differences in health status or differences in access due to racial discrimination (Headen 1992). Race encompasses social, psychological, biological and genetic influences (White-Means 1995). Race therefore pervades socioeconomic status, attitudes and family culture, implying that empirical work should include not merely a dummy variable for race but a fully interacted model. The effect of race may also be related to the opportunity cost of informal care and nursing home care (Headen 1992). For example, if the wage rates of black people are lower than for white people, and the nursing home price is the same, then the opportunity cost of informal care is lower for black people. Headen (1992) found evidence that the opportunity cost of time – measured by labour force participation, education, age and social support – is lower for black informal caregivers than white informal caregivers.

The financial determinants of nursing home demand are the price, the relative price of close substitutes, and the person's income and assets. Nursing home demand will increase when the price falls or when the price of close substitutes rises. Private insurance lowers the out-of-pocket cost of nursing home care, but few elderly people have private insurance, and those who do may still face substantial co-payments and deductibles. Income and assets do not affect nursing home demand in a straightforward way.

The expected rapid rise in the number of elderly persons over the next few decades will greatly increase demand for all types of long-term care. However, two demographic trends may mitigate the problem. The mortality rate has fallen by about 1% per year since 1950, so elderly people are living longer (Cutler 2001). The longevity gender gap has narrowed because the mortality rate for elderly men is falling even faster than for women. In addition, disability rates among the elderly are declining. Therefore, people are living longer and living healthier (Manton and Gu 2001). Lakdawalla and Philipson (2002) argue that these trends help explain much of the decline in the relative growth in nursing home use seen since 1970. Still, overall demand is expected to increase as Baby Boomers enter the prime age for nursing home care.

Insurance

A risk-averse person facing an uncertain and expensive risk of needing long-term care should demand insurance. Indeed, the greatest financial uncertainty for elderly is long-term care expenditures (Norton et al. 2006). It is not food, pharmaceuticals or even inpatient care. In the USA, Medicare insurance is quite complete for inpatient care, outpatient care and pharmaceuticals, especially when considering Medigap and Medicaid policies that help pay co-payments and deductibles. But Medicare coverage of long-term care is quite limited. Medicare not only requires a prior inpatient stay, but requires substantial cost sharing after 20 days, and pays nothing after 100 days. Medicaid coverage of long-term care also requires substantial cost sharing. Roughly speaking, the deductible is most of a person's non-housing wealth and the co-pay is most of her income (Norton 1995). This leaves long-term care as the greatest expenditure risk. In addition to reducing financial risk, the desire to leave a bequest to spouse and children may be a major motive for purchasing long-term care insurance (Bernheim et al. 1985; Hurd 1987, 1989; Bernheim 1991).

Despite the apparent demand for long-term care insurance for the elderly there are many reasons why there is little private long-term care insurance sold. This has been discussed extensively in the literature (for reviews, see Norton (2000) and Brown and Finkelstein (2007, 2011)). Here is a summary of the most important reasons why the private insurance market is small. Adverse selection means that those who are most likely to need long-term care are most likely to want to buy it; insurance companies may target individuals who statistically are least likely to need it. Moral hazard is often a problem in insurance markets. For long-term care there is both standard moral hazard and a version proposed by Pauly (1990) in which elderly people do not buy insurance so that their children, the presumed future decision makers, will not put them in a nursing home. Loading (administrative) costs are high because most sales are made to individuals and because adverse selection requires background and health checks. The load for private long-term care insurance has been estimated to be about 32% (Brown and Finkelstein 2011). It is high because it is mostly sold to individuals and because of the high commission fees paid to the brokers. The high overhead raises the premium and lowers demand.

Medicaid is a close substitute for part of the population who would qualify for Medicaid quickly (Hubbard et al. 1995). But a major reason why there is low demand for private insurance is that the benefit is low. Insurance companies now offer capped daily benefits, instead of paying a fraction of the cost (like most other health insurance), because of the difficulty in predicting future nursing home costs (Cutler 1996). Some policies are limited in the number of days of coverage. People who lapse in their insurance payments forfeit their coverage. These policies all reduce the insurance value of the product and lower its desirability. Some elderly people greatly underestimate their own risk of needing long-term care, again lowering demand. Given all these reasons combined, it is perhaps a wonder anyone buys long-term care insurance.

Conclusion

Nursing homes are an important part of the spectrum of long-term care providers. They are the most expensive form of long-term care and are used extensively by persons unable to live independently. The market is not as competitive as one would expect from an industry with low barriers to entry. Regulated prices and poorly informed consumers make the market less competitive and contribute to the poor overall level of quality of care. Attempts to improve quality of care have recently focused on publicly provided information on quality. Demand for nursing homes is predominantly driven by poor physical and mental health, but also depends on the relative price of close substitutes. The market for private long-term care insurance is hampered not only by the usual problems of adverse selection and moral hazard, but also high loading and poor benefits. The economic issues surrounding nursing homes will continue to be important as the population ages over the next several decades.

See Also

- ▶ Health Economics
- ▶ Health Insurance, Economics of
- ▶ Population Ageing

Bibliography

Arrow, K.J. 1963. Uncertainty and the welfare economics of medical care. *American Economic Review* 53(5): 941–973.

Bernheim, B.D. 1991. How strong are bequest motives? Evidence based on estimates of the demand for life insurance and annuities. *Journal of Political Economy* 99(5): 899–927.

Bernheim, B.D., A. Shleifer, and L.H. Summers. 1985. The strategic bequest motive. *Journal of Political Economy* 93(6): 1045–1076.

Bishop, C.E. 1988. Competition in the market for nursing home care. *Journal of Health Politics, Policy and Law* 13(2): 341–360.

Brown, J.R., and A. Finkelstein. 2007. Why is the market for long-term care insurance so small? *Journal of Public Economics* 91(10): 1967–1991.

Brown, J.R., and A. Finkelstein. 2011. Insuring long-term care in the United States. *Journal of Economic Perspectives* 25(4): 119–141.

Carmichael, F., and S. Charles. 2003. The opportunity costs of informal care: Does gender matter? *Journal of Health Economics* 22(5): 781–803.

Chou, S.Y. 2002. Asymmetric information, ownership and quality of care: An empirical analysis of nursing homes. *Journal of Health Economics* 21(2): 293–311.

Cutler, D. M. 1996. Why Don't Markets Insure Long-term Risk? Harvard mimeo.

Cutler, D.M. 2001. Declining disability among the elderly. *Health Affairs* 20(6): 11–27.

Dusansky, R. 1989. On the economics of institutional care of the elderly in the U.S.: The effects of change in government payment. *Review of Economics* 56: 141–150.

Forder, J., and A. Netten. 2000. The price of placements in residential and nursing home care: The effects of contracts and competition. *Health Economics* 9(7): 643–657.

Gertler, P.J. 1989. Subsidies, quality, and the regulation of nursing homes. *Journal of Public Economics* 38: 33–52.

Gertler, P.J. 1992. Medicaid and the cost of improving access to nursing home care. *Review of Economics and Statistics* 74(2): 338–345.

Grabowski, D.C. 2001a. Medicaid payment and the quality of nursing home care. *Journal of Health Economics* 20(4): 549–570.

Grabowski, D.C. 2001b. Does an increase in the Medicaid payment rate improve nursing home quality? *Journal of Gerontology: Social Sciences* 56B(2): S84–S93.

Grabowski, D.C. 2002. The economic implications of case-mix Medicaid reimbursement for nursing home care. *Inquiry* 39(3): 258–278.

Grabowski, D.C. 2004. A longitudinal study of Medicaid payment, private-pay price and nursing home quality. *International Journal of Health Care Finance and Economics* 4(1): 5–26.

Grabowski, D.C., and J.J. Angelelli. 2004. The relationship of Medicaid payment rates, bed constraint policies, and risk-adjusted pressure ulcers. *Health Services Research* 39(4): 793–812.

Grabowski, D.C., and R.A. Hirth. 2003. Competitive spillovers across nonprofit and for-profit nursing homes. *Journal of Health Economics* 22(1): 1–22.

Grabowski, D.C., and R.J. Town. 2011. Does information matter? Competition, quality, and the impact of nursing home report cards. *Health Services Research* 46(6): 1698–1719.

Grabowski, D.C., R.L. Ohsfeldt, and M.A. Morrissey. 2003. The effects of CON repeal on Medicaid nursing home and long term care expenditures. *Inquiry* 40(2): 146–157.

Grabowski, D.C., J.J. Angelelli, and V. Mor. 2004. Medicaid payment and risk adjusted nursing home quality measures. *Health Affairs* 23(5): 243–252.

Gulley, O.D., and R.E. Santerre. 2003. The effect of public policies on nursing home care in the United States. *Eastern Economic Journal* 29(1): 93–104.

Hamilton, V.H. 1993. The Medicare hospice benefit: the effectiveness of price incentives in health care policy. *RAND Journal of Economics* 24(4): 605–624.

Headen, A.E. Jr. 1992. Time costs and informal social support as determinants of differences between black and white families in the provision of long-term care. *Inquiry* 29: 440–450.

Hirth, R.A. 1999. Consumer information and competition between nonprofit and for-profit nursing homes. *Journal of Health Economics* 18(2): 219–240.

Hubbard, R.G., J. Skinner, and S.P. Zeldes. 1995. Precautionary savings and social insurance. *Journal of Political Economy* 103: 360–399.

Hurd, M.D. 1987. Savings of the elderly and desired bequests. *American Economic Review* 77(3): 298–212.

Hurd, M.D. 1989. Mortality risk and bequests. *Econometrica* 57(4): 779–813.

Intrator, O., and V. Mor. 2004. Effect of state Medicaid payment rates on hospitalizations from nursing homes. *Journal of the American Geriatrics Society* 52(3): 393–398.

Konetzka, R.T., D. Polsky, and R.M. Werner. 2013. Shipping out instead of shaping up: rehospitalization from nursing homes as an unintended effect of public reporting. *Journal of Health Economics* 32(2): 341–352.

Laine, J., M. Linna, U. Hakkinen, and A. Noro. 2005. Measuring the productive efficiency and clinical quality of institutional long-term care for the elderly. *Health Economics* 14(3): 245–256.

Lakdawalla, D., and T. Philipson. 2002. The rise in old-age longevity and the market for long-term care. *American Economic Review* 92(1): 295–306.

Lindeboom, M., F. Portrait, and G.J. van den Berg. 2002. An econometric analysis of the mental-health effects of major events in the life of older individuals. *Health Economics* 11(6): 505–520.

Manton, K.G., and X.L. Gu. 2001. Changes in the prevalence of chronic disability in the United States black and nonblack population above age 65 from 1982 to 1999. *Proceedings of the National Academy of Sciences of the United States of America* 98(11): 6354–6359.

Noguchi, H., and S. Shimizutani. 2006. Do non-profit operators provide higher quality of care? Evidence from micro-level data for Japan's long-term care industry. *Hitotsubashi Journal of Economics* 47(1): 125–135.

Norton, E.C. 1992. Incentive regulation of nursing homes. *Journal of Health Economics* 11(2): 105–128.

Norton, E.C. 1995. Elderly assets, Medicaid policy, and spend-down in nursing homes. *Review of Income and Wealth* 41(3): 309–329.

Norton, E. C. 2000. Long-term care. In: Handbook of health economics, vol. IB (eds. A. J. Culyer and J. P. Newhouse), pp. 956–994. Elsevier Science, New York.

Norton, E.C., H. Wang, and S.C. Stearns. 2006. Behavioral implications of out-of-pocket health care expenditures. *Swiss Journal of Economics and Statistics* 142(Special Issue): 3–11.
Nyman, J.A. 1985. Prospective and cost-plus Medicaid payment, excess Medicaid demand, and the quality of nursing home care. *Journal of Health Economics* 4(3): 237–259.
Nyman, J.A. 1988. Excess-demand, the percentage of Medicaid patients, and the quality of nursing home care. *Journal of Human Resources* 23(1): 76–92.
Nyman, J.A. 1989. Excess demand, consumer rationality, and the quality of care in regulated nursing homes. *Health Services Research* 24(1): 105–127.
O'Neill, C., L. Groom, A.J. Avery, D. Boot, and K. Thornhill. 2000. Age and proximity to death as predictors of GP care costs: Results from a study of nursing home patients. *Health Economics* 9(8): 733–738.
Pauly, M.V. 1990. The rational nonpurchase of long-term-care insurance. *Journal of Political Economy* 98(1): 153–168.
Portrait, F., M. Lindeboom, and D. Deeg. 2000. The use of long-term care services by the Dutch elderly. *Health Economics* 9(6): 513–531.
Scanlon, W.J. 1980. A theory of the nursing home market. *Inquiry* 17: 25–41.
Spector, W.D., T.M. Selden, and J.W. Cohen. 1998. The impact of ownership type on nursing home outcomes. *Health Economics* 7(7): 639–653.
Van Houtven, C.H., and E.C. Norton. 2004. Informal care and health care use of older adults. *Journal of Health Economics* 23(6): 1159–1180.
Van Houtven, C.H., and E.C. Norton. 2008. Informal care and Medicare expenditures: Testing for heterogeneous treatment effects. *Journal of Health Economics* 27(1): 134–156.
Werner, R.M., and D.A. Asch. 2005. The unintended consequences of publicly reporting quality information. *Journal of the American Medical Association* 293(10): 1239–1244.
Werner, R.M., R.T. Konetzka, E.A. Stuart, E.C. Norton, D. Polsky, and J. Park. 2009. Impact of public reporting on quality of postacute care. *Health Services Research* 44(4): 1169–1187.
Werner, R.M., E.C. Norton, R.T. Konetzka, and D. Polsky. 2012. Do consumers respond to publicly reported quality information? Evidence from nursing homes. *Journal of Health Economics* 31(1): 50–61.
White-Means, S.I. 1995. Conceptualizing race in economic models of medical utilization: A case study of community-based elders and the emergency room. *Health Services Research* 30(1): 207–223.
White-Means, S.I. 1997. The continuing significance of race in meeting health care needs of black elderly. In *Race, markets, and social outcomes*, ed. P. Mason and R. Williams. Norwell: Kluwer.

Nutrition

C. Peter Timmer

The economics of nutrition has both demand and supply aspects. Because nutrients for human growth, development and physical activity come almost entirely from food consumed, the demand side of nutrition economics is closely related to food consumption analysis. Because these nutrients interact with the body's health status as well as demands imposed by physical and social activities to produce 'work output', nutrition economics also relates to the burgeoning literature on the formation and productivity of human capital. And because the process of buying foods and transforming them into a family's daily diet involves primarily women's time in the household, nutrition economics also relates to analysis of the productivity of women's activities and to the 'new household economics' paradigm. In addition, biological and economic links have been established between nutrition and fertility. In combination with the influence of maternal and infant nutritional status on mortality rates, these links establish an important connection between nutrition economics and population studies and provide a vehicle for economists to contribute to that field.

At one level, the economics of nutrition touches on nearly all aspects of economic activity through its pervasive influence on demand for commodities, allocation of household time, and resulting productivity and size of a nation's workforce. At another level, nutrition is primarily a non-market issue, and many of the important analytical topics involve unobservable relationships within the household or even within the human body itself. From this viewpoint, it is not surprising that the field of nutrition economics does not contain a coherent set of empirical regularities based on common methodological frameworks and accepted data bases. The field exists as a series of niches in the broader areas of inquiry just noted – in analysis of food demand, in formation of human capital, and in household economics. The purpose of this short article is to draw together, according to an economic perspective, the aspects that specifically relate to nutrition from those diverse fields. The economic perspective is not the only one possible, of course, because nutrition has traditionally been considered primarily a topic in applied biochemistry, where health and medical professionals identify the important problems from the field that need to be solved through bench research in the laboratory.

Only in the past three decades has the bio-medical approach to nutrition – the identification, synthesis and evaluation of physiological significance of nutrients essential for human health and well-being – been broadened to include public health professionals and discipline-based social scientists, including economists. The stimulus came from two major directions.

In the first instance, the documentation of significant hunger in the United States by the Field Foundation in the early 1960s led to the rapid expansion of the Food Stamp Program, followed by the Women, Infants, and Children (WIC) Supplemental Feeding Program. As budget expenditures for the elimination of hunger rose, so too did the concern for understanding basic causes of hunger and its impact on the hungry. Of particular importance in early stages of the programmes was the identification of minimum-cost diets that met nutritional standards so that the value of Food Stamps distributed could be determined. The so-called 'Budget Plan' of diets to be followed by Food Stamp recipients became a hot political issue because its value established the financial cost of the government's most widespread welfare programme. Efforts to expand the constraints needing to be satisfied at minimum cost – eventually to include not only the palatability of the diet but also its social acceptability – led to the inclusion of a broad range of social scientists in programme design and evaluation.

Secondly, a similar but broader set of concerns arose in the 1960s in the economic development profession. Analysis of Food Stamp recipients in the United States showed that poverty was the primary reason for hunger (lack of energy) and malnutrition (imbalance of nutrients, including protein relative to energy). Consequently, nutritional problems were likely to be orders of magnitude worse in poor countries than in rich ones. At the same time, the budgetary and administrative resources available to intervene in the problem were substantially smaller. Two important lines of analysis received attention in this development context: attempts to measure the economic benefits of nutrition interventions, such as supplemental feeding programmes, in order to establish a cost–benefit basis for their expansion (Selowsky and Taylor 1973); and attempts to understand at a highly disaggregated level the demand parameters of those individuals and families likely to be suffering from hunger or malnutrition, or both. Out of this work evolved new directions for programmes and policies for dealing more effectively with these problems (Austin and Zeitlin 1981).

The impact of these studies on policy has often been quite significant, especially in preventing budget cuts to the WIC Program when other welfare programmes were being cut back. Important analytical approaches have been developed to address such policy issues in developed countries, but the most extensive interest in nutrition economics has come from developing countries. The reasons are easy to understand: if societies must wait until they are rich to solve their nutrition problems, then widespread hunger is likely to persist for centuries. As one component of the Basic Needs approach to development, the question was asked whether shortcuts to improved nutritional status were available, at what costs and with what benefits (Streeten et al. 1982).

The starting point for raising and analysing these issues was Alan Berg's volume for Brookings, *The Nutrition Factor: Its Role in National Development* (1973). Growing out of Berg's experiences in managing food aid shipments to India during the 1966–7 food crisis there, the book provided a holistic approach to the role of nutrition in the development process and the potential scope for government interventions. This broad vision of nutrition as a central theme linking agriculture, population, food technology, education and the income dimensions of economic growth evolved into the concept of nutrition planning, which attempted to coordinate all government activities, from macroeconomic policy to agricultural research, with the objective of improving nutritional status (Anderson and Grewal 1976). The field spawned its own journal, *Nutritional Planning,* and several interdisciplinary doctoral programmes.

Attempts to implement nutrition plans, however, ran into serious problems. Even when political commitment to such plans was high, as in Mexico, Sri Lanka and Colombia, and ambitious policy changes were contemplated on behalf of nutritional objectives, economists were unable to specify with any precision what the nutritional outcome of changes in policy (or even programmes and projects) would be. Reutlinger and Selowsky (1976) estimated the number of undernourished people in developing countries based on average calorie requirements, semi-log income elasticity functions for calorie intake and rough income distribution data by region. They concluded that income growth *per se* would not lead to rapid reductions in hunger, but the analysis came under fire from both economists and nutritionists for its aggregative view of the problem. Many factors other than incomes are influential in affecting nutrient intake and health outcomes, especially the prices of important foodstuffs, and the search began for the behavioural parameters that would link variables subject to governmental intervention, such as the distribution of income growth or the prices of basic grains, to decisions at the household level that had an impact on nutritional status.

The first sophisticated attempt to measure these disaggregated parameters was made by Per Pinstrup-Anderson and his colleagues (1976) at CIAT, the International Center for Tropical Agriculture in Cali, Colombia. Their goal was to determine priorities for crop research in terms of its ultimate contribution to improved diets in Colombia. Picking the simplest case to start with, they asked what would happen to nutrient intake by income class in urban areas if it were assumed that technological change would lower market prices for individual commodities. By using the Frisch methodology to estimate a full system of demand parameters and using data from two cross-section surveys to determine money flexibility by income class, Pinstrup-Anderson and his colleagues were able to trace the effects of changes in prices on nutrient intake.

When policy is used to change prices for producers and consumers and when rural as well as urban dietary patterns are investigated, the analysis becomes much more complicated. The price changes have direct and indirect effects on rural incomes. If labour markets are connected, some of the rural dynamics are likely to be transmitted to the urban economy. Even when the question addressed is limited to the impact of food price changes on nutrition, the answer typically requires a general equilibrium approach. This broader concern for effects on production, intersectoral linkages and macroeconomic consequences of food policy, in addition to nutritional outcomes, called for an integrated analytical perspective, such as in Timmer et al. (1983).

Whether the focus is projects, programmes or policies, understanding nutritional impact requires knowledge of matrices of income and price elasticities for specific foodstuffs by income class. The empirical search for these parameters has pushed consumption economics into new areas both methodologically and with respect to data sources (see Waterfield 1985, for a review of this literature and Timmer 1981, for the implications for consumer theory).

The project-oriented and policy-oriented literatures have evolved in somewhat different directions. The former has focused on problems of design and implementation in targeting delivery of services, especially in rural development programmes. Johnston and Clark (1982), for example, focus on the management of integrated delivery systems for nutrition, health, and family planning services in rural areas. At the policy level, attention has generally been focused on prices because they are relatively easy for government trade and exchange rate policies to influence (see Solimano and Taylor 1980; Timmer 1986).

The 'supply-side' dimensions of nutrition economics have been much more difficult to specify and quantify. The long-standing 'nutritional wage' issue has been carefully treated theoretically by Bliss and Stern (1978), but the review by Binswanger and Rosenzweig (1981) found little evidence for a nutritional floor to wages in the South Asian context, which is where the hypothesis arose. The survey of 'health and nutrition' by Behrman and Deolalikar for the *Handbook of Development Economics* concludes rather pessimistically with respect to current knowledge. It has been impossible to specify even basic energy requirements of individuals relative to measurable outcomes of interest to society, much less a direct link between nutrient intake and health status. An influential group of nutritionists and economists has emphasized the very substantial capacity of the human body to cope with low-energy intake through both metabolic and physical adaptations. The result is the hypothesis of Sukhatme (1982), Srinivasan (1981), Seckler (1982) and Payne and Cutler (1984) that people may be 'small but healthy'. Many other nutritionists and economists find this view highly controversial and wish to impose a more normative standard of achieving long-term potential rather than short-term health as the criterion for nutrient intake (Beaton 1983; Reutlinger and Alderman 1980). The result of this intellectual standoff is that nutrition is no longer seen as a 'tangible' marker of development where progress could be stressed (and measured) relatively independently of the broader and slower overall economic development process. Considering the significant connections between factors affecting nutrient intake and the policy environment conditioning the development process, the demise of such a shortcut mentality is perhaps healthy.

See Also

▶ Development Economics
▶ Fecundity
▶ Infant Mortality

Bibliography

Anderson, M.A., and T. Grewal (eds.). 1976. *Nutrition planning in the developing world*. New York: CARE Inc.

Austin, J.E., and M.F. Zeitlin (eds.). 1981. *Nutrition intervention in developing countries*. Cambridge, MA: Oelgeschlager, Gunn & Hain.

Beaton, G.H. 1983. Energy in human nutrition: Perspectives and problems. *Nutrition Reviews* 41(11): 325–340.

Behrman, J.R., and A.B.. Deolalikar. 1987. Health and nutrition. In *Handbook of development economics*, ed. H.B. Chenery and T.N. Srinivasan. Amsterdam: North-Holland.

Berg, A. 1973. *The nutrition factor: Its role in national development*. Washington, DC: Brookings Institution.

Binswanger, H.P., and M.R. Rosenzweig. 1981. *Contractual arrangements, employment and wages in rural labour markets: A critical review*. New York: Agricultural Development Council.

Bliss, C., and N. Stern. 1978. Productivity, wages and nutrition; Parts I and II. *Journal of Development Economics* 5(4): 331–398.

Johnston, B.F., and W.C. Clark. 1982. *Redesigning rural development: A strategic perspective*. Baltimore: Johns Hopkins University Press.

Payne, P., and P. Cutler. 1984. Measuring malnutrition: Technical problems and ideological perspectives. *Economic and Political Weekly* 19(34): 1485, New Delhi.

Pinstrup-Anderson, P., et al. 1976. The impact of increasing food supply on human nutrition: Implications for commodity priorities in agricultural research and policy. *American Journal of Agricultural Economics* 58(2): 131–142.

Reutlinger, S., and H. Alderman. 1980. The prevalence of calorie-deficient diets in developing countries. *World Development* 8: 399–411.

Reutlinger, S., and M. Selowsky. 1976. *Malnutrition and poverty: Magnitude and policy options*, World Bank Occasional Paper, vol. 23. Baltimore: Johns Hopkins University Press.

Seckler, D. 1982. Small but healthy: A basic hypothesis in the theory, measurement, and policy of malnutrition. In *Newer concepts in nutrition and their implications for policy*, ed. P.V. Sukhatme. India: Maharashtra Association for the Cultivation of Science Research Institute.

Selowsky, M., and L. Taylor. 1973. The economics of malnourished children: An example of disinvestment in human capital. *Economic Development and Cultural Change* 22(1): 17–30.

Solimano, G., and L. Taylor. 1980. *Food price policies and nutrition in Latin America*. Tokyo: United Nations University.

Srinivasan, T.N. 1981. Malnutrition: Some measurement and policy issues. *Journal of Development Economics* 8(1): 3–19.

Streeten, P., et al. 1982. *First things first: Meeting basic human needs*. London: Oxford University Press.

Sukhatme, P.V. (ed.). 1982. *Newer concepts in nutrition and their implications for policy*. India: Maharashtra

Association for the Cultivation of Science Research Institute.

Timmer, C.P. 1981. Is there ‘curvature’ in the Slutsky matrix? *Review of Economics and Statistics* 62(3): 395–402.

Timmer, C.P. 1986. *Getting prices right: The scope and limits of agricultural price policy*. Ithaca: Cornell University Press.

Timmer, C.P., W.P. Falcon, and S.R. Pearson. 1983. *Food policy analysis*. Baltimore: Johns Hopkins University Press for the World Bank.

Waterfield, C. 1985. Disaggregating food consumption parameters. *Food Policy* 10(4): 337–351.

Nutrition and Development

Paul Glewwe

Abstract

The nutritional status of children and adults is primarily determined by consumption of foodstuffs that contain macronutrients and micronutrients and by the incidence of gastrointestinal diseases. Insufficient nutrition among young children has particularly severe negative consequences. Factors that lead to better nourished children include better-educated mothers, higher household income, potable water and sanitary toilet facilities. The most effective nutrition programmes target children during their first two years of life; such programmes increase life cycle income by raising children’s levels of education. Economists should focus their research efforts on empirical studies that use panel data and data from randomized trials.

Keywords

Bargaining power; Education; Labour productivity; Malthus, T; Medical care; Nutrition; Panel data; Random experiments; Simultaneity bias

JEL Classifications

O1

Economists have studied human nutrition since Thomas Malthus published his *Essay on the Principle of Population* in 1798. His pessimistic predictions about economic growth and human welfare proved to be incorrect; in today’s developed countries the primary nutrition problem for the majority of the population is obesity, not lack of food. Yet in low-income countries the nutritional status of both children and adults can have a substantial effect on the incomes of individuals and on the rate of economic growth. In addition, the nutritional status of poor children remains a policy concern in almost all developed countries. This article reviews recent research on the factors that affect the nutritional status of children and adults, and the causal impact of child and adult nutrition on income and on other economic outcomes. It focuses on developing countries, where nutritional problems are the most severe and thus their consequences are the largest. For recent studies of nutrition in developed countries, see Kenkel and Manning (1999) and Currie (2000).

Factors That Determine Child and Adult Nutritional Status

Malnutrition can be defined as the lack of sufficient nutrients for human growth and/ or for carrying out daily work and non-work activities. Nutrients can be classified into two broad groups: macronutrients, which are primarily calories and protein; and micronutrients, the vitamins and minerals that are essential for good health. Lack of macronutrients is often caused by insufficient consumption of staple foodstuffs, while lack of micronutrients often reflects an unbalanced diet. The most serious micronutrient deficiencies in developing countries are lack of iron, iodine, vitamin A and zinc. (See Behrman et al. 2004, and the references therein for further details.)

The nutritional status of children and adults in developing countries (and in developed countries) is primarily determined by consumption of foodstuffs that contain macronutrients and micronutrients and by the incidence of gastrointestinal diseases that interfere with the body’s ability to extract macronutrients and micronutrients from those foodstuffs. By far the most serious manifestation of such diseases is the incidence of diarrhoea among very young children. The consumption of foodstuffs is in turn determined by household income and food and non-food prices, as proposed by standard demand theory. Some countries have programmes that provide households with food rations or food coupons (stamps) that can be used to purchase food items, which effectively loosens households' budget constraints. The incidence of gastro-intestinal diseases is mainly due to three factors: exposure to infectious diseases, the health knowledge of both children and adults, and the availability (and prices) of medicines and medical care services. A final consideration is the allocation of foodstuffs and medical care within the household, which is likely to depend on the relative bargaining power of key household members; in particular, several studies have shown that children are better nourished in households in which their mothers have a relatively high level of bargaining power (see Strauss and Thomas 1998, for references).

Many economists, nutritionists and other researchers have attempted to identify the most important causes of malnutrition among children in developing countries. This research is motivated in part by estimates indicating that about 30 per cent of the children in those countries are seriously underweight (de Onis et al. 2004) and about 1.7 million children die every year from malnutrition and diarrhoea (WHO 2003). Careful empirical studies of children's nutritional status in developing countries have provided credible evidence that the following factors have strong causal effects: mother's education; mother's health knowledge; infant breastfeeding; household income; potable water; and modern toilet facilities. Several specific policy interventions have also been shown to have a strong positive impact on child nutrition: oral rehydration therapy (ORT) for children with diarrhoea; monitoring of child growth; programmes that provide health and nutrition information to mothers; and fortification of commonly purchased food items (such as salt and sugar) with selected micronutrients (see World Bank 2004; Filmer 2003, for detailed references).

The factors that determine the nutritional status of adults in developing countries have received less attention from economists and other researchers, primarily because in almost all cases the impact of poor nutrition on adults is thought to be less harmful than the impact of poor nutrition on children. Higher incomes, higher education levels and availability of health care services all have positive impacts on adult health and nutrition. Yet these impacts are not necessarily very strong; for example, the income elasticity of calorie consumption is quite low (Strauss and Thomas 1998), although it is somewhat higher for the poorest households. Similarly, Haddad et al. (2003) found that the income elasticity of the rate of child malnutrition is less than one in 11 of the 12 countries they analysed.

While the results summarized in the previous two paragraphs are intuitively plausible, they can be challenged because formidable econometric problems confound attempts to estimate the determinants of the nutritional status of both children and adults. Several recent studies have carefully attempted to overcome problems of simultaneity bias (for example, food intakes, income and medical treatments are all jointly endogenous), but attenuation bias due to measurement error in the explanatory variables has received less attention. The most convincing studies are those based on either panel data or randomized experiments.

Impact of Poor Nutrition on Socioeconomic Outcomes

The impacts of poor nutritional status on important socioeconomic outcomes vary according to the age of the individual when he or she is malnourished. It is useful to consider separately the following three age ranges: from birth to about five years (before children are enrolled in primary school), from six years to the early teenage years (when most children are enrolled in school and not working), and from the late teenage years through retirement age (the working-age years).

Empirical evidence suggests that poor nutrition in the first few years of life can have substantial negative consequences for educational outcomes and, eventually, for adult income (see World Bank 2004; Glewwe 2005, for recent reviews). For example, Glewwe et al. (2001) show that children in the Philippines who were malnourished during the first two years of their lives start school at a relatively late age and learn less per year while in school. The precise mechanisms are not completely clear, but it is likely that inadequate nutrition during the first years of life affects the physical development of the brain in ways that cannot be easily reversed. For example, iodine deficiency impairs the development of the central nervous system. The reduction in skills obtained from schooling due to poor nutrition in the preschool years almost certainly has large negative impacts on children's incomes when they become adults, and back-of-an-envelope calculations suggest that the benefits (in terms of life cycle income) of programmes to reduce malnutrition among very young children are much higher than the costs (see Glewwe et al. 2001, for an example of such calculations).

Many nutrition programmes in both developed and developing countries are designed to provide nutritious breakfasts and lunches to children on the days they are in school. In developing countries, there is little research on the impact of these programmes on educational outcomes. Almost all of the existing literature suffers from serious estimation problems and/or small sample sizes (see Glewwe 2005, for a detailed discussion). While it may seem obvious that providing breakfasts and/or lunches to students would increase their learning, parents may reduce the amount of food provided at home in response to provision of meals at school; surprisingly, Jacoby (2002) found no evidence that parents respond in this way. Perhaps the best evidence on the impact of the nutritional status of the learning of school age children is a recent randomized study of Kenyan pre-schools. Vermeersch and Kremer (2005) provide evidence of a positive impact of school feeding programme on learning, although only in schools with more experienced teachers. Further research is needed on the impact on education outcomes of child nutrition during their years in school. It may be that improvements in schooling outcomes from school feeding programmes are primarily due to increases in daily attendance brought about by those programmes, as opposed to higher nutritional status among children who participate in those programmes.

Finally, many economists have examined the role of nutrition during adulthood on concurrent labour productivity and labour income. Robert Fogel has studied this phenomenon in the United States and Europe in the 18th and 19th centuries (see Fogel 1999), but the historical data are too incomplete to resolve a host of econometric issues in a convincing way. Some economists have developed 'efficiency wage' models in which low nutrition among adults can lead to involuntary unemployment. Strauss and Thomas (1998) provide a recent review of the empirical evidence on the relationship between adult nutrition, labour productivity and income. There is strong evidence that, *Ceteris paribus,* better- nourished adults (as measured by body mass index) are more productive workers. (There is also evidence that taller workers are relatively more productive, but height is primarily determined by nutritional status during childhood.) Yet Swamy (1997) and others have presented strong evidence that the estimated magnitudes of the effect of current nutritional status on worker productivity are far too small to be a cause of unemployment in developing countries.

Much has been learned in recent years about the relationship between nutrition and economic and social outcomes in developing countries, but even more remains to be learned. The evidence to date suggests that the most effective, and most cost- effective, nutrition programmes are those that are targeted to children during their first two years of life, for whom the main benefits are a higher rate of survival into adulthood and an increase in life cycle income brought about by higher levels of education. The most convincing studies are based on either panel data or randomized trials, but such data are available for only a handful of countries. Indeed, the Cebu Longitudinal Health and Nutrition Survey, which covers only one region of the Philippines, is the data source for many of the most convincing studies based on panel data. While randomized trials, such as Gertler's (2004) assessment of the impact of Mexico's *Progresa* programme on children's nutritional status, can be a very effective method for assessing programme and policy impacts, one must wait many years before long-term impacts can be measured.

Very little is known about the impact of poor nutrition among school-age children on academic performance and, ultimately, adult income, and the same is true of the impact of policies and programmes designed to improve the nutritional status of adults. A very recent policy option that deserves careful study is the development and provision of genetically modified foodstuffs that contain higher levels of essential nutrients. An example of this is 'golden rice' that has been fortified with vitamin A. To provide useful information for policymakers, economists' research efforts in the area of nutrition should not be devoted to developing theoretical models but instead should focus on empirical studies that make careful use of panel data and data from randomized trials. This will require new data collection efforts, but the cost of such data collection is very small compared with the potential benefits.

See Also

- ▶ Anthropometric History
- ▶ Child Health and Mortality
- ▶ Efficiency Wages
- ▶ Fertility in Developing Countries
- ▶ Human Capital, Fertility and Growth
- ▶ Longitudinal Data Analysis
- ▶ Returns to Schooling
- ▶ Treatment Effect

Bibliography

Behrman, J., H. Alderman, and J. Hoddinott. 2004. Hunger and malnutrition. In *Global crises, global solutions*, ed. B. Lomborg. Cambridge: Cambridge University Press.

Currie, J. 2000. Child health in developed countries. In *Handbook of health economics*, ed. A. Culyer and J. Newhouse. Amsterdam: North Holland.

De Onis, M., M. Blössner, E. Borghi, E. Frongillo, and R. Morris. 2004. Estimates of global prevalence of childhood underweight in 1990 and 2015. *Journal of the American Medical Association* 291: 2600–2606.

Filmer, D. 2003. *Determinants of health and education outcomes*. Washington, DC: World Bank.

Fogel, R. 1999. Catching up with the economy. *American Economic Review* 89: 1–21.

Gertler, P. 2004. Do conditional cash transfers improve child health? Evidence from *PROGRESA's* control randomized experiment. *American Economic Review* 94: 336–341.

Glewwe, P. 2005. The impact of child health and nutrition on education in developing countries: Theory, econometric issues and recent empirical evidence. *Food and Nutrition Bulletin* 26: S235–S250.

Glewwe, P., H. Jacoby, and E. King. 2001. Early childhood nutrition and academic achievement: A longitudinal analysis. *Journal of Public Economics* 81: 345–368.

Haddad, L., H. Alderman, S. Appleton, L. Song, and Y. Yohannes. 2003. Reducing child malnutrition: How far does income growth take us? *World Bank Economic Review* 17: 107–131.

Jacoby, H. 2002. Is there an intrahousehold flypaper effect? Evidence from a school feeding program. *Economic Journal* 112: 196–221.

Kenkel, D., and W. Manning. 1999. Economic evaluation of nutrition policy or, there's no such thing as a free lunch. *Food Policy* 24: 145–162.

Strauss, J., and D. Thomas. 1995. Human resources: Empirical modeling of household and family decisions. In *Handbook of development economics*, ed. J. Behrman and T. Srinivasan. Amsterdam/New York/Oxford: Elsevier Science/North Holland.

Strauss, J., and D. Thomas. 1998. Health, nutrition and economic development. *Journal of Economic Literature* 36: 766–817.

Swamy, A. 1997. A simple test of the nutrition-based efficiency wage model. *Journal of Development Economics* 53: 85–98.

Vermeersch, C., and M. Kremer. 2005. *School meals, educational achievement and school competition: Evidence from a randomized evaluation*, Policy Research Working Paper No. 3523. Washington, DC: World Bank.

WHO (World Health Organization). 2003. *The world health report 2003: Shaping the future*. Geneva: WHO.

World Bank. 2004. *World development report 2004: Making services work for poor people*. Washington, DC: World Bank.

Nutrition and Public Policy in Advanced Economies

Janet Currie

Abstract

This article discusses the measurement of nutritional status of populations and examines two classes of tools that policymakers in advanced economies can use to improve nutrition: targeted food and nutrition programmes, and regulation of the food industry. It presents an overview of the economic rationale for providing nutrition programmes (rather than cash assistance), as well as an analysis of some of the difficulties of providing aid in kind – one of the chief difficulties is low take-up of programme benefits by eligible citizens. The overview of regulations suggests that measures aimed at improving nutrition information may be especially attractive.

Keywords

Advertising; Asymmetric information; Food insecurity; Food Stamp Program (USA); Health care regulation; Hunger; In-kind transfers; Medicaid (USA); Nutrition; Nutrition and public policy; Technical change; Well-being

JEL Classifications

H5

Measures of nutritional status such as height, body mass index and the prevalence of nutrient-deficiency diseases are now accepted indicators of well-being. Economic development changes nutritional threats to well-being as populations move from scarcity to abundance. Fogel (1994) links the decline of malnutrition to economic growth, and highlights improvements in nutrition as an engine of growth. Cutler et al. (2003) highlight technological change as a factor in reducing the cost of the production and distribution of food: the average household in the United States spent one-third of its income on food in 1960, but spends less than half that amount on food today.

As a result, public policymakers now struggle against a rising tide of obesity and related diseases such as type 2 diabetes using policy tools that were formulated largely to combat the effects of scarcity. The incidence of type 2 diabetes has doubled since 1995 in the United States, where 30 per cent of adults over the age of 20 are obese. Even in countries like France, which historically had little obesity, rates are increasing rapidly (World Health Organization 2005). Surprisingly, many people in the United States are both overweight and consuming diets that are deficient in fibre, calcium, potassium, magnesium and vitamin E. This juxtaposition suggests that an excess of calories and a deficit of nutrients may in fact be closely related and reflect poor food choices rather than food scarcity.

This article considers the difficulties involved in tracking the nutritional status of populations and examines two classes of tools that policymakers in advanced economies can use to improve nutrition: targeted food and nutrition programmes and regulation of the food industry.

Measuring Nutrition

Tracking the nutritional status of a population over a long period of time is difficult. Much of Fogel's work relies on the records of army veterans, largely because the veterans represent a large group for whom anthropometric measures are available. Birth weight is also available in many populations over long periods of time (cf. Currie and Moretti 2007).

Going beyond anthropometric measures is generally expensive. Data on food consumption is often collected using food diaries, in which subjects are asked to record everything that they ate (and the amount that they ate) over some specified period such as a day or a week. These entries must then be converted into data about the number of calories from various sources. Clearly, there is likely to be a great deal of measurement error in this type of data, so large sample sizes are needed to uncover any systematic relationships between food intakes and outcomes.

A few data sets such as the National Health and Nutrition Examination Survey (NHANES) in the United States collect information about the levels of specific nutrients using blood and urine tests as well as food diaries. This information is collected as part of a complete physical examination conducted in a mobile clinic. Each wave takes several years to collect, as the mobile examination units travel to interview sites around the country. The expense of collecting the data means that the survey is mounted approximately once a decade. The long intervals between surveys raise additional problems because best practices in terms of ways to measure nutritional status often change between the surveys. Hence, while one can use the NHANES to track changes in body mass index over time, it is difficult to use these data to examine changes in the prevalence of specific nutritional deficiencies.

A fourth source of information about nutrition comes from health surveillance data. Doctors are often required to report the prevalence of specific conditions in their practices to central health agencies. These central agencies in turn can determine how many cases of something like iron deficiency anemia occur in a given population. One suspects that such surveillance systems will tend to underestimate the extent of nutritional deficiencies to the extent that people go untreated, or doctors fail to meet reporting requirements.

Finally, developed countries often produce statistics about the number of people suffering from ‘hunger’. It is important to realize that in advanced economies hunger is a social construct that is not directly related to measures of actual nutritional deficiency. In 1968, a group of physicians issued ‘Hunger in America’, a landmark report documenting appalling levels of malnutrition among poor children. Outright malnutrition is now extremely rare in the developed world. In the United States, people are now classified as hungry if they respond affirmatively to a series of questions in the current population survey. These questions ask whether households are worried about having the money to pay for food, whether there are times that households go without food because they lack money to pay for it, and whether specific household members go without food. These ‘food insecurity’ questions are inexpensive to ask and can be asked more frequently and consistently than the direct measures of nutritional status can be collected in more episodic surveys.

However, once poverty is controlled for, food insecurity is predictive of poorer nutritional outcomes among older household members, but not among children (Bhattacharya et al. 2004). To say that food insecurity is not a direct measure of nutritional deficiency does not mean that it is unimportant. Food insecurity has been linked to higher levels of hyperactivity, absenteeism, aggression and tardiness as well as impaired academic functioning among children, although these linkages may not be causal.

Targeted Food and Nutrition Programmes

Most advanced economies prefer income support to targeted food and nutrition programmes as a way of improving the nutrition (and overall well-being) of their poorest citizens. In contrast, the United States has an array of food and nutrition programmes targeted to specific low-income groups. School meal (or milk) programmes are an exception, in that they are widespread in advanced economies. Apparently the paternalism involved in creating a feeding programme is acceptable when dealing with children, but not (in many countries) when dealing with adults.

Apart from paternalism, economists have developed an array of rationales for providing benefits (including food) in kind, rather than in cash. One common rationale for government intervention in kind is that malnourished citizens create negative externalities for other citizens, through the psychological distress of those who interact with them, burdens on social programmes and health care systems, or their own inability to work.

A second set of arguments has to do with informational asymmetries. Since the government cannot perfectly identify those who need help, it must create schemes that will encourage self-selection. Such schemes often involve penalizing recipients through stigma or through the imposition of non-trivial transactions costs (see Blackorby and Donaldson 1988; Besley and Coate 1991, 1995).

A final rationale is more dynamic: the government fears that cash aid will not be spent as intended, and that recipients will return again and again. The problem is that the government cannot credibly commit to cut off starving people, even if the needy person has squandered past aid (Bruce and Waldman 1991).

These models shed some light on the question of why in-kind programmes are set up as they are, with often substantial barriers to entry and consequent lack of take-up by the neediest people (see Currie 2006b, for a discussion of the take-up of these programmes, and of factors that affect it).

A complete survey of the literature assessing US in-kind food and nutrition programmes is beyond the scope of this article, but see Currie (2006a, ch. 3) for more details about the programmes discussed here and evidence regarding their effectiveness. These programmes take various forms and target various groups. The largest and most studied include the Food Stamp Program (FSP), the Supplemental Nutrition Program for Women, Infants, and Children (WIC) and the National School Lunch Program (NSLP). These three programmes have adopted very different approaches to improving nutrition in disadvantaged families.

The NSLP (and the smaller School Breakfast programme) provide free or reduced-price meals conforming to certain nutritional guidelines directly to their target population. The programme is available in most US government-sponsored schools and serves approximately 27 million lunches every day at a cost of about six billion dollars annually. The FSP provides electronic debit cards that can be redeemed for food with few restrictions on the types of foods which can be purchased. The programme serves about 20 million households at a cost of roughly 19 billion dollars. WIC offers coupons that may be redeemed only for specific types of food (often specific brands), to women, infants, and children under five who are certified to be at nutritional risk. WIC also involves a significant nutrition education component, which is largely absent from the other two programmes. This programme serves about eight million people each month, at a cost of approximately four billion dollars.

WIC packages are tailored to the nutritional needs of each of the target groups. The programme has been credited with virtually eliminating iron deficiency anemia among infants and young children and with improving birth weight and birth outcomes among the most disadvantaged mothers in an extremely cost-effective manner. There is less research available about WIC's effects on young children. In the past, WIC promoted bottle over breast-feeding by giving mothers free infant formula. Ongoing strenuous efforts are being made to promote breast-feeding and give nursing mothers food packages of equal value to those received by mothers getting formula.

The near-unanimous consensus regarding the positive effects of WIC on infant outcomes has been disturbed in recent years by those who argue that there may be unobserved factors that are correlated both with positive infant health outcomes and with WIC participation. While this is true in theory, careful analyses of selection in the WIC programme suggest that it is the most disadvantaged eligible women who participate, and it is unlikely that they have other positive unobserved characteristics that are driving the findings – that is, selection is probably leading to underestimates of the effects of WIC.

At some points WIC has also generated controversy by enrolling more infants than the government estimated to be eligible. A National Academy of Sciences report on the subject found, however, that the number of those eligible was underestimated, and that the programme fell a long way short of full take-up (National Research National Research Council 2003). The fact that many eligible people do not participate in food and nutrition programmes remains a far more significant problem than participation by ineligible people.

FSP benefits are available to all households with incomes less than 130 per cent of the poverty threshold. FSP benefits can be used to purchase virtually any foods at almost all grocery stores. Since the benefits are generally less than the household's food budget, economic theory suggests that the benefit should be treated in the same way as a cash transfer. But several food stamp 'cash-out' experiments in which treatment households were given cash instead of food stamps while control households continued to receive food stamps suggested that the cash-out reduced spending on food.

However, Whitmore (2002) re-analysed data from one such experiment and found that only households whose benefits exceeded their food budgets initially reduced spending in response to the cash-out. Thus it appears that the FSP may in fact be no different from a cash transfer. It is thus worth asking whether the FSP plays any role other than serving as an indirect cash safety net that is available to the many US households that do not qualify for any other form of assistance. Given that virtually any type of food can be purchased, the FSP should not be expected to have much impact on the quality of the diet, other than via relaxation of the budget constraint. Evidence that people buy and sell stamps (often doing both within a month) further suggests that FSP benefits are treated like cash.

Studies of the FSP shed a good deal of light on the question of take-up and again suggest that lack of participation by eligible people is a greater problem than participation by those who are ineligible. Enrolments in the FSP grew rapidly in the early 1990s following the expansion of the federal Medicaid programme. Households could sign up for Medicaid and the FSP at the same office, so households that were attracted by Medicaid also signed up for FSP. Conversely, the 1996 welfare reform in the United States was accompanied by a decline in FSP participation even among those who remained eligible. Those who lost eligibility for cash benefits were no longer automatically eligible for the FSP and the fact that people were now required to go through enrolment procedures for the FSP and to repeat those procedures every three to six months drove many eligible people away. These examples suggest that transactions costs are an important deterrent to enrolment in means-tested transfer programmes.

The NLSP operates in a way that is similar to school meal programmes in many other countries. In the United States the poorest children are eligible for free meals, while slightly better off children are eligible for reduced-price meals, and other children can purchase school meals at 'full price'. The meals are subject to US government dietary guidelines, which were revised in 1994 to limit the amount of fat and sodium.

Evaluations suggest that the NLSP has successfully raised the consumption of important nutrients. At the same time, meals have been roundly criticized for being high in calories, fat and sodium. Still, the evidence suggests that many American children have extremely unhealthy diets which are improved somewhat through participation in the NLSP.

Like other food and nutrition programmes, the NLSP has been criticized for serving too many ineligible children. The US government conducted several studies of this issue, experimenting with different ways to tighten controls on eligibility. In every case, 'reforms' were more likely to discourage eligible children from applying than they were to reduce programme use by ineligible children (see Neuberger and Greenstein 2003). As a result of these policy experiments, the US government adopted several measures designed to make it easier for poor families to document and maintain eligibility when these programmes were re-authorized in 2004.

Regulation

Traditionally, regulation of the food industry has aimed to ensure the safety of the food supply. However, regulation has been increasingly used as a tool to improve the quality of the diet. Governments in advanced economies have mandated the inclusion of important nutrients such as iodine in salt (which has eliminated goitre), vitamin D in milk (which has helped to eliminate rickets), and folic acid in flour (which has greatly reduced the incidence of neural tube birth defects). Increasingly, regulation is being targeted at the information available to consumers, through labelling and advertising.

There is a good deal of evidence that consumers respond to food labels. Ippolito and Mathios (1990) examine the effect of a US government decision to allow cereal makers to advertise the link between fibre and cancer reduction. The change led to increased advertisement of fibre content, as well as other content information, and to increases in the consumption of high-fibre cereals. Ippolito and Mathios (1995) found that consumption of fat had been declining secularly, but that it declined more rapidly after manufacturers were allowed to advertise health claims associated with low-fat products.

It is however, unclear whether food labels have allowed consumers to make food choices that are healthier overall. Marketing studies suggest that few consumers consult labels assiduously and that many are unaware of the nutritional contents of items in their food baskets. Moreover, food-away-from-home constitutes a large and growing fraction of total consumption and is largely exempt from labelling regulations.

Low socio-economic status households have higher propensities to suffer from nutrition related disorders and are also least likely to use labels. However, some labelling requirements have encouraged manufacturers to reformulate their products in ways that will benefit all consumers, whether or not they read labels. For example, a recent US requirement that manufacturers label ‘transfats’ has led many producers of products such as crackers to substitute transfats with less harmful fats.

Governments have also acted directly in the limitation of the consumption of unhealthy foods. Many US school districts have removed ‘junk food’ and soft drinks from vending machines, and federal legislation that would require this of all school districts has been introduced. France and the United Kingdom have taken similar measures nationally and the UK is going further by banning burgers and processed sausages in schools and requiring two servings of fruit and/or vegetables per day.

The UK also banned the use of celebrities to advertise junk food during children’s television programming and the use of film tie-in advertisements in 2006 (*Guardian* 2006). Several studies indicate that the majority of food advertising directed at children is for relatively unhealthy foods, and a recent report from the National Academy of Sciences (Institute of Medicine 2006) concluded that children’s preferences are significantly swayed by such advertising, and called for either voluntary or regulatory controls on the advertising of food to children.

Given our increasing knowledge about the links between poor food choices and future health, and the rising social costs of providing health care for nutrition-related conditions such as diabetes, additional future regulation is likely. Government intervention can be viewed as a way of reducing the externalities created by poor individual choices, which in turn may be encouraged by food producers who do not bear the social costs created by their products. Economists can contribute to this important public health debate by analysing the costs and benefits of regulation.

See Also

- ▶ Health Outcomes (Economic Determinants)
- ▶ Nutrition and Development
- ▶ Poverty Alleviation Programmes

Bibliography

Bhattacharya, J., J. Currie, and S. Haider. 2004. Poverty, food insecurity, and nutritional outcomes in children and adults. *Journal of Health Economics* 23: 839–862.

Besley, T., and S. Coate. 1991. Public provision of private goods and the redistribution of income. *American Economic Review* 81: 979–984.

Besley, T., and S. Coate. 1995. The design of income maintenance programs. *Review of Economic Studies* 62: 187–221.

Blackorby, C., and D. Donaldson. 1988. Cash versus kind, self-selection and efficient transfers. *American Economic Review* 78: 691–700.

Bruce, N., and M. Waldman. 1991. Transfers in kind: Why they can be efficient and nonpaternalistic. *American Economic Review* 81: 1345–1351.

Currie, J. 2006a. *The invisible safety net: Protecting poor children and families*. Princeton: Princeton University Press.

Currie, J. 2006b. The take-up of social benefits. In *Poverty, the distribution of income, and public policy,* ed. A. Auerbach, D. Card, and J. Quigley. New York: Russell Sage.

Currie, J., and E. Moretti. 2007. Biology as destiny? Short and long-run determinants of intergenerational transmission of birth weight. *Journal of Labor Economics* 25: 231–264.

Cutler, D., E. Glaeser, and J. Shapiro. 2003. Why have Americans become more obese? *Journal of Economic Perspectives* 17(3): 93–118.

Fogel, R. 1994. Economics growth, population theory, and physiology: The bearing of long-term processes on the making of economic policy. *American Economic Review* 84: 369–395.

Guardian. 2006. Supermarkets feel junk food ad ban bite. *The Guardian,* November 27.

Institute of Medicine. 2006. *Food marketing to children and youth: Threat or opportunity?* ed. M. McGinnis, G. Jennifer, and K. Vivica. Washington, DC: National Academies Press.

Ippolito, P., and A. Mathios. 1990. Information, advertising and health choices. *RAND Journal of Economics* 21: 459–480.

Ippolito, P., and A. Mathios. 1995. Information and advertising: The case of fat consumption in the United States. *American Economic Review* 85: 91–95.

National Research Council. 2003. *Estimating eligibility and participation in the WIC program*. Washington, DC: National Research Council.

Neuberger, Z., and R. Greenstein. 2003. *What have we learned from FNS' new research findings about overcertification in the school meals programs?* Washington, DC: Center on Budget and Policy Priorities.

Whitmore, D. 2002. What are food stamps worth? Working Paper No. 468. Industrial Relations Section, Princeton University.

World Health Organization. 2005. *The challenge ofobesity in the WHO European Region. Fact sheet EURO/13/05*. Geneva: World Health Organization.